Universidad San Jorge

Facultad de Ciencias de la Salud

Grado en Bioinformática

Proyecto Final

Implementación de código para realizar alineamientos de pares de bases, alineamientos múltiples y árboles filogenéticos adaptado a las necesidades de análisis de Exopol

Autor del proyecto: Silvia del Caso Yagüe

Director del proyecto: Beatriz Ranera Beltrán

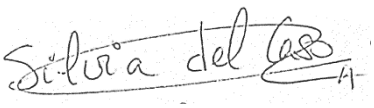
Zaragoza, 9 de septiembre de 2021

Este trabajo constituye parte de mi candidatura para la obtención del título de Graduado o Graduada en Bioinformática por la Universidad San Jorge y no ha sido entregado previamente (o simultáneamente) para la obtención de cualquier otro título.

Este documento es el resultado de mi propio trabajo, excepto donde de otra manera esté indicado y referido.

Doy mi consentimiento para que se archive este trabajo en la biblioteca universitaria de Universidad San Jorge, donde se puede facilitar su consulta.

Firma



Fecha

9 de septiembre de 2021

Dedicatoria y agradecimientos

A la Dra. Beatriz Ranera por haberme enseñado una nueva profesión durante estos tres años y haberme dirigido este proyecto.

Al Dr. Alfredo Benito, por tu confianza y tu gran ayuda en este proyecto y en mis inicios profesionales en Exopol.

A Iñaki Albizu, por cederme tu código para presentarlo en este proyecto y permitirme hacer todos estos cambios.

A Sofía Lázaro, por tu paciencia para enseñarme hasta el último detalle de Exopol y ayudarme en este proyecto y en todo lo que te pido.

A todo el equipo de Exopol por darme esta gran oportunidad profesional, permitirme realizar este proyecto en la empresa y, sobre todo, por toda la ayuda y ánimos que me dais día a día.

A todos mis amigos por estar siempre ahí, a pesar de que hace mucho tiempo que no nos vemos.

A Clara Isabel y Eduardo por tratarme como una hija, muchas gracias por todo.

A mi hermana Alicia por ser mi referente, darme buenos consejos y conocerme tan bien, compartir la vida contigo es el mejor regalo. A Diego por estar siempre dispuesto a ayudarme y ser ya como mi hermano. A Sofía por tus abrazos y primeras palabras que me dan tanta alegría y ganas de vivir.

A mis padres María Isabel y José Luis por ser mi gran apoyo y refugio, gracias por enseñarme con vuestro ejemplo a trabajar duro y con dedicación, pero también que lo más importante siempre es la familia. Gracias por confiar y creer en mí, y cuidarme tanto. Es un orgullo ser vuestra hija.

A ti Eduardo, porque cada día contigo es como el primero, gracias por animarme y apoyarme incondicionalmente. Nuestra vida juntos es nuestro mejor proyecto y no ha hecho más que empezar.

Índice

Resumen

La bioinformática proporciona herramientas para realizar un seguimiento de la evolución genética y el control de virus, como Influenza A porcino o el virus del Síndrome Reproductivo y Respiratorio Porcino (PRRS), que ocasionan pérdidas económicas al sector ganadero, y pueden comprometer la salud humana. Los objetivos de este trabajo son: realizar un estudio filogenético de las secuencias de Influenza A porcino para implementar en R mejoras para este análisis en Exopol, y seleccionar los parámetros para calcular la identidad de secuencias de PRRS, e implementarlos en R para Exopol. Del estudio filogenético realizado con MEGA-X para Influenza A porcino, se han seleccionado los parámetros con menos coste computacional para implementarlos en R: MUSCLE, modelo de evolución Tamura-Nei 1993, Unión de Vecinos y *bootstrap.* Existen diferencias entre los árboles obtenidos con MEGA-X y R causadas por no implementar árboles consenso, ni utilizar el mejor modelo de evolución por su coste computacional. Los parámetros seleccionados para calcular la identidad en secuencias de PRRS con R, están basados en la herramienta LALIGN: alineamiento Smith-Waterman, para nucleótidos una matriz de sustitución que incluye nucleótidos ambiguos, para aminoácidos la matriz BLOSUM50. Los resultados obtenidos con la implementación en R son iguales que los de LALIGN.

Palabras clave: árbol filogenético, alineamiento de pares de secuencias, alineamiento múltiple, Influenza A porcino, Virus del Síndrome Reproductivo y Respiratorio Porcino (PRRS).

Abstract

Bioinformatics provides tools to track the genetic evolution and control of viruses, such as swine Influenza A or the Porcine Reproductive and Respiratory Syndrome virus (PRRS), which cause economic losses to the livestock sector, and can compromise human health. The objectives of this work are: to carry out a phylogenetic study of the swine Influenza A sequences to implement in R improvements for this analysis in Exopol, and to select the parameters to calculate the identity of PRRS sequences, and to implement them in R for Exopol. From the phylogenetic study carried out with MEGA-X for swine Influenza A, the parameters with the least computational cost have been selected for implementation in R: MUSCLE, Tamura-Nei 1993 evolution model, Neighbour-Joining and bootstrap. There are differences between the trees obtained with MEGA-X and R caused by not implementing consensus trees, not using the best evolution model due to its computational cost. The parameters selected to calculate the identity in PRRS sequences with R, are based on the LALIGN tool: Smith-Waterman alignment, for nucleotides a substitution matrix that includes ambiguous nucleotides, for amino acids the BLOSUM50 matrix. The results obtained with the implementation in R are the same as those of LALIGN.

Key words: phylogenetic tree, sequence pair alignment, multiple alignment, swine Influenza A, Porcine Reproductive and Respiratory Syndrome (PRRS).

1. Introducción

1.1 Alineamientos

La bioinformática representa un nuevo campo científico que aporta herramientas informáticas que ayudan a revelar los mecanismos fundamentales que subyacen en los problemas de la biología. La bioinformática desarrolla bases de datos y algoritmos para analizar proteínas, genes y genomas, intentado también gestionar las enormes cantidades de datos que se generan en los proyectos de secuenciación del genoma, proteómica y otros proyectos a gran escala [1].

Una de las preguntas básicas en biología molecular es si un gen o una proteína está relacionado con cualquier otro gen o proteína. Si existe esta relación entre dos proteínas a nivel de su secuencia, sugiere que son homólogas. Y, por tanto, también sugiere que pueden tener funciones comunes y que comparten una ascendencia evolutiva común. Estos análisis de relación de proteínas o genes se llevan a cabo alineando las secuencias y comparándolas. El alineamiento de secuencias es fundamental para todo el campo de la bioinformática. En los años setenta y ochenta, se describieron los primeros algoritmos de alineamientos de pares de secuencias, realizándose en sus inicios manualmente. Posteriormente en la década de los noventa, se desarrollaron nuevos métodos de alineamiento que tienen en cuenta, por ejemplo, aspectos estadísticos [1].

Es posible realizar alineamientos de secuencias de ADN o de la proteína que codifica, aunque suele ser más informativo comparar las secuencias de proteínas. Hay varios motivos como:

- Muchos cambios en una secuencia de ADN, sobre todo si son en la tercera posición del codón, no cambian el aminoácido en la secuencia proteínica (imagen 1).
- Muchos aminoácidos comparten propiedades biofísicas, por ejemplo, la lisina y la arginina son aminoácidos básicos, y el intercambio de uno por otro pueden no modificar la función o estructura de la proteína resultante.
- Es posible evaluar mediante alineamientos basados en sistemas de puntuación (se describirán más adelante) el cambio por un aminoácido relacionado, pero no coincidente.
- Las comparaciones de secuencias de proteínas pueden identificar secuencias homólogas, pero las secuencias de ADN correspondientes no pueden.

IMAGEN 1. Código genético: se muestran los 64 posibles codones junto con su frecuencia de aparición en la naturaleza y el aminoácido que codifican. Se muestran las cuatro bases A, C, G, T. Cada codón está compuesto por tres bases, por lo tanto, hay 4^3 = 64 combinaciones. Imagen modificada del libro *Bioinfomatics and Functional Genomics* [1]

Por todo ello, las secuencias de ADN son menos informativas en estos aspectos y es necesario, si es una secuencia codificante, estudiar su proteína traducida. Sin embargo, en otros muchos casos, es más apropiado comparar directamente las secuencias de nucleótidos, por ejemplo: comprobar la identidad de una secuencia de ADN en una base de datos, buscar polimorfismos, analizar fragmentos de ADNc (ADN complementario) clonados, comparar regiones reguladoras de genes, etc. [1]. Siendo la identidad el grado en el que dos secuencias de aminoácidos o nucleótidos son invariantes, y su cálculo consiste en contar el número de residuos idénticos al comparar dos secuencias.

Dos secuencias son homólogas si comparten un ancestro evolutivo común. No existen grados de homología, o las secuencias son homólogas o no lo son [2]. Las proteínas homólogas suelen presentar:

- una estructura tridimensional significativamente relacionada.
- sus secuencias de aminoácidos o nucleótidos suelen compartir una identidad significativa.

La homología es una inferencia cualitativa, es decir, o es homóloga o no lo es, pero la identidad y la similitud son cuantitativas, por lo que describen el grado de relación de las secuencias. En el caso de que los residuos que se están comparando no sean idénticos, pero estén relacionados estructural o funcionalmente, se puede hablar de similitud. El porcentaje de similitud de dos

secuencias es la suma de coincidencias idénticas y similares. Dos moléculas pueden ser homólogas sin compartir una identidad de aminoácidos o nucleótidos estadísticamente significativa, esto podría ser resultado de una divergencia tan acusada que no comparten una identidad de secuencia reconocible. Sin embargo, la estructura tridimensional diverge mucho más lentamente que la coincidencia de aminoácidos en las secuencias de proteínas [3]. Es un desafío bioinformático reconocer este tipo de homología.

Por otro lado, se han denominado proteínas análogas aquellas que no son homólogas, pero que comparten alguna similitud por casualidad, y, por tanto, se presupone que no descienden de un ancestro común.

El descubrimiento de la homología de las secuencias con respecto a una proteína o familia de proteínas a menudo proporciona las primeras pistas de la función de un gen recién secuenciado [4]. Este descubrimiento, unido a la revolución de la secuenciación del ADN, ha acelerado el estudio de los genomas en las últimas décadas, y del mismo modo se ha impulsado el desarrollo de herramientas bioinformáticas cada vez más potentes y versátiles para su análisis.

La obtención de secuencias biológicas planteó la necesidad de crear bases de datos que las albergaran, teniendo como ejemplo más claro la creación del Centro Nacional de Información Biotecnológica (NCBI) en Estados Unidos en 1988. A medida que las bases de datos de secuencias de aminoácidos, ADN y ARN fueron acumulando cada vez más información, se hizo también necesario el desarrollo de softwares para gestionar esta información y poder analizar las relaciones biológicas entre las secuencias mediante alineamientos. Se desarrollaron herramientas bioinformáticas basadas en algoritmos de puntuación para poder evaluar si estas relaciones biológicas son significativas o son similitudes fortuitas. Estos algoritmos asignan distintas puntuaciones en función de si hay una inserción, una deleción o una sustitución, y calculan una alineación entre dos secuencias que corresponde con el resultado menos costoso para tales mutaciones (imagen 2). Estas puntuaciones se aplican mediante matrices de puntuación, que recogen los valores teniendo en cuenta la probabilidad de que ocurra esa mutación y la distancia evolutiva. Esta alineación podría considerarse como una minimización de la distancia evolutiva o una maximización de la similitud entre las secuencias comparadas [4].

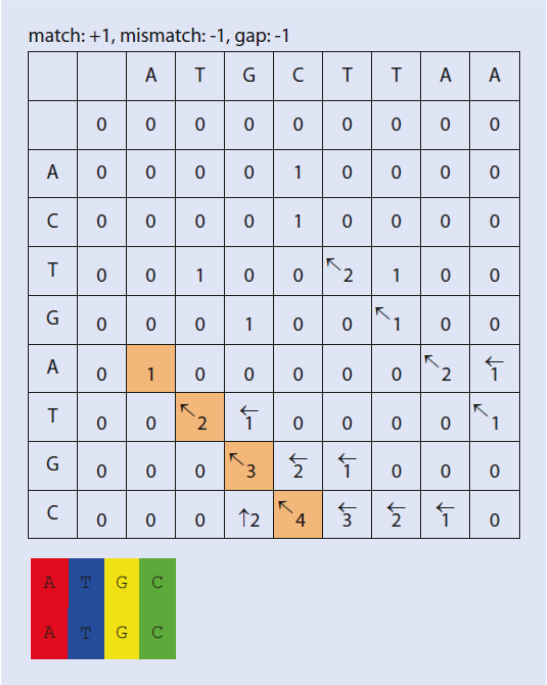


IMAGEN 2. Matriz de puntuación completada mediante el algoritmo Smith y Waterman (1981). El rastreo comienza en el valor más alto de la matriz y solo se alinean las subcadenas de las secuencias. Imagen obtenida del libro *Phylogenomics: An introduction* [5].

1.2. Matrices de puntuación

1.2.1. Matrices de puntuación para alineamientos de proteínas

Las matrices de puntuación, también denominadas matrices de sustitución, son un parámetro clave en los alineamientos. Estas matrices recogen valores específicos para el caso en el que los aminoácidos coincidan, denominado en inglés *match,* y en el caso en el que no coincidan, denominado *mismatch* (imagen 3). En el caso de que no haya coincidencia, se podría hablar de una mutación por sustitución. La primera matriz de puntuación fue obtenida por Margaret Dayhoff (1966, 1978) que elaboró un modelo de reglas basado en los cambios evolutivos que ocurren en las proteínas. Este modelo es la base del sistema de puntuación para los alineamientos de pares de secuencias entre proteínas, ya estén cerca o lejos evolutivamente hablando. Para desarrollar este modelo, se analizaron 34 superfamilias de proteínas divididas en 71 alineaciones. Estudiaron qué sustituciones específicas de aminoácidos se observan cuando se alinean dos secuencias de proteínas homólogas [6].

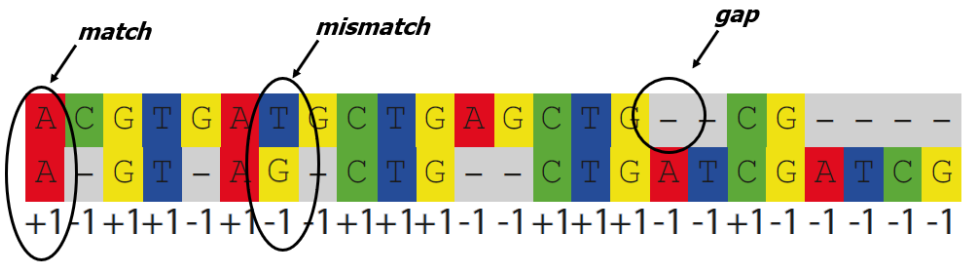


IMAGEN 3. Se considera *match* cuando en dos secuencias alineadas los nucleótidos o aminoácidos a comparar son el mismo. Cuando son distintos es *mismatch* y cuando se insertan huecos en una secuencia son inserciones, y en la otra secuencia son deleciones, es lo que se considera como *gap*. A cada uno de estos eventos se le asigna una puntuación, este caso *match* = 1, *mismatch* = -1 y *gap* = -1. Debajo del alineamiento aparece la puntuación para cada posición, la puntuación final es la suma total de todas ellas, que este ejemplo es cero. Imagen obtenida del libro *Phylogenomics: An introduction* [5].

Definieron una mutación puntual aceptada (PAM) como el reemplazo de un aminoácido en una proteína por otro aminoácido que ha sido aceptado por la selección natural. Esto ocurre cuando un gen sufre una mutación de ADN que codifica un aminoácido diferente, y toda la especie adopta ese cambio como la forma predominante de la proteína. Hay algunos reemplazos que son más frecuentes como la serina por treonina, pero para determinar todos los posibles cambios y definirlos como «aceptados» se basaron en sustituciones observadas empíricamente y realizando un análisis filogenético. Para ello, en lugar de comparar dos residuos de aminoácidos directamente, los compararon con el ancestro común inferido de esas secuencias. También calcularon la mutabilidad relativa, es decir, cuán frecuente es que un aminoácido cambie en un periodo evolutivo corto. Por ejemplo, los residuos que mutan menos suelen ser importantes estructuralmente o funcionalmente para la función de la proteína, por lo tanto, una mutación de ese residuo puede ser perjudicial para ese organismo.

Con los datos sobre las mutaciones aceptadas y las probabilidades de ocurrencia de cada aminoácido generaron una matriz de probabilidad de mutación. Cada elemento de esta matriz muestra la probabilidad de que un aminoácido original sea reemplazado por otro durante un intervalo evolutivo definido, en este caso un PAM, esta matriz resultante se denomina PAM1. Un PAM es la unidad de divergencia evolutiva en la que se ha cambiado el 1% de los aminoácidos entre las dos secuencias de proteínas. Se puede producir una divergencia del 1% en lapsos de tiempo muy diferentes, debido a que las familias de proteínas se someten a sustituciones a diferentes velocidades.

Dayhoff asumió que las mutaciones de aminoácidos son igual de probables en cualquier dirección. Este modelo fue elaborado con proteínas con una relación cercana, pero si queremos comparar dos secuencias más distantes evolutivamente, podemos extrapolar otras matrices a partir de la matriz PAM1. En los casos en los que se dan 250 cambios en una secuencia de 100 aminoácidos de longitud, lo que se hace es multiplicar la matriz PAM1 por sí misma, cientos de veces. De este

modo obtenemos, las matrices PAM10, PAM60 o PAM250, en función del número de veces que la multiplicamos. La precisión de la matriz que se obtiene depende de la precisión que tenía la PAM1, porque se pueden propagar errores, pero si PAM1 está bien calculada es considerada válida. A esta matriz se le aplica el cálculo del logaritmo de la razón de momios llamado en inglés *log-odds*, obteniéndose la llamada Log-Odds matrix, cuyos valores serán los que se utilicen como matriz de puntuación.

En resumen, podemos hacer uso de las matrices PAM10 - PAM60 para secuencias homólogas y las PAM mayores (PAM250) para secuencias más divergentes (imagen 4). Si se comparan dos secuencias muy divergentes con una matriz PAM10, se estará penalizando demasiado un *mismatch,* porque en esta matriz este suceso tiene un valor más negativo que una matriz PAM250. Es cierto que se deben penalizar los cambios, pero si se es muy estricto no se encontrará ninguna similitud cuando sí que la hay. Por el contrario, en secuencias homólogas sí se debe ser conservador, y penalizar mucho las diferencias, para poder encontrarlas. Si no se penalizan, como son secuencias tan iguales, no se detectarán. Sin embargo, en muchas ocasiones no es posible saber cuánta relación hay entre las secuencias por lo que es conveniente probar varias matrices, o utilizar matrices que apliquen puntuaciones más intermedias, como la BLOSUM62 (imagen 4). Las matrices BLOSUM son matrices de sustitución de bloques. Henikoff y Henikoff (1992, 1996) usaron la base de datos BLOCKS (Block Substitutions Matrix) [7]. En esta base de datos, encontraron más de 500 grupos de regiones conservadas –que se denominaron bloques–, en proteínas que estaban relacionadas lejanamente entre sí. Este esquema de puntuación también utiliza el cálculo con *log-odds ratio*, a partir de los valores de probabilidad de sustitución de un aminoácido por otro. Las frecuencias de sustitución para la matriz BLOSUM62 están influenciadas con más fuerza por bloques de secuencias de proteínas que tienen menos del 62% de identidad, por lo tanto, es útil para alineamientos de proteínas que tienen una identidad menor del 62% [7]. La matriz BLOSUM62 es la matriz que usa el programa de búsqueda de proteínas BLAST (Basic Local Alignment Search Tool) del NCBI.

Henikoff y Henikoff (1992) probaron la capacidad de una serie de matrices BLOSUM y PAM para detectar proteínas en búsquedas con BLAST en bases de datos. Descubrieron que BLOSUM62 funcionaba ligeramente mejor que BLOSUM60 o BLOSUM70, y mejor que las matrices PAM en la identificación de varias proteínas. Sus matrices fueron especialmente útiles para identificar alineaciones de puntuación débil [7]. A pesar de esto, las matrices PAM pueden ser de utilidad para identificar la conservación entre dos proteínas estrechamente relacionadas, pero que solamente presentan valores altos de identidad en algunas regiones de la secuencia, pero no en toda.

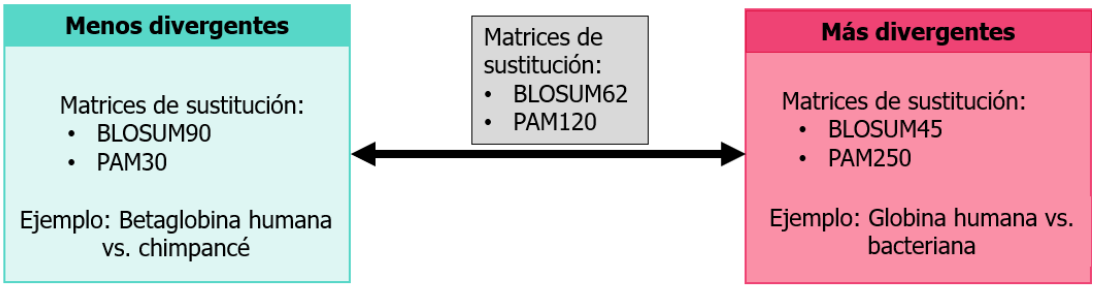


IMAGEN 4. Resumen de las matrices PAM y BLOSUM. Las matrices BLOSUM de número más alto, como por ejemplo BLOSUM 90, y las matrices PAM de número más bajo, como PAM30, son más apropiadas para estudiar proteínas muy conservadas de especies próximas. Las matrices BLOSUM de número bajo o matrices PAM con número alto son más adecuadas para alinear proteínas provenientes de especies más lejanas. La matriz BLOSUM62 es útil cuando no se conoce el grado de relación. Imagen inspirada del libro *Bioinformatics and Functional Genomics* [1].

1.2.2. Matrices de puntuación para alineamientos de secuencias de DNA

Cuando se trata de hacer alineamientos de secuencias de DNA, asignamos una puntuación diferente en función de si los nucleótidos coinciden *(match)* o no coinciden *(mismatch),* (ver imagen 3), y a partir de esos valores se confecciona la matriz de sustitución. Se considera que la probabilidad de que, por ejemplo, una adenina cambie a una adenina, y la probabilidad de que una adenina cambie a cualquier otro nucleótido es la misma. Aplicándose en el primer caso el valor de puntuación de *match*, y en el segundo el valor de *mismatch* [1]. Se desarrolla en el siguiente apartado el modo de uso de estos valores para realizar los alineamientos.

1.3. Algoritmos de alineamiento

Existe una gran variedad de algoritmos de alineamiento como: la programación dinámica que es un método lento, pero de optimización, o los métodos heurísticos o probabilísticos eficientes, que no son exhaustivos, y están diseñados para la búsqueda a gran escala en bases de datos [8].

Un algoritmo de programación dinámica es útil para comparar dos secuencias, pero no lo es para comparar millones de secuencias. Si quisiéramos comparar un número elevado de secuencias mediante un análisis tan exhaustivo, necesitaríamos un ordenador con una memoria enorme, o tardaría en resolverlo un tiempo tan elevado como inaceptable.
Sin embargo, un algoritmo heurístico es aquel que hace aproximaciones de la mejor solución sin considerar exhaustivamente todos los resultados posibles. De este modo un algoritmo heurístico es capaz de resolver estos problemas en un segundo. Un ejemplo de este tipo de algoritmos es el conocido, BLAST (Basic Local Alignment Search Tool), que es un tipo de alineamiento local [4].

Estos algoritmos se utilizan para los distintos métodos que existen de alineamiento en función de si se comparan dos o más secuencias:

- **La alineación de secuencias por pares:** se emplea para identificar regiones de similitud que pueden indicar relaciones funcionales, estructurales y/o evolutivas entre dos secuencias biológicas (proteínas o nucleótidos).
- **La alineación de secuencias múltiples:** es la alineación de tres o más secuencias biológicas de longitud similar. A partir del resultado obtenido de este alineamiento múltiple, se puede inferir la homología y la relación evolutiva entre las secuencias analizadas.
- **La alineación de secuencias frente a secuencias almacenadas en bases de datos:** una de las herramientas de búsqueda más conocidas es BLAST; mediante un alineamiento local compara una secuencia problema con un gran número de secuencias identificadas que se encuentran almacenadas en una base de datos.

1.3.1. Alineamientos de pares de secuencias

Los algoritmos computacionales para realizar alineamientos de dos secuencias se dividen en dos tipos principales: los alineamientos globales y los locales, ambos basados en algoritmos de programación dinámica.

Los alineamientos globales intentan que el alineamiento ocupe la longitud total de las secuencias a comparar. El algoritmo que realiza este tipo de alineamientos es el desarrollado por Needleman-Wunsch en 1970 [9]. El objetivo es utilizarlo cuando se desea comparar la similitud de las secuencias en toda su extensión.
Sin embargo, los alineamientos locales buscan regiones similares dentro de las secuencias que se comparan, no es necesario que las secuencias sean similares en toda su extensión, de hecho, con frecuencia las secuencias que se comparan son muy divergentes entre sí. El algoritmo capaz de llevar a cabo esto es el desarrollado por Smith-Waterman (1981) [10].

Ambos métodos tienen cosas en común, se basan en algoritmos de programación dinámica y emplean matrices de puntuación o sustitución, para asignar valores a las puntuaciones de *match*, *mismatch* o *gap* (imagen 3). Cuando una mutación da como resultado una inserción o una deleción, es decir, se agrega o se elimina un residuo, se representa en el alineamiento como un hueco, en inglés *gap,* mediante un guión en una de las dos secuencias [5]. También es frecuente añadir estos huecos al inicio o al final de una secuencia para igualar la longitud de las dos secuencias que se están comparando, o formando brechas en medio de las secuencias [5].

Se necesita un algoritmo apropiado para realizar la alineación, debido a que hay una enorme cantidad de posibles alineaciones que aumenta exponencialmente con la longitud de la secuencia.

La programación dinámica permite identificar la ruta de alineación óptima. Para ello va calculando la puntuación de las distintas rutas de alineamiento posibles, y en cada paso va tomando la decisión de qué subruta, es decir qué alineamiento, tiene la mejor puntuación (imagen 2) [5].

1.3.1.1. Alineamiento global

El método Needleman-Wunsch está basado en la programación dinámica de alineamientos globales [9]. Lo que se quiere obtener es la mejor puntuación para el alineamiento. Esto se lleva a cabo mediante recursividad, donde la puntuación máxima se obtiene de la maximización de la función F (i, j); ver función 1.

$$F\ (i,j) = \max \begin{cases} F(i-1,j-1) + s\left(X_i, Y_j\right), \\ F(i-1,j) - g, \\ F(i,j-1) - g. \end{cases}$$

Función 1. F (i, j), donde *g* es la penalización del *gap.*

Primero, se calcula la puntuación de la última columna de la matriz de puntuación o sustitución, que es s (X_i, Y_j) en caso de *match* o *mismatch* del par de bases. En el caso de que haya un *gap* en cualquiera de las secuencias, se aplica – g. La puntuación de cada una de las otras columnas es F (i, j), F (i-1, j) o F (i, j-1), dependiendo del camino que se sigua. La puntuación de la alineación óptima es la suma de las puntuaciones de alineación [11].

Para llevar a cabo el método Needleman-Wunsch y maximizar esta función se siguen tres pasos: inicialización de la matriz, llenado de la matriz y rastreo. La inicialización de la matriz consiste, como podemos ver en la imagen 5, en colocar las secuencias a lo largo de los ejes y añadir una fila y una columna. La primera posición de esta fila y columna se contabiliza como cero, y el resto con las múltiples puntuaciones de los *gaps*. También se añade una flecha que señala al cero inicial para marcar la posición de la que viene [5].

		A	C	G	T	G	A	T	G
	0	←-1	←-2	←-3	←-4	←-5	←-6	←-7	←-8
A	↑-1								
G	↑-2								
T	↑-3								
A	↑-4								
C	↑-5								
T	↑-6								
G	↑-7								
C	↑-8								

IMAGEN 5. Inicialización de una matriz por el método de alineamiento global de Needleman and Wunsch (1970). En los ejes están las secuencias, en este caso de DNA, que se van a comparar. La primera fila y la primera columna de la matriz están completadas con la puntuación de los *gaps,* con cada posición aumenta en una unidad el valor de la penalización, además las flechas apuntan al 0. Imagen obtenida del libro *Phylogenomics: An introduction* [5].

En el segundo paso (imagen 6.a) que es el rellenado de la matriz, basándonos en el sistema de puntuación que se haya elegido (por ejemplo: *match* =1, *mismatch* =-1, *gap* = -1), se calculan tres valores para cada celda. Cada valor es el proveniente de cada posición, arriba, izquierda y diagonal arriba, más la puntuación de si es *match*, *mismatch* o *gap*, lo que corresponda. Finalmente se elige el valor más alto para esa celda y se indica con una flecha de dónde proviene. En el caso de que varios de los tres valores sean iguales, se introducen múltiples flechas o se elige una al azar [5].

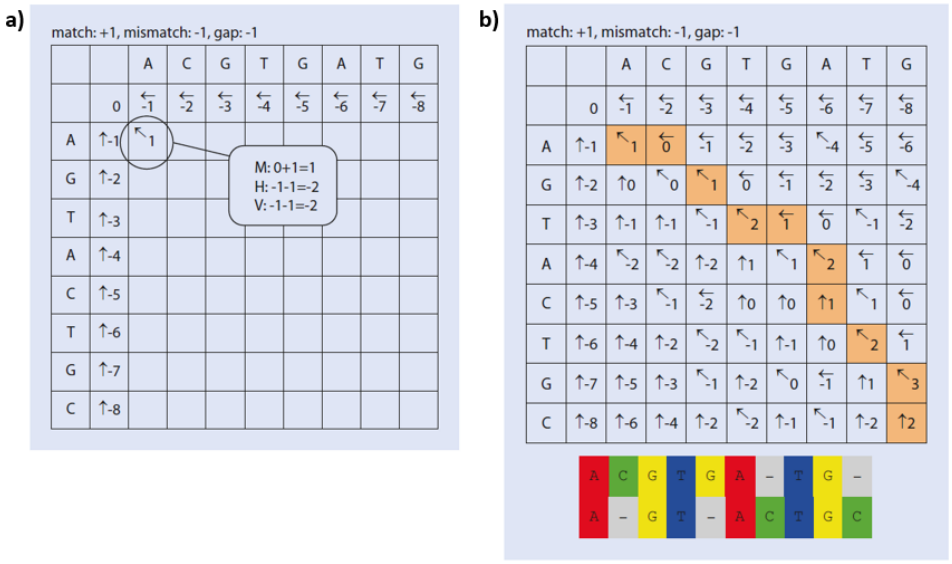


IMAGEN 6. Llenado de la matriz por el método de alineamiento global de Needleman and Wunsch (1970). Sistema de puntuación elegido *match* = 1, *mismatch* = -1, *gap* = -1, y se calculan tres valores para cada casilla. **a)** En la posición rodeada, se puede ver el cálculo de los tres valores, M es coste de *match* o *mismatch*, en este caso se suma al valor de la diagonal (0), la puntuación de match (+1), en total +1. La puntuación del *gap* horizontal, H, es el valor de la izquierda (-1), más la puntuación de gap (-1), en total -2. La puntuación del *gap* vertical, se calcula con el valor de la casilla superior (-1), más la puntuación del *gap* (-1), en total -2. Se elige el valor más alto para llenar la casilla, en este caso, +1 y una flecha indica de dónde proviene, diagonal hacia arriba. **b)** En naranja, están marcadas las celdas del rastreo óptimo. Comenzando por la celda inferior derecha, las flechas marcan la ruta hacia la parte superior izquierda. Debajo de la matriz se muestra el alineamiento óptimo de las secuencias. Imagen obtenida del libro *Phylogenomics: An introduction* [5].

El tercer paso es el rastreo (imagen 6.b), se comienza por la celda inferior derecha, que es el valor de puntuación óptimo del alineamiento, y se van siguiendo las flechas hacia la parte superior izquierda. Si es una flecha diagonal significa que los dos nucleótidos están alineados, si es vertical se alinea con un hueco en la secuencia que está situada en la parte superior (eje X), y si es una flecha horizontal el hueco se coloca en la otra secuencia (eje Y). Si se introdujeron varias flechas en una misma celda, se pueden obtener distintas alineaciones igualmente óptimas.

En el caso de alinear secuencias de proteínas, las puntuaciones se basan generalmente en las matrices de puntuación de las que se ha hablado anteriormente en el apartado 1.2. de esta introducción. Las matrices más utilizadas son BLOSUM y PAM (Henikoff y Henikoff, 1992) que

incorporan las preferencias evolutivas para ciertas sustituciones sobre otros tipos de sustituciones [12]. Para la programación dinámica, estas matrices son la matriz s (X_i, Y_j) en la función a maximizar (función 1). Las puntuaciones en estas matrices se dan como logaritmos de probabilidades, que se pueden usar directamente como parámetros de los esquemas de puntuación de alineación. Las puntuaciones positivas significan que encontramos emparejamientos de aminoácidos con más frecuencia de lo esperado (sustituciones conservadas); los valores negativos indican aquellos que ocurren con menos frecuencia como se esperaba (sustituciones no conservadas) [13]. Alternativamente, se puede implicar una matriz que cuente los pasos para las sustituciones de aminoácidos inferidas del código genético. En este caso, las puntuaciones son −1 (se necesita un cambio en el triplete de codones), −2 (se necesitan dos cambios) o −3 (se necesitan tres cambios). Obviamente, la elección de la función de puntuación y sus parámetros tiene una gran influencia en la selección de la mejor alineación por pares.

1.3.1.2. Alineamiento local

Los alineamientos locales pueden parecer una tarea más compleja que el alineamiento global, ya que habría que realizar muchos alineamientos globales probando con distintos inicios y finales de las secuencias. Sin embargo, Smith y Waterman (1981) crearon una adaptación del alineamiento de Needleman y Wunsch (1970) que resuelve este tipo de alineamientos de una manera sencilla.

La función a maximizar, F (i, j), mediante programación dinámica en este caso es la función 2.

$$F\ (i,j) = \max \begin{cases} 0, \\ F(i-1,j-1) + s\ (X_i, Y_j), \\ F(i-1,j) - g, \\ F(i,j-1) - g. \end{cases}$$

Función 2. F (i, j), siendo *g*, el valor de penalización cuando hay un hueco o *gap.*

De manera similar al algoritmo que se acaba de explicar, se crea una matriz basada en la longitud de las secuencias, y todas las celdas se llenan según un sistema de puntuación (imagen 2), pero la fila y la columna adicionales de la parte superior y lateral izquierdo, se rellenan de ceros. La manera de rellenar la matriz es la misma, con la excepción de que siempre que se obtiene un valor negativo, la celda se rellena en su lugar por un cero. El valor cero, hace que se comience otro alineamiento, es mejor comenzar uno nuevo que seguir con uno en el que no hay coincidencia. Otra diferencia es que las flechas solo se asignan si provienen de un valor positivo. El rastreo también es un poco distinto, se comienza en los valores más altos de dentro de la matriz, siguiendo las flechas hasta que se alcanza un cero (imagen 2). El tiempo que cuesta ejecutar este algoritmo crece linealmente con el producto de la longitud de las dos secuencias comparadas, de este modo, aunque es un método óptimo y relativamente rápido, no es lo suficientemente rápido si necesitamos comparar secuencias muy largas, o si necesitamos

comparar muchas secuencias como en una aplicación de búsqueda de una base de datos, como por ejemplo el GenBank del NCBI [5].

Un ejemplo de herramienta desarrollada con este tipo de alineamiento es LALIGN. Compara dos secuencias, ya sean de aminoácidos o nucleótidos, en busca de similitudes internas calculando alineaciones locales que no sean inserciones [14]. Esta herramienta fue desarrollada por X. Huang y W. Miller en 1991 [14], y se puede utilizar en la web del Instituto Europeo de Bioinformática (EBI) [15].

1.3.2. Alineamiento de secuencias múltiple

El alineamiento de secuencias múltiples se utiliza para extraer la homología y las relaciones evolutivas entre más de dos secuencias, partiendo del supuesto de que existe un ancestro común [16]. Al realizar un alineamiento simultáneo de las secuencias de, por ejemplo, un gen, se pueden estudiar patrones dentro de las secuencias que han estado sujetas a alteraciones debidas a la evolución. Esta técnica nos permite aprender sobre la estructura y función de las moléculas, presentando un gran potencial para la biología molecular. Por ello, se han desarrollado distintos métodos computacionales para llevar a cabo un alineamiento óptimo, aunque no ha sido una tarea fácil. Cuando se plantea el alineamiento, se considera que es óptimo cuando se registra el mayor número de caracteres similares en la misma columna de alineación, como se ha descrito en los alineamientos de pares de secuencias. Sin embargo, computacionalmente el alineamiento múltiple presenta varios desafíos [17]:

1. Encontrar un alineamiento óptimo de más de dos secuencias que tenga en cuenta: *match*, *mismatch* y *gaps*, además del grado de variación en todas las secuencias al mismo tiempo.
2. Identificar un método razonable para obtener una puntuación acumulativa para las sustituciones en una columna de un alineamiento múltiple.
3. Encontrar un método que permita gestionar la ubicación y puntuación de los *gaps* en las diversas secuencias.

En referencia al primer desafío, el algoritmo de programación dinámica utilizado para la alineación óptima de pares de secuencias puede extenderse a tres secuencias, pero para más de tres secuencias, solo puede analizarse un pequeño número de secuencias relativamente cortas. Esto se debe a que el aumento del número de pasos computacionales que hay que realizar y la cantidad de memoria requerida crecen exponencialmente con el número de secuencias a analizar [17]. Como se ha descrito antes, para alinear dos secuencias se crea una matriz de puntuación en la que cada posición proporciona la mejor alineación posible; si tuviese que alinear tres

secuencias con una longitud considerable, la matriz tendría forma de cubo. Por lo tanto, se emplean métodos aproximados para poder alinear más secuencias, que incluyen [17]:

- una alineación global progresiva de las secuencias, comenzando con una alineación de las secuencias más parecidas y luego construyendo una alineación agregando más secuencias;
- métodos iterativos que hacen una alineación inicial de grupos de secuencias y vuelven a realizar el alineamiento para lograr un resultado más razonable;
- alineaciones basadas en patrones conservados localmente que se encuentran en el mismo orden en las secuencias;
- uso de métodos estadísticos y modelos probabilísticos de las secuencias.

Al utilizar métodos aproximados, no es posible demostrar que el alineamiento que se ha obtenido sea el óptimo [18]. Por ello, se han ido desarrollando mejoras; actualmente, se dispone de distintas herramientas de alineación múltiple, que se diferencian en función del volumen de datos que pueden alinear, la precisión y la velocidad:

- **CLUSTAL OMEGA**: la última versión permite alinear cientos de miles de secuencias en unas pocas horas. Emplea diferentes algoritmos consiguiendo una máxima precisión: calcula árboles para utilizarlos como guía y hace uso del modelo oculto de Markov (HMM) [19].
- **T-COFFEE**: a diferencia del resto, es más apropiado para realizar pequeñas alineaciones. T-coffee es un acrónimo de función objetivo de coherencia basada en árboles para la evaluación de la alineación. Primero, se sirve de una librería de alineamientos de pares de secuencias, y después con un método de optimización encuentra el alineamiento múltiple que mejor se adapta a las alineaciones de pares de secuencias que ha encontrado en la librería, para ello utiliza un árbol de guía [20].
- **MAFFT** (Alineamiento múltiple usando la rápida transformada de Fourier): realiza alineaciones de conjuntos de datos medianos a grandes. Se basa en la transformada rápida de Fourier para la detección de regiones homólogas y es uno de los alineamientos múltiples más precisos y rápidos. Hace uso de un método iterativo, divide el problema en sub-alineamientos mediante árboles, después estos sub-alineamientos se vuelven a alinear por grupos. Este proceso se repite una y otra vez con el propósito de mejorar la puntuación final [21].
- Una de las herramientas más utilizadas es **MUSCLE** (Multiple Sequence Comparison by Log- Expectation): que se recomienda para realizar alineamientos con datos de tamaño mediano. Este método es preciso y rápido, permite alinear 1000 secuencias de proteínas, con una longitud media de 282 residuos en 21 segundos en un ordenador de sobremesa. Hace uso de una gran variedad de métodos para calcular el alineamiento: alineamientos

de pares de secuencias para obtener resultados de similitud, calcula de manera iterativa árboles como guía y calcula matrices de sustitución siguiendo distintos modelos evolutivos, entre otros [22].

Una vez que se ha encontrado el alineamiento múltiple, el número o los tipos de cambios en los residuos de las secuencias alineadas pueden usarse para realizar un análisis filogenético. Cada columna de la alineación predice las mutaciones que ocurrieron en ese sitio durante la evolución de la familia de secuencias (imagen 7). Dentro de cada columna, podemos ver caracteres que aparecieron más temprano que otros (en la imagen 7 se muestran en verde, amarillo y morado), produciendo sustituciones. Sin embargo, en otras posiciones que son importantes para la función de la proteína, vemos que no hay cambios de caracteres (en la imagen 7 se muestran en rojo y azul), y son estas posiciones conservadas las más útiles para producir una alineación [17].

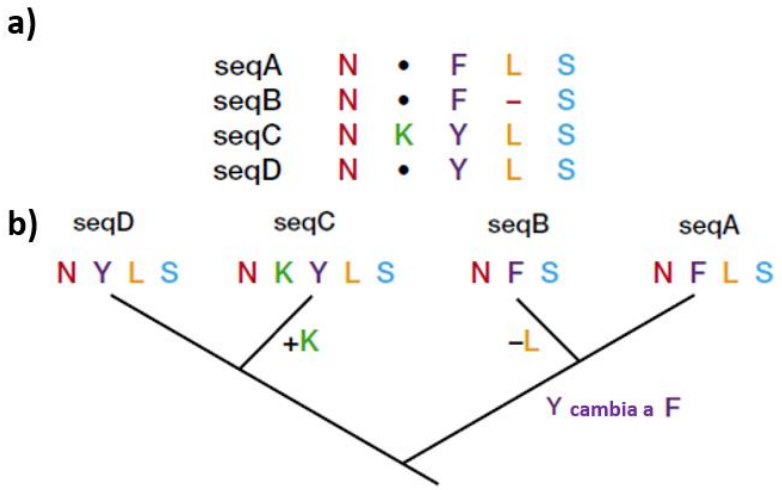


IMAGEN 7. Relación entre el alineamiento de secuencias múltiple y la construcción de los árboles evolutivos. **a)** Se muestra el alineamiento múltiple de cuatro secuencias de proteínas. En la primera (rojo) y en la última columna (azul) de alineamiento aparecen aminoácidos conservados. En la tercera columna (morado) hay una sustitución. En la segunda columna (verde) una inserción de K y en la cuarta (amarillo) una deleción de L; **b)** Hipotético árbol evolutivo que podría haber generado estos cambios en la secuencia, aunque la alineación podría explicarse por otros árboles diferentes. En cada extremo de la rama representa una de las secuencias. La rama más profunda y antigua es la de la secuencia D, seguida por A, luego por B y C. La alineación óptima de varias secuencias puede considerarse como una minimización del número de pasos mutacionales en un árbol. Imagen obtenida del libro *Bioinformatics: sequence and structural analysis* [17].

1.3.3. Alineación de secuencias frente a secuencias almacenadas en bases de datos

Cuando se obtiene una nueva secuencia ya sea de nucleótidos o aminoácidos, lo primero que se suele hacer es buscarla en una base de datos que almacena secuencias que ya han sido caracterizadas y están identificadas. De este modo, se puede comprobar si ya existe, o si existe alguna secuencia similar para poder inferir su función. BLAST es la más importante, se puede utilizar para realizar búsquedas sencillas de manera rápida a través de su web; para búsquedas masivas, se puede instalar localmente.

En 1990, Altschul desarrolló el algoritmo de BLAST, que es de tipo heurístico [23]. Este método identifica regiones de alta similitud antes de calcular la puntuación del alineamiento de pares de secuencias. Los métodos heurísticos, a diferencia de la programación dinámica, no garantizan encontrar la alineación óptima ya que, como su nombre indica, buscan la solución del problema mediante un método no tan riguroso, similar al tanteo. El algoritmo de BLAST es dos órdenes de magnitud más rápido que el algoritmo de Smith y Waterman, esto lo consigue buscando solamente dentro de la región de la secuencia que tiene una alta similitud. BLAST realiza las búsquedas por palabras, k-mers, de una longitud k (k=3 aminoácidos y k=11 para nucleótidos), que existen en la secuencia de consulta [17]. Se enumeran las palabras similares de alta puntuación, según una matriz de sustitución, para cada palabra de la matriz de consulta. Las palabras similares de su «vecindad» se alinean y se ordenan según la puntuación obtenida. Para las sustituciones de aminoácidos, se suele utilizar la matriz BLOSUM62. Todas las palabras encontradas con este procedimiento se guardan en una lista y se buscan coincidencias exactas en las secuencias de la base de datos. Cada coincidencia se denomina «par de secuencias de alta puntuación» (HSP), que se emplea como «semilla» para la alineación de la secuencia local. La alineación se extiende a la izquierda y a la derecha de la semilla, y la puntuación de alineación se calcula después de cada extensión según la matriz de sustitución [5]. Comenzando desde la puntuación máxima encontrada en cualquier punto durante la alineación, el algoritmo deja de extenderse cuando la puntuación disminuye hasta un valor determinado. Se mantiene la puntuación final para cada alineación local y se descartan todas las alineaciones con una puntuación por debajo de un valor umbral. BLAST se desarrolló inicialmente para alineaciones sin espacios *(gap)* [4], pero actualmente está disponible para alineaciones que incluyen espacios [24].

A partir de este método, se han desarrollado variaciones para adaptarse a los distintos tipos de secuencias que se quieren comparar, como: TBLASTN que compara una proteína frente a una base de datos de nucleótidos que se traduce en seis fragmentos de lectura [25].

1.4. Estudio de la evolución

Es posible utilizar alineamientos tanto de ADN como de proteínas para estudiar la evolución, sin embargo, las secuencias de proteínas poseen ventajas particulares sobre el ADN en la identificación de los genes y sus relaciones evolutivas. De este modo, para analizar entidades más divergentes es preferible hacer uso de alineamientos de proteínas, pero si queremos analizar individuos cercanos se obtiene más información estudiando los alineamientos de ADN [16].

La evolución representa cambios graduales en el contenido genético de un organismo durante generaciones sucesivas [26]. Desde que Charles Darwin documentó científicamente el proceso evolutivo y postuló que era resultado de la selección natural [27], su teoría se ha ido ampliando. Se combinó con la genética mendeliana a principios del siglo XX, que condujo al nacimiento de la síntesis evolutiva moderna, que es el concepto central de la actual teoría evolucionista [28]. Aunque se pueden dar numerosas modificaciones en un organismo a lo largo del tiempo, estos cambios puede que no resulten en una evolución, esta solo se producirá si se heredan en las generaciones siguientes a través del material genético, fijándose de esta manera la sustitución. Por todo ello la evolución es un tema candente en el panorama científico. Se utiliza para estudiar diferentes aspectos de los organismos vivos, desde investigar las diferencias en una proteína expresada por distintas especies, hasta los mecanismos de resistencia antibiótica de los microorganismos [16].

Para analizar estas relaciones evolutivas entre diferentes organismos, se hacen estudios comparativos de las secuencias del genoma de estas entidades, que suelen realizarse con alineamientos múltiples. Con este análisis se obtiene una estimación de las diferencias y similitudes entre las secuencias y, con la ayuda de diferentes algoritmos, se puede inferir información evolutiva y presentarla en forma de árbol, denominado árbol filogenético [29].

1.4.1 Características de los árboles filogenéticos

Un árbol filogenético (imagen 8) es una representación arborescente que ilustra las relaciones evolutivas de taxones, situados en los nodos externos o terminales, que pueden ser especies, individuos, genes, etc., que se conectan mediante ramas a los nodos internos. Se denomina clado a las unidades monofiléticas que están compuestas por un ancestro (nodo interno) y todos sus descendientes (imagen 8). De un árbol se puede extraer información de la topología y de la longitud de las ramas.

Originalmente, los árboles filogenéticos se utilizaban con datos morfológicos para describir las relaciones entre las especies que residen en diferentes taxones. Sin embargo, gracias a los avances en la secuenciación de ADN y de proteínas, se realizan árboles para describir relaciones entre diferentes grupos de organismos.

La topología define las relaciones entre los taxones que están representados. En la imagen 8, se puede ver un ejemplo, en este caso los taxones C y D forman un grupo monofilético, que pertenecen a un linaje hermano. El clado formado por estos dos taxones proviene de un antepasado común que se representa con un nodo interno. De este antepasado común también provienen los taxones E y F que forman otro grupo monofilético distinto. De este modo se ilustran las relaciones de los diferentes taxones del A al F. Hay que tener en cuenta que la rotación del eje de los nodos internos no cambia la información topológica [5].

Si el árbol se ha calculado en escala, como en el caso de los **filogramas o árboles ultramétricos**, la longitud de sus ramas puede mostrar el grado de relación de los taxones y nodos internos. Esta longitud es proporcional a la cantidad de cambios evolutivos o divergencia evolutiva, y representa el número de cambios de nucleótidos o aminoácidos que se han producido en esa rama. Cuantos más cambios entre las secuencias, más larga será la rama. Sin embargo, los **cladogramas o dendogramas** no están a escala, solamente nos muestran información topológica, podemos ver los clados que se forman pero la longitud de las ramas no es proporcional.

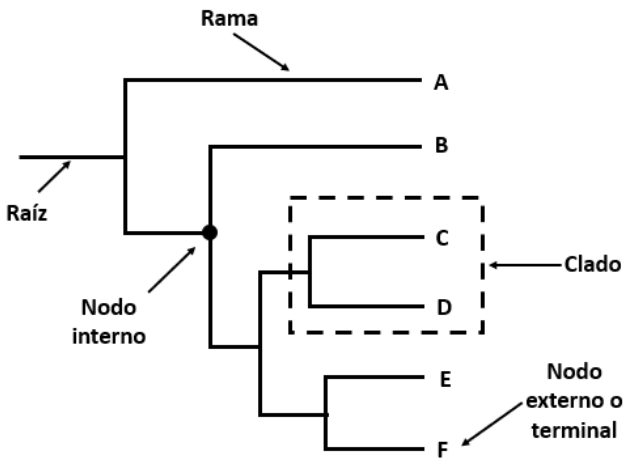


IMAGEN 8. Términos que describen la topología de un árbol filogenético. Imagen modificada obtenida del libro *Phylogenomics: An introduction* [5].

Otra característica de los árboles filogenéticos es que pueden estar enraizados o no. Si no tiene raíz, el camino evolutivo no tiene una dirección, se muestran las posiciones relativas de los taxones, y no se asume que existe un ancestro común. Por otro lado, los enraizados como los

cladogramas y dendogramas tienen un nodo que es la raíz, del que diverge el resto del árbol. Esta raíz es considerada el ancestro común universal. Hay dos maneras de enraizar un árbol:

- Usando un taxón de un grupo externo: una secuencia homóloga de las secuencias que se están analizando, obtenida de una fuente de información independiente. Por ejemplo, utilizar una secuencia de aves (grupo externo) como raíz para el análisis filogenético de mamíferos (grupo interno).
- Enfoque de enraizamiento del punto medio: se asigna como raíz el punto medio de los dos grupos más divergentes considerando todas las longitudes de las ramas. Este tipo de enraizamiento sigue la hipótesis del reloj molecular.

1.4.2. Teorías de la evolución

Aunque probablemente exista un árbol de la vida verdadero que describa con precisión la evolución de todas las especies, ese árbol es imposible de generar. Lo que se construyen son árboles inferidos que hipotetizan las versiones de los eventos que pudieron ocurrir.
Para explicar y calcular la evolución de los genes y proteínas se han descrito dos teorías en los últimos 50 años: la hipótesis del reloj molecular y la teoría neutral de la evolución.

1.4.2.1. Hipótesis del reloj molecular

- Permite traducir el número de sustituciones en una secuencia en términos de tiempo. Cuanto más larga sea la rama del árbol, más tiempo de divergencia se ha dado.
- Sustitución se considera cuando se ha producido una mutación y se ha fijado en la población mediante selección.
- Implica que la tasa de sustitución en las secuencias es constante.
- Es posible modelar cuándo el aminoácido o nucleótido cambió entre un grupo de secuencias.
- Si las secuencias evolucionan a velocidades constantes, pueden usarse para estimar el tiempo en que las secuencias divergieron.

Esta hipótesis puede explicar cómo han evolucionado algunos genes y proteínas, sin embargo, presenta limitaciones:

- No se puede aplicar a todas las proteínas y a todos los genes. Debido a que la tasa de evolución varía entre diferentes organismos y entre distintos genes.
- Solamente es aplicable cuando el gen conserva su función durante el tiempo evolutivo.
- Se puede aplicar la prueba de Tajima para determinar si dos secuencias exhiben tasas evolutivas iguales.

1.4.2.2. Teoría neutral de la evolución

Existe una gran cantidad de polimorfismos distintos de ADN en todas las especies, que es difícil de explicar mediante la selección natural convencional. La mayor parte de los polimorfismos observados en el ADN son neutrales o casi neutrales, y no parece estar bajo selección natural en la mayoría de los casos.

Hasta la década de 1960, el modelo predominante de evolución molecular era que la mayoría de los cambios en los genes se seleccionaban a favor o en contra en un sentido darwiniano. Pero Motoo Kimura (1968, 1983) propuso un modelo diferente para explicar la evolución a nivel del ADN. Kimura señaló que la tasa promedio de sustitución de aminoácidos es aproximadamente un cambio cada $28 \cdot 10^6$ años por proteína de 100 residuos [30]. Además, estimó que la tasa correspondiente de sustitución de nucleótidos debe ser extremadamente alta (un promedio de un par de bases de ADN reemplazado en el genoma de una población cada dos años). La conclusión de Kimura fue que la mayoría de las sustituciones de ADN observadas deben ser neutrales o casi neutrales, y que la principal causa del cambio evolutivo (o variabilidad) a nivel molecular es la deriva aleatoria de alelos mutantes. La mayoría de las mutaciones no sinónimas son perjudiciales y, por tanto, no se observan como sustituciones en la población. Bajo este modelo, llamado teoría neutral de la evolución, la selección darwiniana positiva juega un papel extremadamente limitado.
De hecho, la existencia de un reloj molecular tiene sentido en el contexto de la hipótesis neutral porque la mayoría de las sustituciones de aminoácidos son neutrales. Desde su publicación de 1983, la teoría neutral continúa siendo probada en una variedad de organismos [30].

Bajo el supuesto de una tasa evolutiva constante a lo largo del tiempo, debe esperarse un aumento lineal del número de sustituciones de nucleótidos después de la divergencia de un par de secuencias. Sin embargo, como puede haber sustituciones hacia atrás, sustituciones múltiples o sustituciones convergentes, la comparación de distancias observadas (distancias p) entre pares de secuencias mostrará un nivel de saturación después de algún tiempo de divergencia [31]. Para corregir esta saturación, se emplean modelos probabilísticos de evolución de secuencia para calcular las distancias esperadas. La mayoría de los métodos de reconstrucción filogenética para secuencias de nucleótidos se basan en modelos de evolución de secuencias formulados explícitamente. Estos modelos se incorporan dentro de los métodos de distancia, máxima verosimilitud e inferencia bayesiana.
Los modelos de sustitución de nucleótidos que se utilizan para la inferencia filogenética hacen varias suposiciones para modelar las sustituciones como un proceso estocástico [17]:

I. Para cada sitio de una secuencia, se supone que la tasa de cambio de una base a otra es independiente de la historia de este sitio (propiedad de Markov).

II. Se supone que las tasas de sustitución no cambian con el tiempo (homogeneidad).
III. Se asume el equilibrio de las frecuencias base (estacionariedad).

Los modelos que se ajustan a esta descripción se denominan modelos de Markov estacionarios continuos y homogéneos en el tiempo. Dichos modelos resumen las tasas de sustitución en una matriz de tasas (o matriz Q), donde cada entrada especifica la probabilidad de cualquier posible sustitución de nucleótidos [32]. Por lo general, los modelos utilizados en filogenética molecular son reversibles en el tiempo, por lo que se supone, además, que la tasa de cambio de una base i a otra base j es idéntica a la tasa de cambio de j a i (j e i pueden ser todas las bases posibles, pero deben ser bases diferentes). Estas tasas y probabilidades se calculan mediante los resultados obtenidos mediante un alineamiento múltiple de las secuencias. El modelo más general de sustituciones de nucleótidos es el modelo General Reversible en el Tiempo (GTR) [33], que se resume la matriz Q [5]:

$$Q = \begin{pmatrix} -\mu(\alpha\pi_C + \beta\pi_G + \gamma\pi_C) & \mu\alpha\pi_C & \mu\beta\pi_G & \mu\gamma\pi_T \\ \mu\alpha\pi_A & -\mu(\alpha\pi_A + \delta\pi_G + \epsilon\pi_T) & \mu\delta\pi_G & \mu\epsilon\pi_T \\ \mu\beta\pi_A & \mu\delta\pi_C & -\mu(\beta\pi_A + \delta\pi_C + \eta\pi_T) & \mu\eta\pi_T \\ \mu\gamma\pi_A & \mu\epsilon\pi_C & \mu\eta\pi_G & -\mu(\gamma\pi_A + \epsilon\pi_C + \eta\pi_G) \end{pmatrix}$$

Matriz Q: Esta matriz resume el modelo General Reversible en el Tiempo (GTR).

En la tabla 1, se muestran las características de otros de los modelos de sustitución de nucleótidos que existen. Los parámetros que se utilizan para describir estos modelos son:

- Frecuencias de las bases: son las frecuencias de cada nucleótido en la secuencia. Denominados: π_A, π_C, π_G, π_T.
- Tasa global de sustitución: es el número total de sustituciones en una secuencia por unidad de tiempo. Nombrada como μ.
- Tasa de transición y transversión: por unidad de tiempo, son las probabilidades de que A se sustituya por G, y viceversa, A por C o T; C por A o G; G por A o T; y T por G o A, y viceversa. La tasa de transición y transversión se denomina: κ.
- Tasa de transición (contexto de Markov): son las frecuencias de posibles sustituciones de cada nucleótido, es decir, que un nucleótido «i» pase a ser «j», y viceversa. Son denominados: α, δ, γ, η.

Tabla 1. Modelos de sustitución de nucleótidos.

Modelo	Suposición	Parámetros del modelo
Jukes - Cantor 1969	• Todos los nucleótidos tienen la misma probabilidad de ser sustituidos.	μ
Kimura 1980	• Transiciones ($\beta = \varepsilon$) y transversiones ocurren con diferente probabilidad ($\alpha = \delta = \gamma = \eta$). • Todos los nucleótidos pueden ser sustituidos con la misma probabilidad ($\pi_A = \pi_C = \pi_G = \pi_T = 0{,}25$)	Modelo original α era la tasa de transiciones y β la única transversión, (κ y μ).
Felsestein 1981	Es una extension de Jukes-Cantor 1969: • todos los nucleótidos pueden ser susitutidos con diferentes probabilidades (π_A ! = π_C ! = π_G ! = π_T , con valores entre 0 y 1) • Transiciones y transversiones ocurren con la misma probabilidad ($\alpha = \beta = \delta = \gamma = \varepsilon = \eta$ =1).	μ, π_A, π_C, π_G, π_T
Tamura 1992	Es una extension de Kimura 1980: • Las transiciones ($\beta = \varepsilon$) y transversiones ocurren con diferente probabilidad ($\alpha = \delta = \gamma = \eta$). • Los nucleótidos son sustituidos con una probabilidad ajustada al contenido de GC en la secuencia de DNA ($\pi_C = \pi_G = \pi_{GC} / 2$; $\pi_A = \pi_T = (1 - \pi_{GC})/2$).	μ, π_{GC}
Hasegawa-Kishano-Yano 1985	Es una combinación de las extensions de Kimura 1980 y Felsestein 1981: • Transiciones ($\beta = \varepsilon$)y transversiones ocurren con diferente probabilidad ($\alpha = \delta = \gamma = \eta$). • Los nucleótidos pueden ser sustituidos con diferente probabilidad (π_A ! = π_C ! = π_G ! = π_T, con valores entre 0 y 1).	μ, κ, π_A, π_C, π_G, π_T
Tamura - Nei 1993	• La probabilidad de transición A<->G (β) es diferente de la probabilidad de transición C<->T (ε), y también la transversión ocurre con diferente probabilidad, pero todos los tipos de transversiones ocurren con la misma ($\alpha = \delta = \gamma = \eta$). • Los nucleótidos pueden ser sustituidos con diferente probabilidad (π_A ! = π_C ! = π_G ! = π_T, con valores entre 0 y 1).	μ, κ_1, κ_2, π_A, π_C, π_G, π_T
Generalised time-reversible 1986	• Es el modelo más complejo. • Los nucleótidos pueden ser sustituidos con diferente probabilidad (π_A ! = π_C ! = π_G ! = π_T, con valores entre 0 y 1). • La probabilidad de las transiciones y transversiones es distinta en cada combinación (α ! = β ! = δ ! = γ != ε != η).	α, β, δ, γ, ε, η, π_A, π_C, π_G, π_T

Estos modelos están relacionados unos con otros, ya que añaden o quitan parámetros. Todos ellos asumen que la tasa de evolución es la misma para cada posición del alineamiento de las secuencias. Aunque la verdad es que hay variación entre las posiciones, o sitios. Esta propiedad es denominada como heterogeneidad. Por ejemplo, en los siguientes casos sería necesario aplicar esta propiedad:

- Las tasas de sustitución difieren para distintas posiciones de codón: la tercera posición en un codón muta mucho más rápido que las otras dos.
- Algunas regiones de una proteína tienen dominios conservados, la tasa de sustitución de sus correspondientes nucleótidos codificantes sería baja.
- El ARN no codificante, como el ARNr, a menudo tiene limitaciones funcionales con posiciones muy conservadas, por lo que la tasa de sustitución sería inferior.

La manera de incluir esta heterogeneidad es a través del modelo gamma, Γ, que se trata de un tipo de distribución estadística que se define mediante el parámetro α; α debe determinarse para ajustarse al modelo gamma para un conjunto de datos dado. Por lo general, α es bastante

pequeño (<1) [34], lo que da como resultado una distribución L sesgada, lo que refleja que la mayoría de los sitios muestran tasas de sustitución bajas o son invariables, pero algunos de ellos tienen una tasa de sustitución muy alta [35]. Los valores altos de α (>1) darían como resultado una distribución en forma de campana donde la mayoría de los sitios evolucionan a un ritmo similar, presentan tasas de sustitución intermedias, y unos pocos sitios son los que tienen tasas de sustitución más altas [34]. Los modelos de evolución de secuencia, como los modelos de sustitución de nucleótidos, que incorporan la distribución gamma, están marcados con «+ Γ» o «+ G». La inclusión de una distribución gamma requiere mucho tiempo y memoria computacionalmente, lo que puede ser un problema para los análisis filogenéticos a gran escala [5].

Otra modificación para tener en cuenta la heterogeneidad es la incorporación de la proporción de sitios invariantes en modelos de evolución de secuencias [36]. Los modelos que utilizan esta modificación están marcados con un «+ I». Si todos los sitios de alineación cambiaran a la misma velocidad, como suponen todos los modelos discutidos aquí, el número de sustituciones debería seguir una distribución de Poisson. Los conjuntos de datos reales generalmente no se ajustan a esta distribución. Sin embargo, la exclusión de sitios invariantes permite un mejor ajuste. Los modelos que incluyen ambas modificaciones (+ I + Γ) asumen que una proporción de sitios es invariable, mientras que las tasas de los sitios restantes tienen distribución gamma [37]. Se ha discutido si este tipo de modelos deberían usarse juntos, dado que la cantidad de sitios invariables ya está incluida en la distribución gamma [38]. Sin embargo, los estudios de simulación y las comparaciones de conjuntos de datos reales encontraron que los modelos que incluyen ambos parámetros (+ I + Γ) a menudo mejoran los análisis filogenéticos [39,40].

También, existen modelos de sustitución de aminoácidos, en este caso hay que distinguir entre los empíricos y los mecanicistas [35]. El tipo de modelos empírico utiliza matrices de sustitución calculadas a partir de la comparación de secuencias. Las primeras matrices fueron publicadas por Margaret Dayhoff y su equipo [6]. Estas matrices son la PAM1 y la PAM250 de las que se ha hablado anteriormente. Años más tarde, Jones-Taylor-Thornton 1992 (JTT) publicó una actualización de las matrices basada en una base de datos de secuencias mucho más grande [41]. Ya en los años 2000, se desarrollaron dos enfoques nuevos que evitan problemas de las anteriores matrices, se trata de las matrices de Wheland y Goldman 2001 (WAG) [42], y por último la de Le Gascuel 2008 (LG) [43].
Los modelos mecanicistas incluyen supuestos sobre procesos biológicos como, por ejemplo, clasificar los aminoácidos en clases según sus propiedades químicas, o son calculados a nivel de codones. En especial, se ha demostrado que los modelos de codones superan a los modelos empíricos, pero son computacionalmente mucho más costosos de calcular [44].

Se utiliza una versión simplificada del modelo de codones propuesto por Goldman y Yang (1994) [45]. En este modelo, los parámetros se estiman para las comparaciones de pares de codones. Se aplica una tasa de 0 para codones que difieren en dos o tres posiciones. Se estiman tasas separadas para codones que difieren en una sola posición [45]. En este caso, se estiman diferentes tasas para las transversiones sinónimas y no sinónimas, así como para las transiciones sinónimas y no sinónimas [46].

Una vez vistos todos los tipos de modelos que hay, una de las preguntas más frecuentes es cuál de todos los modelos es más apropiado y si deben tener en cuenta la heterogeneidad de la tasa (+ Γ) y los sitios invariantes (+ I), así como la frecuencia de los aminoácidos (+ F). Obviamente, se necesitan métodos para seleccionar un modelo que se ajuste a los datos, mientras se intenta evitar la parametrización excesiva. Los métodos más empleados para elegir entre modelos son la prueba de razón de verosimilitud jerárquica (hLRT), el criterio de información bayesiano (BIC) y el criterio de información de Akaike (AIC) [5].

1.4.3. Métodos para la reconstrucción filogenética

Existen distintos métodos para la reconstrucción filogenética, los cuatro más importantes son: utilizando distancias por pares entre secuencias, el método de unión de vecinos (NJ), o basándose en caracteres discretos, los métodos de máxima parsimonia (MP), máxima verosimilitud (ML) e inferencia bayesiana (BI). Se han propuesto otros métodos, por ejemplo: UPGMA, evolución mínima, pero ya no se usan en la filogenia molecular moderna. Por lo general, con distancias, el mejor árbol se reconstruye por agrupamiento, mientras que los métodos basados en caracteres aplican un criterio de optimización para elegir el mejor árbol entre todas las topologías de árboles posibles [29]. Históricamente, los primeros análisis de secuencias a menudo se basaban en distancias [36]. Sin embargo, hoy en día, son los métodos basados en caracteres los que más se utilizan.

Inferir árboles por **Unión de Vecinos** (NJ) consta de dos pasos: construcción de una matriz de distancias por pares que se emplea para un agrupamiento posterior de un árbol utilizando el algoritmo NJ, el algoritmo elige el árbol con la menor suma de longitudes de rama [47]. Por lo general, las distancias entre secuencias se calculan considerando un modelo evolutivo de los descritos anteriormente. Esta matriz se agrupa en un árbol mediante el algoritmo NJ, que hace uso de la descomposición en estrella. El algoritmo comienza con un árbol de estrellas sin resolver, y se une sucesivamente a un par de terminales basándose en la matriz de distancias hasta que el árbol se resuelve por completo. De manera iterativa, los terminales se eligen de manera que se minimice la longitud total de la rama del árbol. NJ es computacionalmente muy rápido, ya que

el tiempo para analizar grandes conjuntos de datos aún se puede medir en milisegundos. Sin embargo, se ha demostrado que los métodos basados en distancias, en general, son propensos a problemas con errores sistemáticos [48]. Aunque este método se implementa a menudo cuando se necesita un árbol rápido, por ejemplo, árboles para guiar alineaciones, o árboles iniciales para búsquedas heurísticas de métodos basados en caracteres.

Máxima Parsimonia (MP) es un método de inferencia filogenético que emplea un criterio de optimalidad para decidir qué árboles son los mejores entre todos los árboles posibles. Como el número de árboles posibles para un mayor número de secuencias analizadas es demasiado grande para ser analizado exhaustivamente, se utilizan métodos heurísticos para reducir el espacio de los árboles buscados. La idea que se sigue es que la mejor hipótesis para explicar una observación es la que requiere la menor cantidad de supuestos [49], «la navaja de Ockham». Hoy en día, hay distintas variantes de MP en uso que, por ejemplo, difieren en la forma en que las transformaciones de caracteres se ponderan u ordenan [50]. Solo los caracteres que producen diferentes números de pasos en las topologías se consideran informativos, mientras que todos los demás caracteres se excluyen del análisis. Los caracteres informativos son aquellos que tienen, como mínimo, dos estados de carácter diferentes, que aparecen, al menos, en dos terminales cada uno. MP es un método fácil de entender y, debido a su simplicidad, se dispone de algoritmos de análisis eficientes y rápidos [29]. Sin embargo, la falta de un uso explícito de modelos evolutivos es un gran inconveniente de este método. La mayoría de los estudios de simulación muestran que los enfoques basados en modelos basados en inferencias de ML superan a la MP en la reconstrucción filogenética molecular [51]. Sin embargo, los métodos de MP se utilizan con frecuencia para la reconstrucción filogenética de patrones de ausencia / presencia de caracteres, a nivel del genoma, retrotransposones o microARN [5].

La función de **Máxima Verosimilitud** (ML) se define como la probabilidad de observar los datos dados unos parámetros. Fue desarrollada originalmente por el estadístico R. A. Fisher en la década de 1920. En un contexto filogenético, una topología de árbol representa un modelo, mientras que la longitud de la rama de esta topología y los parámetros de sustitución son parámetros de este modelo [29]. En un análisis de ML, se busca la topología del árbol y su conjunto de longitudes de rama. La topología que produjo el mejor valor de probabilidad se elige finalmente mediante el criterio de optimalidad. Los análisis de ML son el estado del arte de la filogenética, y la mayoría de las publicaciones en este campo utilizan este enfoque. A principios de la década de 2000, el uso de ML seguía siendo a menudo computacionalmente difícil. Sin embargo, con las mejoras de la tecnología informática (la disponibilidad de clústeres informáticos de alto rendimiento y, especialmente, el software que aprovecha este desarrollo), los análisis de ML se volvieron factibles incluso para conjuntos de datos muy grandes. Algunos de los programas tienen la limitación de que para los análisis de nucleótidos solo se puede elegir el modelo GTR.

Mientras que la probabilidad describe la posibilidad de observar los datos dada una hipótesis, mediante el uso de la **Inferencia Bayesiana (BI)**, se describe la probabilidad de la hipótesis dada la información. Es decir, se aplica el teorema de Bayes en un contexto filogenético. Para BI, se deben distinguir las probabilidades previas y las probabilidades posteriores. Las probabilidades previas son suposiciones hechas antes de los análisis de BI. Estas probabilidades previas se actualizan después de acuerdo con los datos analizados, y las probabilidades posteriores son el resultado de BI. Resolver analíticamente el teorema de Bayes es computacionalmente demasiado intensivo. Sin embargo, una aproximación de probabilidades posteriores mediante el uso de un enfoque de Monte Carlo de cadena de Markov (MCMC) hizo factible la BI de filogenias [52]. Al usar una cadena de Markov, se genera una serie de variables aleatorias y la distribución de probabilidad de los estados futuros solo depende del estado actual en cualquier punto de la cadena. Para la inferencia de filogenias, la cadena de Markov comienza con un árbol generado aleatoriamente que incluye las longitudes de las ramas. El siguiente paso en la cadena es generar un nuevo árbol, que se basa en el árbol anterior. A esto se le llama propuesta. El nuevo árbol propuesto se acepta dada una probabilidad específica basada en el algoritmo Metropolis-Hastings [53]. Si se acepta el árbol propuesto, se convertirá en el nuevo estado actual para proponer el siguiente paso en la cadena. Si se rechaza el árbol propuesto, el árbol actual permanece y se debe proponer un árbol nuevo para el siguiente paso. Un software ampliamente utilizado para BI de filogenias es MRBAYES [54]. Para análisis estándar con el objetivo de obtener una topología de árbol, ML parece ser la mejor alternativa, ya que computacionalmente suele ser más rápido, y los resultados suelen ser similares [1].

1.4.4. Evaluación de la fiabilidad del árbol

Los análisis filogenéticos siempre darán como resultado una topología de árbol, lo que plantea cuánta confianza se puede depositar en él. La medida más común de apoyo a las filogenias se deriva de los análisis ***bootstrap***. El *bootstrap* es una técnica de remuestreo comúnmente usada en estadística para valorar la variabilidad de una estimación [55], que fue introducida por Felsenstein (1985) en la filogenia [56].

Para realizar este análisis, el conjunto de datos original debe volver a muestrearse con reemplazo. Hablamos de las denominadas pseudorreplicaciones que contienen el mismo número de sitios de alineación que la alineación original. Cada sitio en estos pseudorreplicados está lleno de sitios de la alineación original. Como este muestreo se realiza con reemplazo, los pseudorreplicados pueden incluir algunos sitios originales varias veces, y otros podrían faltar. Normalmente, se generan 100 o 1000 pseudorreplicados, que luego se analizan como en el análisis filogenético original (por ejemplo, con ML, MP o NJ). Un algoritmo de *bootstopping* puede estimar el número de réplicas necesarias [57]. Todos los árboles resultantes de estos análisis se resumen como un

árbol de consenso basado en la regla de la mayoría, y las frecuencias se muestran en los nodos. Si se encuentra una rama en todas las réplicas, el soporte es del 100%. En estadística, estos valores se interpretan de la manera típica, donde los valores > 95% son estadísticamente significativos. Los análisis de *bootstrap* requieren mucho tiempo computacional, y se han publicado varios enfoques que pueden aproximar rápidamente los valores de *bootstrap* para grandes conjuntos de datos [1].

A pesar de toda esta complejidad para elaborar e interpretar los análisis filogenéticos, tienen una importante y amplia variedad de aplicaciones:

- Ayudan a clasificar las nuevas especies y proporcionan un gran apoyo para definir las categorías taxonómicas.
- Determinación de la paternidad forense, identificación de personas con delitos penales.
- Epidemiología.
- Políticas de conservación de especies en peligro de extinción.
- Estudio de los roles que desempeñan dominios conservados de secuencias en la función de las proteínas.

2. Antecedentes

Una de las aplicaciones de estas herramientas bioinformáticas es la epidemiología y el estudio de la evolución de microorganismos patógenos. En especial, el estudio de bacterias y virus, con el fin de realizar un seguimiento y vigilancia, trazar el origen rutas de transmisión, la introducción y expansión de genes y variantes de resistencia a fármacos, etc. En el ámbito veterinario y específicamente en el sector porcino, el uso de estas herramientas sería de gran utilidad para dos agentes virales muy relevantes en esta especie animal: el virus de Influenza A o gripe porcina, y el virus del síndrome respiratorio y reproductivo porcino.

2.1. Influenza A

El virus de la Influenza A es un patógeno zoonótico del género *influenzavius A*, al que se está prestando mucha atención dada su relevancia en la salud pública. Este virus puede infectar a humanos, aves, cerdos y muchas otras especies animales. En los porcinos esta enfermedad cobra gran relevancia, debido a la susceptibilidad que tienen los cerdos a la infección con virus de influenza A de distintas especies animales (aviar, porcino o humano). Esto facilita el surgimiento de nuevas variantes del virus potencialmente pandémicas en los cerdos y que se pueden extender a los humanos. Además, en el caso de los cerdos, la Influenza A es uno de los patógenos respiratorios más importantes pues es causa de fiebre, pérdida de peso, tos, abortos, entre otros síntomas. Tiene una alta tasa de contagio, pero la mortalidad es muy baja. Por todo ello, la gripe porcina causa pérdidas económicas importantes a través de la disminución de la producción, el aumento de los costes por la vacunación y tratamiento, e incluso el aumento de la mortalidad por co-infecciones bacterianas o víricas [58].

La gran diversidad genética que se ha observado en el virus de Influenza A ha llevado a buscar formas de clasificarlo y una de ellas es la determinación del subtipo viral que se realiza mediante la identificación serológica o molecular de dos proteínas de superficie llamadas hemaglutinina (HA) y neuraminidasa (NA). Existen distintos subtipos de cada una de estas proteínas y varían incluso dependiendo de la especie animal. Los tres subtipos más comunes que afectan al cerdo son: H1N1, H1N2 y H3N2 [59]. Hay, además, diferentes cepas o linajes dentro de cada subtipo, con diferente patogenicidad (capacidad de producir enfermedad) y sin inmunidad cruzada [60].

El primer subtipo endémico de influenza A porcino se originó a partir de la pandemia de gripe española de 1918, que dio lugar a los virus actualmente clasificados como H1N1 «porcino clásico», que se puede observar en la imagen 9 en color morado [61]. Además, con el paso del tiempo, se han ido originando otros cuatro subtipos que han co-circulado en varios países:

- H1N1 porcino de tipo aviar, también denominado, porcino euroasiático (se muestra en color azul en la imagen 9): este linaje de tipo aviar "H1avN1" resultó en 1979 de la propagación de virus H1 de aves silvestres en Europa, que posteriormente se exportó a Asia [58].

- H3N2 porcino recombinante tipo humano (se muestra en color verde en la imagen 9): un virus de la influenza H3N2 de origen humano se introdujo en los cerdos europeos poco después de la pandemia de influenza humana de Hong-Kong en 1968, pero solo se generalizó después de que se recombinase con el virus H1N1 aviar porcino (porcino euroasiático) a principios de la década de 1980, adquiriendo seis nuevos genes [58]. Este H3N2 porcino recombinante similar al humano (H3N2) es ahora el genotipo dominante del virus H3N2 para porcino en Europa.

- H1N2 porcino recombinante similar al humano (se muestra en color naranja en la imagen 9): proviene de un reordenamiento del H3N2 porcino con el gen HA de un virus H1N1 humano estacional dando origen al linaje de tipo humano "H1huN2" [62].

- El virus pandémico H1N1 (H1N1pdm) de origen porcino es el más reciente (se muestra en color naranja en la imagen 9); y su aparición en 2009 alteró la su situación bastante estable en las piaras de cerdos europeos que se observaba con los anteriores subtipos. El H1N1pdm parece haberse generado a partir de múltiples recombinaciones y está constituido por genes de origen porcino, humano y aviar [61]. El virus H1N1pdm se detectó en humanos varias semanas antes de que se informara del primer caso de zoonosis inversa en una piara de cerdos canadiense [61]. Se demostró que los cerdos eran muy susceptibles a este virus humano de origen porcino virus que provocó rápidamente numerosos brotes en todo el mundo, incluida Europa [61]Posteriormente, se demostró la adaptación de H1N1pdm a los cerdos y la propagación en curso en estos animales.

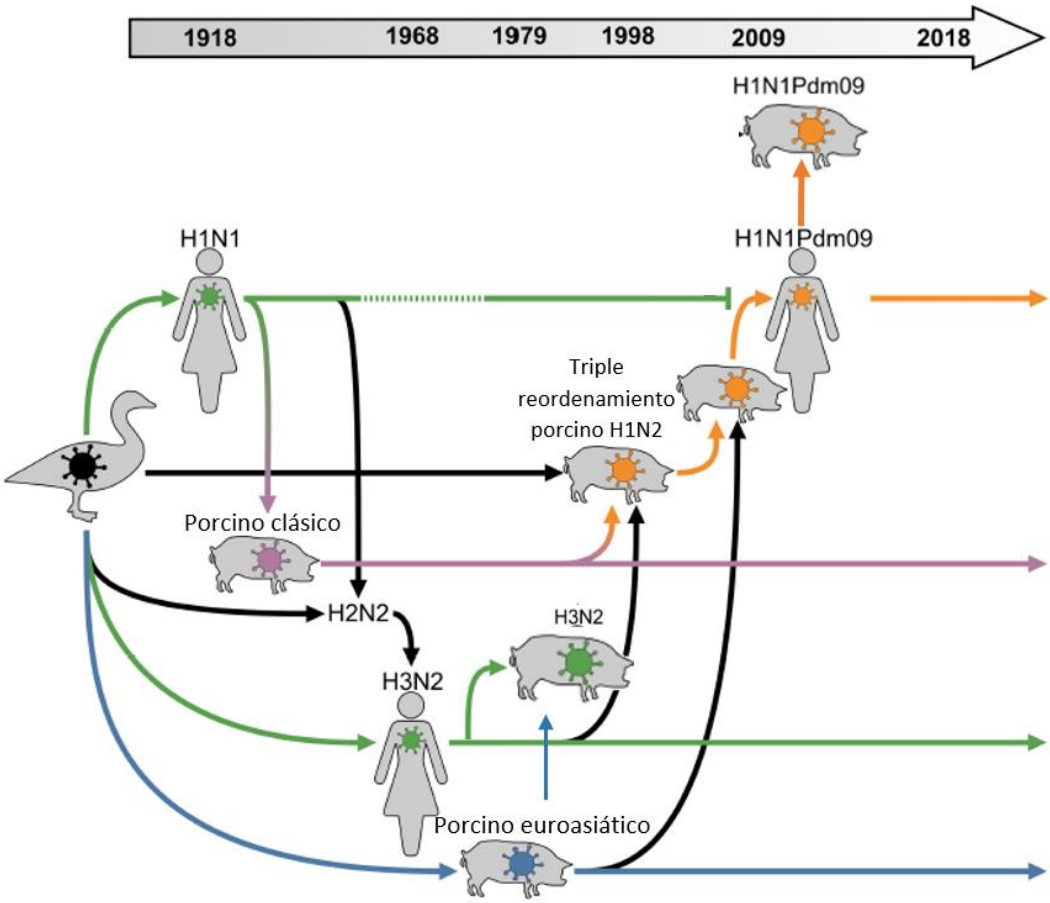


IMAGEN 9. Relación genética de los linajes de virus humanos y porcinos. Las flechas verdes, naranjas, violetas y azules siguen la evolución de la hemaglutinina, mientras que las negras marcan otros segmentos. Los linajes humanos estacionales son verdes, el linaje porcino pandémico H1N1pdm09 es naranja, el linaje porcino clásico es morado y el linaje porcino euroasiático es azul. En 2009, el virus de la influenza A / H1N1Pdm09 desplazó al anterior virus A / H1N1 estacional de los humanos, que continúa circulando en los cerdos. Imagen basada de la obtenida en el artículo *Influenza A Virus Field Surveillance at a Swine-Human Interface* de Benjamin L. Rambo-Martin [61].

Dentro de estos cuatro subtipos principales, numerosos clados genéticos de HA han evolucionado dentro de regiones geográficas específicas. Por lo tanto, la comprensión de los patrones de la diversidad genética de la influenza porcina permite la identificación de nuevas cepas o linajes virales, proporcionando criterios para la intervención racional en el sector porcino y facilitando la preparación para una pandemia de salud pública.

En el artículo *Diversity of influenza A viruses retrieved from respiratory disease outbreaks and subclinically infected herds in Spain (2017-2019)* de Sosa Portugal, se muestran los resultados de su estudio de la diversidad de Influenza A que circula por las granjas de la Península Ibérica [59]. Estos investigadores utilizaron muestras de granjas con brotes de enfermedad respiratoria, y también de granjas seleccionadas al azar sin patología respiratoria manifiesta, para realizar un análisis filogenético de los virus que identificaron. Los genotipos más frecuentes fueron H1avN2hu (av: aviar, y hu: humano, hacen referencia a la procedencia de la variante) y H1avN1av, aunque

se observó una gran diversidad en el análisis filogenético; incluso se encontraron otros subtipos poco comunes en los cerdos. En general, este estudio indica que el virus Influenza A es un agente etiológico muy común en los brotes de enfermedades respiratorias en las granjas porcinas españolas. La diversidad genética de este virus se expande continuamente con cambios claros en los subtipos y linajes predominantes en períodos de tiempo relativamente cortos. El esquema de genotipado actual debe ampliarse o incluir los nuevos genotipos que podrían encontrarse en el futuro.

La infección por el virus podría presentar diferentes patrones, desde un brote epidémico hasta una forma endémica con diferentes oleadas de infecciones con menor incidencia [63], según el estudio *Swine influenza virus infection dynamics in two pig farms; results of a longitudinal assessment.*

Todo esto recalca la importancia del control y seguimiento de los subtipos del virus de Influenza A, por el bienestar animal, por las pérdidas económicas que supone y por el elevado riesgo de desencadenar una pandemia con graves consecuencias para los humanos. Por ello la comunidad científica de todo el mundo está pendiente de la evolución de Influenza A, e incluso han desarrollado una herramienta online, *Influenza Research Database* [64], para facilitar el estudio genético de este virus. También, los laboratorios de diagnóstico veterinario juegan un papel clave, ya que, pueden realizar junto a sus clientes, un seguimiento más local de la evolución de este virus en las granjas.

2.2. Síndrome reproductivo y respiratorio porcino

El síndrome reproductivo y respiratorio porcino (PRRS) está causado por un virus RNA del género *arterivirus*. Este virus se clasifica en dos genotipos: norteamericano y europeo (o Lelystad). Es muy susceptible al ambiente y se caracteriza por tener una alta tasa de mutación que se traduce en una gran diversidad genética.

El virus del PRRS tiene como células diana a los macrófagos alveolares del porcino. Los macrófagos son células inmunitarias que ingieren y eliminan virus y bacterias. El virus es ingerido por los macrófagos alveolares que se encuentran en el pulmón. Dentro del macrófago, el virus se multiplica, provocando que la célula muera en el proceso de replicación. Es capaz de destruir hasta un 40% de los macrófagos del animal, reduciendo considerablemente su defensa inmunitaria, lo que conlleva a que otras bacterias y virus proliferen y agraven la infección respiratoria [65].

El virus, que puede afectar a animales de todas las edades, presenta dos formas clínicas distintas: la reproductiva y la respiratoria. La reproductiva afecta a cerdas reproductoras, verracos y

animales jóvenes; presenta un desarrollo epidémico con una elevada respuesta inmune. La forma respiratoria afecta a todas las edades y tiene un patrón endémico, con una escasa respuesta inmune, junto con una gran variabilidad en la gravedad de los síntomas clínicos. Produce entre otros síntomas: abortos, signos respiratorios o diarreas. Aunque es muy infeccioso, no es muy contagioso y suelen pasar varios meses, incluso casi un año, hasta que se da la primera infección en todos los animales de la granja. También influye en la propagación del virus la edad de los animales: los adultos excretan el virus durante 14 días; sin embargo, los cerdos en crecimiento lo excretan durante 1-2 meses, incluso más de 5 meses en algunos animales. El gran problema del virus del PRRS es que, tras finalizar el brote, el virus persiste en la granja y, tiempo después, vuelve a aparecer otro brote.

Muchas granjas modernas disponen de un sistema de bioseguridad para la prevención del PRRS y, así, evitar las pérdidas económicas que supone, ya que la situación clínica puede variar mucho de una granja a otra, debido a que hay una enorme cantidad de cepas diferentes [65,66]. Aunque se secuencie el virus, es imposible predecir la virulencia, presentación clínica o inmunidad cruzada entre cada cepa identificada.

Los genotipos virales descritos, el americano y el europeo [67] presentan una similitud genética entre ambos de aproximadamente el 65%, y existe además gran diversidad dentro de cada uno de ellos. Algunas investigaciones han demostrado que incluso diferentes grupos de cepas europeas pueden coexistir en el mismo rebaño [68]. Estos hechos son de gran relevancia para el diseño de vacunas, ya que estas variaciones pueden afectar potencialmente su eficacia en condiciones de campo.

El virus consta de un genoma de aproximadamente 15.000 bases que se dividen en 8 ORFs (*open reading frames,* marcos abiertos de lectura). Aunque es posible realizar secuenciación del genoma completo, el método más utilizado para realizar análisis filogenéticos y estudios de similitud es el gen del ORF 5, que codifica la glicoproteína de la envoltura del virus (GP5). Esta proteína está expuesta en la parte externa del virión, así que está sometida a una presión selectiva ejercida por anticuerpos de los animales infectados o vacunados [69]. Esto produce que el gen sea polimórfico, lo que es útil para estudiar la diversidad y la evolución genética del virus, además de ser una posible diana para el diseño de vacunas [69].

En el estudio de 2003, *Genetic diversity and phylogenetic analysis of glycoprotein 5 of European-type porcine reproductive and respiratory virus strains in Spain,* se analizaron secuencias de ORF5 obtenidas de aislamientos de muestras clínicas procedentes de España [70]. Al analizar estas secuencias, encontraron que en la mayoría de los casos la similitud con el virus Lelystad (prototipo de PRRS europeo) fue inferior al 90%. Tras el análisis de las secuencias de proteínas predichas,

algunos aislamientos mostraron, además, un bajo grado de similitud con el virus Lelystad, por debajo del 50%. Estos resultados evidencian la existencia de cepas variantes del virus PRRS en España, con características que podrían ser ventajosas para la evasión de la respuesta inmune [70].

Cinco años más tarde, en el estudio *Influence of time on the genetic heterogeneity of Spanish porcine reproductive and respiratory syndrome virus isolates,* se mostró que no existe una correlación estricta entre la zona geográfica o el tiempo de aislamiento y el grado de parentesco de un grupo de cepas de PRRS de España [69]. También, vieron que la diversidad de los aislamientos parecía estar aumentando con el tiempo, comprometiendo así la efectividad de las vacunas y la inmunidad cruzada, elementos clave para el control de esta infección [69].

En este año 2021, el estudio *Effect of PRRSV stability on productive parameters in breeding herds of a swine large integrated group in Spain,* obtuvo, por primera vez, resultados que demuestran la mejora de la producción debido al logro de la estabilidad del virus PRRS en las granjas de recría [66]. Tras mantener las medidas de control, lograron una estabilidad en la granja durante un año, lo que se tradujo en un aumento de hasta 1,28 lechones destetados por cerda y año [66]. La monitorización diagnóstica del virus fue la principal medida de control y fue realizada mediante la técnica de reacción en cadena de la polimerasa (PCR) en tiempo real. Las muestras positivas se secuenciaron y se realizó un estudio de la diversidad genética del gen ORF 5, diferenciando también las cepas que están presentes en las vacunas [66].

Tanto la alta diversidad genética del virus, como la posible mejora del rendimiento de la granja mediante el control rutinario de la infección viral, sustentan la necesidad de realizar un buen diagnóstico y un estudio filogenético; con el fin de comparar el grado de similitud de las secuencias obtenidas de PRRSV, en cada nuevo caso, con secuencias de casos anteriores, cepas vacunales, y de otras procedencias. Estas medidas pueden ayudar al veterinario a diferenciar las cepas vacunales de las de campo, a valorar la gravedad de la situación y la necesidad de tomar medidas de control más exhaustivas, así como a decidir el inicio de un nuevo plan vacunal o un plan de erradicación.

3. Objetivos

Los objetivos de este proyecto se han basado en las necesidades de Exopol, S. L., laboratorio de diagnóstico veterinario:

- Estudio filogenético con una herramienta de referencia de las secuencias de Influenza A porcino de los subtipos H1 y H3 del gen HA, y los subtipos N1 y N2 del gen NA, para conocer el modelo de evolución y el método más apropiado para hacer árboles filogenéticos.
- Implementación de código mediante lenguaje R del modelo más óptimo para los subtipos H1 y H3 del gen HA, y los subtipos N1 y N2 del gen NA de Influenza A porcino, para realizar árboles filogenéticos con las secuencias obtenidas de las muestras clínicas que llegan a Exopol.
- Selección de los parámetros para el estudio de la identidad de secuencias del gen ORF 5 de PRRS.
- Implementación de código mediante lenguaje R para realizar alineamientos de pares de secuencias del gen ORF 5 de PRRS tanto de ADN como de proteínas adaptado a las necesidades de Exopol.

4. Metodología

El proyecto se desarrolló en el laboratorio de diagnóstico veterinario Exopol, S. L., en el área de Investigación y Desarrollo del Departamento de Biología Molecular. Con el fin de alcanzar los objetivos anteriormente detallados, la metodología se divide en:

4.1. Estudio filogenético de las secuencias de Influenza A porcino

Se manejaron todas las secuencias de ADN obtenidas de muestras clínicas de porcino provenientes de España recibidas en Exopol. Las secuencias son del virus influenza A porcino de los genes HA y NA, en concreto, de los subtipos H1 y H3 del gen HA, y los subtipos N1 y N2 del gen NA. La mayor parte de las secuencias no se obtuvieron completas, así que se manejaron las secuencias más largas que abarcan, sobre todo, la segunda mitad del gen. Se incluyeron secuencias de referencia obtenidas del estudio realizado por Sosa (2020) [59] que están disponibles en Genbank [71]. Se puede consultar, en el anexo 1, el listado de las secuencias. En la tabla 2, se puede ver el número de secuencias estudiadas para cada variante.

Tabla 2. Número de secuencias utilizadas para el estudio de cada subtipo de HA y NA, y la procedencia de ellas.

	Exopol	**Genbank**
H1	21	19
H3	12	11
N1	12	14
N2	24	12

Para el análisis filogenético, se empleó como herramienta de referencia el software MEGA-X versión 10.2.2, desarrollado por S. Tamura, G. Stecher y S. Kumar en 1993 [72]. Para la implementación de las mejoras en la realización de los árboles en Exopol, se utilizó R Studio (Versión 1.4.1103, 2009-2021, para Windows) mediante lenguaje de programación R. Las librerías que utilizadas fueron: ape [73], msa [74], stats [75], utils [76], spider [77], seqinr [78], phytools [79]. Se ha manejado, como base, el código elaborado por Iñaki Albizu (informático de Exopol) para realizar árboles filogenéticos y el sistema de comunicación con *FileMaker* (aplicación multiplataforma de bases de datos relacionales).

Se realizó con MEGA-X un estudio filogenético de cada subtipo de secuencias, en base al resultado obtenido se implementó en R el modelo de evolución y el método para elaborar árboles más

eficientes computacionalmente. Mediante la comparación de los árboles obtenidos con R y MEGA-X, teniendo en cuenta los valores de *bootstrap,* se valoró si los árboles eran equivalentes.

Los pasos para realizar el análisis filogenético son los siguientes, también se puede ver un resumen en el anexo 2:

1.º Elegir el marcador molecular: se eligió ADN porque las secuencias son muy similares, o de genes altamente conservados entre especies.

2.º Realizar el alineamiento múltiple de las secuencias: es un paso crítico, el alineamiento tiene que ser correcto para que se pueda inferir un árbol filogenético correcto. Se realizó con Clustal Omega y MUSCLE, ya que ambos están disponibles en MEGA-X y en R.

3.º Elección del modelo de evolución: el resultado del alineamiento elaborado con MEGA-X lo utiliza el mismo software para realizar un análisis de qué modelo de evolución es más apropiado. Se obtuvo un listado de los modelos de evolución ordenados por los valores AIC y BIC. El que tiene menor puntación de AIC y BIC es el más apropiado para esas secuencias. En este listado también indica si además del modelo de evolución hay que aplicar el parámetro gamma e *Invariant*.

4.º Cálculo de la matriz de distancias: con el resultado del alineamiento múltiple y el modelo de evolución elegido, más gamma e *Invariant* si fuesen necesarios, se calcula la matriz de distancias. Se realiza en MEGA-X y en R. Esta matriz se utiliza para hacer el árbol.

5.º Determinar el método de construcción de árboles: se eligió dependiendo del modelo de evolución que había que aplicar y el coste computacional, entre un método basado en la distancia como el de Unión de vecinos (NJ), o un método basado en los caracteres como el de Máxima verosimilitud (ML). Para inferir el árbol hay que utilizar la matriz de distancias calculada, tanto en MEGA-X como en R.

6.º Evaluación final del árbol: se realizó un análisis estadístico computacional que consiste en volver a muestrear las muestras originales para crear nuevos árboles denominado *bootstrapping.* Este análisis en MEGA-X se selecciona cuando se elige el método de construcción de árboles, y se calcula en el mismo proceso. En R hay que introducirle como parámetro un árbol ya calculado y la función que calcula los árboles para que la ejecute las veces que sea necesario. En ambos casos se pudo introducir el número de repeticiones que se consideró oportuno, que puede ser entre 100 y 1000.

7.º Árbol resultante: finalmente se obtuvo un árbol que en el caso de MEGA-X aparece representado gráficamente, es un árbol consenso con los valores de *bootstrap* obtenidos en los

nodos. En el caso de R se mostró en dos formatos: *newick* para poder ser procesado por *FileMaker*, y la representación gráfica en R para su interpretación en este proyecto. En ambas representaciones de R, la organización del árbol se realizó con el que ha comparado la función *bootstrap* todas las repeticiones que ha calculado, y los valores que ha devuelto esta función aparecen en los nodos.

8.º Interpretación de los árboles: comparación de los árboles visualmente, y respecto de los obtenidos en la literatura científica.

4.2. Alineamiento de pares de secuencias del gen ORF5 del virus PRRS

Se recurrió, para hacer pruebas, a secuencias de ADN del gen ORF 5 del virus PRRS porcino, obtenidas de muestras clínicas de porcino provenientes de España recibidas en Exopol. En concreto, se eligió un caso recibido en agosto de 2021 de cerdos lactantes en el que el cliente había enviado anteriormente al laboratorio otras 12 muestras más. Se realizó el alineamiento de pares de secuencias de la nueva secuencia del gen ORF 5 de PRRS, con las otras 12 secuencias, una secuencia de referencia de PRRS europeo (Lelystad) y 5 secuencias de vacunas comercializadas: Pyrsvac, Unistrain, Porcillis-DV, Reprocyc B. I. y Suvaxyn.

La herramienta de la que se obtuvieron los resultados de referencia ha sido LALIGN del Instituto Europeo de Bioinformática (EBI) [15]. Para traducir las secuencias a aminoácidos, se utilizó la herramienta on-line Expasy Translate del Instituto Suizo de Bioinformática [80]. La implementación de código para Exopol se llevó a cabo en R Studio (Versión 1.4.1103 2009-2021 para Windows), utilizando el lenguaje de programación R. La librería necesaria fue Biostrings [81], y los paquetes: BiocGenerics [82], parallel [83]. Se ha manejado, como base, el código elaborado por Iñaki Albizu para realizar alineamientos y su sistema de comunicación con *FileMaker.*

Se realizó un estudio de los parámetros que se aplican en la herramienta web de LALIGN, para realizar alineamientos de pares de secuencias de nucleótidos y aminoácidos: tipo de alineamiento, matriz de sustitución, puntuaciones, fórmula para calcular la identidad, y posteriormente se implementaron los mismos parámetros en R. Para validar la implementación en R, se compararon los resultados con los obtenidos de LALIGN.

El cálculo de la identidad se comienza realizando un alineamiento de pares de secuencias. La secuencia de ADN obtenida en la muestra clínica recibida en agosto de 2021, es la secuencia problema. Esta secuencia se alinea con cada una de las otras 17 secuencias de ADN, de una en una. En el caso de LALIGN durante el alineamiento se calcula el porcentaje de identidad que

aparece junto con el resultado del alineamiento. En el caso de R, primero se realiza el alineamiento y con el resultado que da, se calcula el porcentaje de identidad mediante otra función.

Para el cálculo de la identidad de las secuencias de aminoácidos, primero hay que traducir las secuencias de ADN. LALIGN, no las traduce, así que hay que copiarlas manualmente en la herramienta web de Expasy Translate. Las secuencias resultantes se copiaron en LALIGN, y eligiendo la opción de proteína, se calculó el alineamiento y la identidad. Para realizar este paso en R, se realiza todo igual que con ADN, pero en la función *pairwiseAlignment* que realiza el alineamiento, se introducen las secuencias aplicándoles la función *translate* para traducirlas, ambas funciones pertenecen a la librería Biostrings.

5. Desarrollo

5.1. Influenza A porcino

5.1.1. Análisis filogenético con MEGA-X

Primero, se realizó el estudio filogenético de los 4 subtipos: H1, H3, N1 y N2. Para ello se llevó a cabo la exportación de las 69 secuencias de nucleótidos de los casos clínicos de Exopol desde *FileMaker.* Después, se agruparon por subtipos con las 56 secuencias obtenidas del Genbank (tabla 2), en un archivo tipo FASTA distinto para cada uno.

Para llevar a cabo el estudio filogenético con MEGA-X, se realizó para cada subtipo el alineamiento mediante Clustal Omega y MUSCLE; se pueden ver los parámetros seleccionados en la imagen 10, para valorar cómo afectaba el tipo de alineamiento en los árboles resultantes.

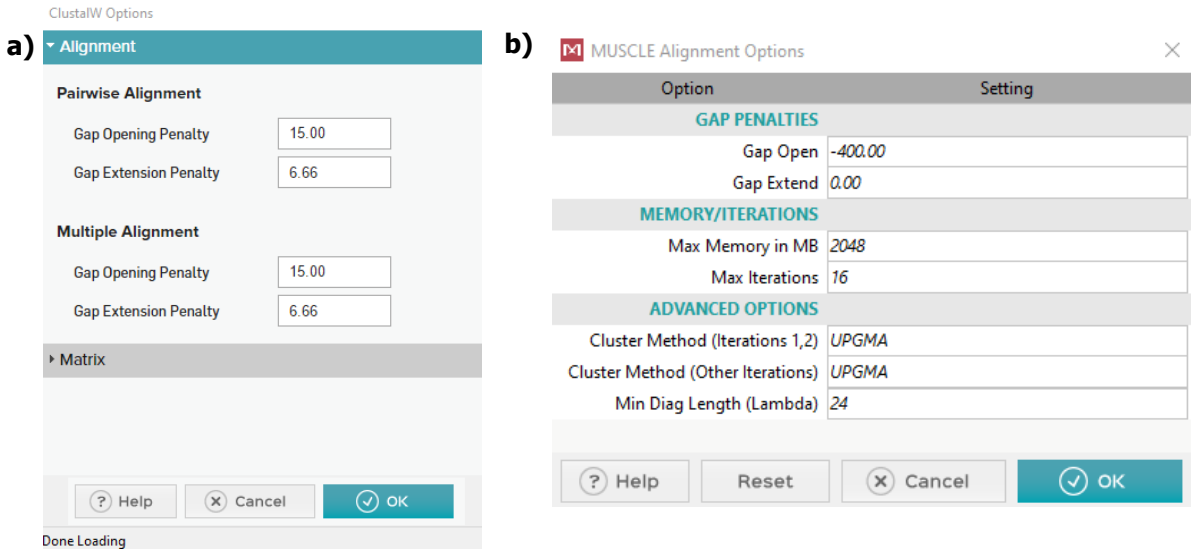


IMAGEN 10: Captura de pantalla de los parámetros seleccionados para realizar los alineamientos a) mediante Clustal Omega, y b) mediante MUSCLE en MEGA-X.

A continuación, se buscó el mejor modelo de evolución para cada alineamiento y variante génica; en la imagen 11, aparecen los parámetros seleccionados para esta búsqueda. Para ejecutar este análisis, se calcula un árbol inicial mediante Unión de vecinos (NJ), y el algoritmo combina aspectos del método de Máxima verosimilitud (ML) con el NJ. Va realizando los árboles con cada modelo de evolución, así como calcula el valor del criterio de información bayesiano (BIC) y el criterio de información de Akaike (AIC) para cada modelo. Se valoró si el modelo que tenía menor puntuación de AIC y BIC era el mismo para los dos tipos de alineamiento. En base a este modelo, se calculó la matriz de distancias de evolución y se elaboró el árbol filogenético siguiendo el método apropiado para el modelo de evolución seleccionado. Así mismo, se tuvieron en cuenta

otros modelos de evolución de los que se ha obtenido una buena puntuación BIC y AIC que no son tan exigentes computacionalmente. Con estos métodos, también se calculó la matriz de distancias de evolución y, con todo ello, se elaboró el árbol. Posteriormente, se compararon los distintos árboles obtenidos teniendo en cuenta el tipo de alineamiento múltiple, el modelo de evolución y el método de elaboración de árboles.

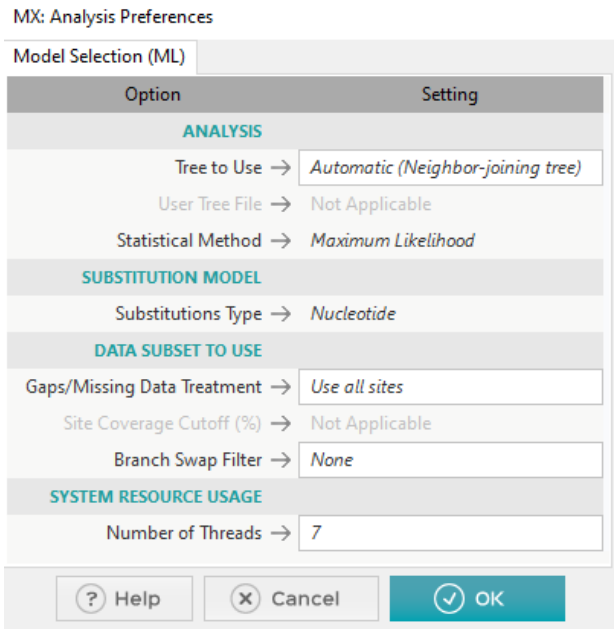


IMAGEN 11: Captura de pantalla de los parámetros seleccionados para realizar la búsqueda del mejor modelo de evolución.

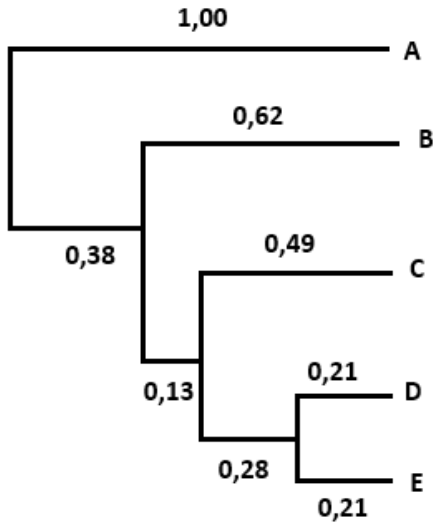
Tras saber qué modelos de evolución eran los idóneos, se calcularon las matrices distancia y se representaron los árboles, que se pueden ver en el apartado de resultados de este trabajo.
Con toda la información obtenida del estudio filogenético realizado con la herramienta de referencia MEGA-X, se decidió qué alineamiento múltiple, modelo de evolución, método para elaborar árboles y parámetros eran los que se iban a implementar en R. Para tomar esta decisión, se tuvo sobre todo en cuenta el coste computacional intentando no perder precisión en el cálculo, y siendo en todo momento conscientes de los errores que se pueden cometer si se elegían las opciones menos precisas en el cálculo de los árboles.

5.1.2. Implementación en R

Tras el estudio filogenético detallado en el apartado anterior y en el de resultados, se seleccionó el alineamiento múltiple y el modelo de evolución más eficiente computacionalmente intentando perder la menor precisión a la hora de inferir los árboles, y se implementó con lenguaje R. El tiempo de ejecución es un aspecto importante, ya que el número de secuencias que se pueden comparar puede ser variable y reducir el tiempo de las tareas es fundamental para desarrollar el trabajo diario en la empresa. Se empleó como base el código ya implementado en R con el que

se trabajaba en Exopol, que, además, permitía la comunicación con *FileMaker,* así que se mantuvo la estructura de comunicación y alguna de las partes. A partir de este código, se desarrollaron las mejoras oportunas.

Este código efectuaba un alineamiento múltiple mediante MUSCLE y, con la matriz obtenida del alineamiento, se elaboraba el árbol mediante NJ una sola vez, sin tener en cuenta ningún modelo de evolución ni calcular ninguna matriz de distancias. Después, se procedía a copiar en *Clipboard,* mediante formato *newick,* la estructura y distancias del árbol. Y, con una herramienta de *FileMaker,* se dibujaba el árbol para que se pudiera integrar en los informes para los clientes de Exopol. El formato *newick*, permite representar árboles usando paréntesis y comas para agrupar los clados, utiliza dos puntos para señalar que lo siguiente que aparece es valor de la distancia al antecesor, y termina el árbol con un punto y coma. Este formato se puede copiar como un texto plano, y al tener una estructura fija se puede leer por otro programa para representarlo gráficamente, en este caso la herramienta de *FileMaker* para graficar. Un ejemplo de este formato sería el que se muestra en la imagen 12.



Formato newick: ((((E: 0,21, D: 0,21): 0,28, C: 0,49): 0,13, B: 0,62): 0,38, A: 1,00);

IMAGEN 12. Ejemplo del formato newick para representar árboles y su representación gráfica. Los clados se ponen entre paréntesis, y se van anidando formando la estructura del árbol. Después del nombre del taxón y los dos puntos se escribe la distancia evolutiva. El árbol se finaliza con punto y coma.

Las mejoras que se querían implementar abarcaban tres aspectos:

- Introducir el modelo de evolución apropiado para estas secuencias.
- Realizar *bootstrap* y mostrar el resultado en cada nodo del árbol.
- Todo ello sin aumentar el tiempo de ejecución.

La función *boot.phylo* utilizada para hacer el *bootstrap* comparaba un árbol realizado por NJ con otros 100 realizados por NJ también, con lo que se obtiene el número de veces que ha salido ese mismo nodo. Se incluyó el enraizamiento de punto medio dentro de la función que hace el árbol NJ, de ese modo, todos los árboles que comparaba estaban enraizados igual y no habría discrepancias a la hora de colocar en los nodos correspondientes la frecuencia de recurrencia de ese nodo.

Las mejoras que este trabajo aportó fueron:

- Representar el árbol obtenido mediante NJ.
- Utilizar una matriz de distancias calculada con un modelo de evolución.
- Representar en los nodos de ese árbol los valores de *bootstrap* obtenidos de pseudorreplicar el árbol, permitiendo valorar la confianza que se puede depositar en el árbol filogenético.

El resultado se devolvió en formato *newick* incluyendo, además de las distancias y la organización de los nodos, el valor obtenido en el *bootstrap* para poder incluirlo en el dibujo en cada nodo. Estos valores se situarn después de los dos puntos y antes de la distancia del clado.
Debido a que añadir los valores del *bootstrap* en la representación gráfica del árbol desde el formato *newick* en *FileMaker* lleva varias horas de programación por parte del informático, todavía no se ha realizado y los árboles resultantes que se muestran en este trabajo están graficados con R Studio.

5.2. Estudio de los parámetros para el cálculo de la identidad para PRRS

Para el cálculo de la identidad, se realizó un análisis del tipo de secuencias a comparar y la herramienta on-line de la que se obtenían resultados óptimos, concretamente, LALIGN del EBI. En este caso, las secuencias del gen ORF 5 del virus PRRS son secuencias que sufren mutaciones, pero no son secuencias demasiado variantes; normalmente, presentan porcentajes de identidad altos. Además, las secuencias obtenidas de los casos clínicos en Exopol son de longitud variable.

La herramienta LALIGN utiliza el algoritmo the Smith-Waterman (SSE2, Michael Farrar, 2006) (7.2 Nov 2010) que es un alineamiento local. Los parámetros que se utilizaron fueron los que aparecen por defecto en la web, que son los que se utilizan por el Departamento.

Las puntuaciones de los parámetros para alineamientos de pares de secuencias de DNA son: *match:* +5; *mismatch:* -4; *gap opening:* -12; *gap extension:* 0.

Para alineamientos de pares de secuencias de aminoácidos, utiliza una matriz BLOSUM50, *gap opening:* -12; *gap extension:* -2.

Con esta información, se trató de incluir todos estos parámetros en el código de R. En este caso, en el código original se estaban realizando alineamientos globales y no se tenía en cuenta ninguna penalización ni matriz de sustitución; se utilizaban las funciones sin aplicar ningún parámetro.

En este caso, el alineamiento más apropiado es un alineamiento local, ya que la longitud de las secuencias es variable, y se estaba penalizando esto en los porcentajes de identidad al llevar a cabo un alineamiento global, que buscar repartir la secuencia más corta por toda la longitud de la secuencia más larga.

La identidad se calculó en R mediante la función *pid* con el argumento *type="PID1",* que utiliza la siguiente ecuación:

$$\%\ de\ identidad\ de\ la\ secuencia = \frac{100 \cdot\ (posiciones\ idénticas)}{(posiciones\ alineadas\ +\ posiciones\ con\ gap\ interno)}$$

En las secuencias de nucleótidos, se incluyen 11 letras distintas de los nucleótidos habituales: A, C, G, T; estas letras se incluyen en lugar de uno de los cuatro nucleótidos cuando hay ambigüedad en la secuenciación. Dependiendo del grado de ambigüedad, se categoriza con una letra u otra, y se le da una puntuación mayor o menor cuando se realiza el alineamiento, ver tabla 3. Por ejemplo: se escribe una S, cuando hay duda entre C y G; una B cuando se sabe que no es A; o una N cuando puede ser cualquiera de los cuatro nucleótidos. Por lo tanto, con todo esto se puede generar una matriz de sustitución que contemple estas ambigüedades. Esto no se estaba teniendo en cuenta en el alineamiento de pares de secuencias de ADN, por ello, se incluyó en el código R la lectura de un archivo de texto con la matriz de sustitución que incluía las ambigüedades recogidas por la IUPAC (International Union of Pure and Applied Chemistry), con los mismos valores de *match:* +5 y *mismatch:* -4, que LALIGN.

TABLA 3. Código elaborado por la IUPAC para nombrar los nucleótidos ambiguos en secuencias de nucleótidos.

Nucleótido ambiguo	Significado
S	C o G
W	A o T
R	A o G
Y	C o T
K	G o T
M	A o C
B	No es A
V	No es T
H	No es G
D	No es C
N	Cualquiera

En la función *pairwiseAlignment* de R, se incluyó además del tipo de alineamiento: local, la matriz de sustitución (con las ambigüedades de la IUPAC que incluye los valores de *match:* +5 y *mismatch:* -4), y los valores de *gap opening:* -12, *gap extension:* 0, como en LALIGN.

También, se realiza el alineamiento de pares de secuencias de las mismas secuencias, pero traducidas a aminoácidos, para ello se utiliza la función *translate* de R. Una vez traducidas se utiliza la función *pairwiseAlignment* de R, con el tipo de alineamiento local, y se puso como matriz de sustitución la BLOSUM50 como la utilizada en LALIGN.

Para comprobar si los resultados obtenidos con las modificaciones en R eran iguales a los de LALIGN, se utilizó como secuencia problema, la obtenida del caso clínico recibido en agosto de 2021. Esta secuencia se comparó con las otras 12 secuencias obtenidas en otros casos clínicos del mismo cliente y con las 5 secuencias de vacunas comercializadas.

Se eligió el caso desde *FileMaker,* las secuencias seleccionadas se exportan en un archivo .txt, este archivo lo lee el código de R. Se calculan los resultados de identidad y el código devuelve el resultado que se muestra en *FileMaker* mediante *writeClipboard.*

Para realizar la validación de estos resultados en LALIGN se introdujeron las secuencias manualmente en la web. La secuencia problema era la obtenida en el caso clínico seleccionado, y se comparó con el resto de secuencias del mismo cliente. Se seleccionó la opción de ADN para analizar las secuencias de nucleótidos, y se dejaron los parámetros por defecto, que son los anteriormente citados. Para realizar el alineamiento de las secuencias traducidas, se copiaron manualmente las secuencias de ADN en la herramienta web de Expasy Translate. Después las secuencias de aminoácidos resultantes se copiaron en LALIGN. Se seleccionó la opción de proteínas, y se mantuvieron los parámetros por defecto, que son los nombrados anteriormente.

En la parte superior del alineamiento que realiza LALIGN aparece el resultado del cálculo de la identidad, que fue lo que se comparó con el resultado que aparecía en *FileMaker*.

6. Estudio económico

Para efectuar el estudio económico de este proyecto, se ha tenido en cuenta: el coste de la mano de obra (tabla 4), el coste de la infraestructura en la que se ha realizado el trabajo (tabla 5) y el coste del material utilizado (tabla 6).

Tabla 4: Coste de la mano de obra.

Puesto de trabajo	Horas totales	Coste salario por hora	Coste total
Bioinformático	150 h	20 €	3.000 €
		Total:	**3.000 €**

La oficina de trabajo está situada dentro de las instalaciones de Exopol, S. L.; el coste estimado del tiempo que duró el proyecto se detalla en la tabla 5.

Tabla 5: Coste infraestructura por Oficina dentro de Exopol.

	Duración proyecto	Coste mensual	Coste total
Oficina Exopol	20 días	150 €	150 €
Electricidad	20 días	40 €	40 €
Internet	20 días	25 €	25 €
		Total:	**215 €**

Considerando que el ordenador que se ha usado tiene una vida útil media de 7 años, es decir, unas 17.000 horas, e incluye el teclado y el ratón, el coste para las horas dedicadas al proyecto es el que se muestra en la tabla 6.

Tabla 6: Coste del material utilizado.

	Precio	Tiempo de uso	Coste
Ordenador de sobremesa Acer Aspire C24-963, 23,8" [84]	749 €	150 h	6,60 €
		Total:	**6,60 €**

Finalmente, en la tabla 7, se presenta el presupuesto final para este proyecto, sumando los costes que se han detallado y aplicando un margen de beneficio.

Tabla 7: Presupuesto final del proyecto.

Mano de obra	3.000 €
Infraestructura	215 €
Material	6,60 €
Total	**3.221,60 €**
Margen de beneficio	30 %
Coste total presupuestado	**4.188,08 €**

En este caso, el beneficio principal es la fidelización del cliente al obtener un resultado más preciso y fiable. Del mismo modo, también mejora la imagen de la empresa para atraer nuevos clientes. Exopol invierte mucho en I+D+i y publica numerosas comunicaciones científicas, realizar estas mejoras y conocer mejor las secuencias con las que se trabaja, también mejorará aspectos de estas publicaciones y aumentará el conocimiento científico. Esto puede contribuir, por ejemplo, a dar publicidad a la empresa, o generar oportunidades para iniciar nuevos proyectos.

7. Resultados

7.1. Influenza A porcino

7.1.1. Alineamiento múltiple de las secuencias y búsqueda del modelo de evolución

Se decidió trabajar con las secuencias de nucleótidos y no con secuencias de aminoácidos, debido a que se están comparando genes de una misma especie de virus muy variables, y con los aminoácidos perderíamos información de muchas mutaciones que pueden ser decisivas para que en el árbol filogenético se formen los nodos.

Se compararon los alineamientos múltiples realizados con Clustal Omega y con MUSCLE, los resultados son muy similares entre ellos para los cuatro subtipos de secuencias de Influenza A: H1, H3, N1 y N2. En la imagen 13 se puede ver un fragmento del alineamiento de las secuencias H1 mediante Clustal Omega y en la imagen 14 el realizado con MUSCLE. Marcado en la parte superior en amarillo, aparecen señalados con un punto los sitios conservados, y en ambos fragmentos están alineados igual.

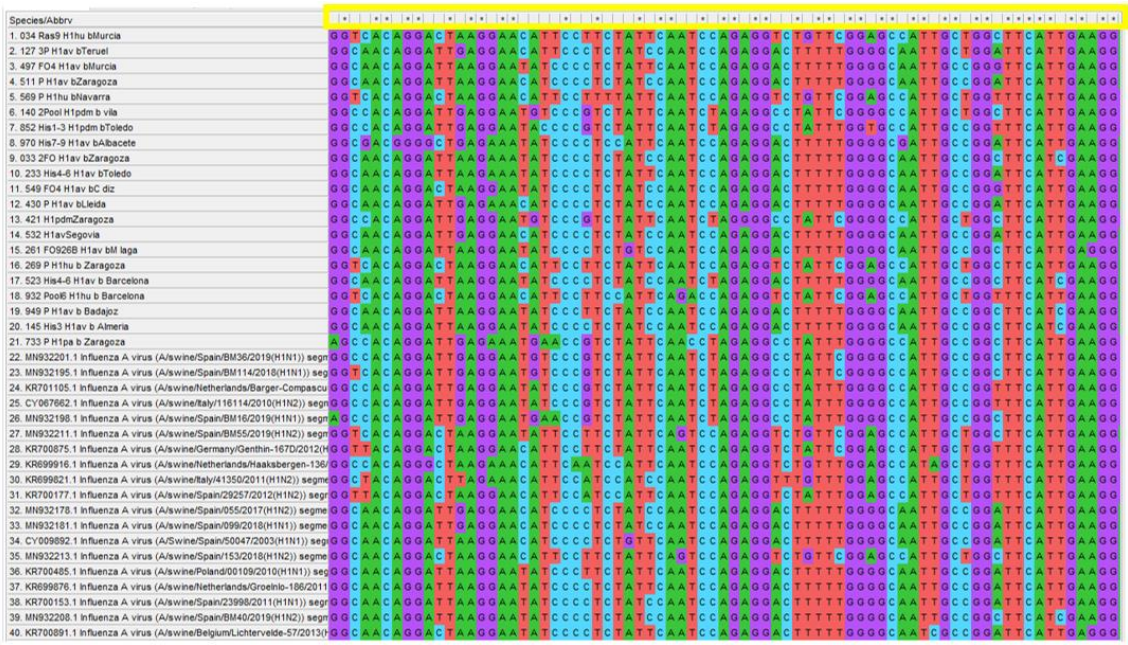


IMAGEN 13. Fragmento del alineamiento múltiple realizado con Clustal Omega de las secuencias del subtipo H1 de Influenza A porcino. En la parte superior de la imagen, aparecen unos puntos que señalan que esa posición está conservada.

IMAGEN 14. Fragmento del alineamiento múltiple realizado con MUSCLE de las secuencias del subtipo H1 de Influenza A porcino. En la parte superior de la imagen, aparecen unos puntos que señalan que esa posición está conservada.

Los modelos de evolución obtenidos con un Clustal Omega y con MUSCLE fueron para H1, H3, N1 y N2 exactamente los mismos y en el mismo orden de valoración de BIC y AIC. Hay alguna diferencia en el valor de BIC, AIC y gamma (+*G*), pero como se puede comparar entre la imagen 15 –que es el resultado obtenido por Clustal Omega para H1–, y el resultado de la imagen 16 –que es el obtenido tras alinear las secuencias H1 con MUSCLE–, la diferencia entre los valores BIC y AIC es de aproximadamente 9 unidades de diferencia, lo que no supone ninguna variación en el orden y tampoco afecta a la hora de decidir cuál es modelo idóneo. Para el valor gamma, la diferencia es 0,01 así que tampoco afecta. En el anexo 3, se pueden ver las tablas obtenidas para H3, N1 y N2, en las que sucedió lo mismo o incluso no se encontró ninguna diferencia. Por lo tanto, el algoritmo de alineamiento múltiple no ha influido, en este caso, en la selección del modelo de evolución.

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(
GTR+G+I	87	27632.123	26862.851	-13344.276	0.41	2.46	4.18	0.350	0.249	0.182	0.219	0.019	0.030	0.166	0.027	0.160	0.021	0.058	0.220	0.005	0.
HKY+G+I	83	27668.896	26934.980	-13384.354	0.41	2.58	4.20	0.350	0.249	0.182	0.219	0.024	0.017	0.177	0.034	0.147	0.021	0.034	0.202	0.021	0.
GTR+G	86	27690.417	26929.984	-13378.846	n/a	0.44	4.17	0.350	0.249	0.182	0.219	0.020	0.031	0.162	0.028	0.164	0.019	0.059	0.225	0.005	0.
TN93+G+I	84	27692.999	26950.244	-13390.982	0.41	2.37	4.99	0.350	0.249	0.182	0.219	0.020	0.015	0.160	0.029	0.177	0.018	0.029	0.243	0.018	0.
T92+G+I	81	27723.467	27007.229	-13422.485	0.41	2.45	4.13	0.300	0.300	0.200	0.200	0.028	0.019	0.162	0.028	0.162	0.019	0.028	0.243	0.019	0.
GTR+I	86	27725.377	26964.944	-13396.326	0.46	n/a	4.00	0.350	0.249	0.182	0.219	0.020	0.029	0.167	0.029	0.156	0.024	0.057	0.215	0.006	0.
HKY+G	82	27733.743	27008.666	-13422.200	n/a	0.44	4.24	0.350	0.249	0.182	0.219	0.024	0.017	0.177	0.033	0.147	0.021	0.033	0.202	0.021	0.
TN93+G	83	27737.532	27003.616	-13418.672	n/a	0.44	4.18	0.350	0.249	0.182	0.219	0.024	0.017	0.163	0.033	0.162	0.021	0.033	0.223	0.021	0.
HKY+I	82	27757.199	27032.122	-13433.928	0.46	n/a	3.98	0.350	0.249	0.182	0.219	0.025	0.018	0.175	0.035	0.145	0.022	0.035	0.200	0.022	0.
TN93+I	83	27763.054	27029.138	-13431.433	0.46	n/a	3.96	0.350	0.249	0.182	0.219	0.025	0.018	0.166	0.035	0.155	0.022	0.035	0.214	0.022	0.
T92+G	80	27780.849	27073.450	-13456.599	n/a	0.44	4.16	0.300	0.300	0.200	0.200	0.028	0.019	0.163	0.028	0.163	0.019	0.028	0.244	0.019	0.
T92+I	80	27817.205	27109.807	-13474.777	0.45	n/a	3.91	0.300	0.300	0.200	0.200	0.030	0.020	0.161	0.030	0.161	0.020	0.030	0.241	0.020	0.
K2+G+I	80	27921.810	27214.412	-13527.079	0.40	2.05	5.16	0.250	0.250	0.250	0.250	0.020	0.020	0.209	0.020	0.209	0.020	0.020	0.209	0.020	0.
K2+G	79	27947.841	27249.282	-13545.518	n/a	0.44	4.13	0.250	0.250	0.250	0.250	0.024	0.024	0.201	0.024	0.201	0.024	0.024	0.201	0.024	0.
K2+I	79	28008.855	27310.296	-13576.025	0.46	n/a	3.87	0.250	0.250	0.250	0.250	0.026	0.026	0.199	0.026	0.199	0.026	0.026	0.199	0.026	0.
GTR	85	29176.435	28424.841	-14127.278	n/a	n/a	3.12	0.350	0.249	0.182	0.219	0.024	0.033	0.152	0.034	0.153	0.021	0.064	0.210	0.021	0.
TN93	82	29201.649	28476.572	-14156.153	n/a	n/a	3.37	0.350	0.249	0.182	0.219	0.028	0.020	0.152	0.039	0.160	0.025	0.039	0.219	0.025	0.
HKY	81	29214.378	28498.141	-14167.941	n/a	n/a	3.38	0.350	0.249	0.182	0.219	0.028	0.021	0.169	0.040	0.140	0.025	0.040	0.193	0.025	0.
T92	79	29235.606	28537.047	-14189.400	n/a	n/a	3.38	0.300	0.300	0.200	0.200	0.033	0.022	0.156	0.033	0.156	0.022	0.033	0.233	0.022	0.
K2	78	29427.381	28737.662	-14290.711	n/a	n/a	3.36	0.250	0.250	0.250	0.250	0.029	0.029	0.193	0.029	0.193	0.029	0.029	0.193	0.029	0.
JC+G+I	79	29925.190	29226.632	-14534.192	0.40	2.67	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.
JC+G	78	29975.634	29285.914	-14564.837	n/a	0.49	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.
JC+I	78	30006.719	29316.999	-14580.379	0.45	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.
JC	77	31318.472	30637.592	-15241.679	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+*G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+*I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 40 nucleotide sequences. There were a total of 1761 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

IMAGEN 15. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias H1 mediante Clustal Omega en MEGA-X.

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G+I	87	27641.199	26871.927	-13348.814	0.41	2.45	4.18	0.350	0.249	0.182	0.219	0.019	0.030	0.166	0.027	0.160	0.021	0.058	0.220	0.005	0.265	0.024	0.004
HKY+G+I	83	27677.922	26944.006	-13388.867	0.41	2.58	4.21	0.350	0.249	0.182	0.219	0.024	0.017	0.177	0.033	0.147	0.021	0.033	0.202	0.021	0.283	0.024	0.017
GTR+G	86	27699.013	26938.580	-13383.144	n/a	0.44	4.18	0.350	0.249	0.182	0.219	0.020	0.030	0.162	0.028	0.164	0.019	0.059	0.225	0.005	0.260	0.022	0.004
TN93+G+I	84	27702.218	26959.463	-13395.592	0.41	2.35	5.00	0.350	0.249	0.182	0.219	0.020	0.015	0.160	0.029	0.177	0.018	0.029	0.244	0.018	0.256	0.020	0.015
T92+G+I	81	27732.534	27016.296	-13427.019	0.41	2.44	4.13	0.300	0.300	0.200	0.200	0.028	0.019	0.162	0.028	0.162	0.019	0.028	0.243	0.019	0.243	0.028	0.019
GTR+I	86	27735.164	26974.731	-13401.220	0.46	n/a	4.01	0.350	0.249	0.182	0.219	0.020	0.029	0.167	0.029	0.156	0.024	0.057	0.215	0.006	0.267	0.027	0.005
HKY+G	82	27742.246	27017.169	-13426.452	n/a	0.44	4.24	0.350	0.249	0.182	0.219	0.024	0.017	0.177	0.033	0.147	0.021	0.033	0.202	0.021	0.284	0.024	0.017
TN93+G	83	27746.063	27012.147	-13422.937	n/a	0.44	4.18	0.350	0.249	0.182	0.219	0.024	0.017	0.163	0.033	0.162	0.021	0.033	0.223	0.021	0.261	0.024	0.017
HKY+I	82	27767.073	27041.996	-13438.865	0.46	n/a	3.99	0.350	0.249	0.182	0.219	0.025	0.018	0.175	0.035	0.145	0.022	0.035	0.200	0.022	0.280	0.025	0.018
TN93+I	83	27772.894	27038.978	-13436.353	0.46	n/a	3.96	0.350	0.249	0.182	0.219	0.025	0.018	0.166	0.035	0.156	0.022	0.035	0.214	0.022	0.266	0.025	0.018
T92+G	80	27789.404	27082.006	-13460.876	n/a	0.44	4.17	0.300	0.300	0.200	0.200	0.028	0.019	0.163	0.028	0.163	0.019	0.028	0.244	0.019	0.244	0.028	0.019
T92+I	80	27827.071	27119.672	-13479.710	0.45	n/a	3.92	0.300	0.300	0.200	0.200	0.029	0.020	0.161	0.029	0.161	0.020	0.029	0.241	0.020	0.241	0.029	0.020
K2+G+I	80	27930.320	27222.921	-13531.334	0.40	2.04	5.18	0.250	0.250	0.250	0.250	0.020	0.020	0.210	0.020	0.210	0.020	0.020	0.210	0.020	0.210	0.020	0.020
K2+G	79	27955.727	27257.168	-13549.461	n/a	0.44	4.14	0.250	0.250	0.250	0.250	0.024	0.024	0.201	0.024	0.201	0.024	0.024	0.201	0.024	0.201	0.024	0.024
K2+I	79	28018.077	27319.518	-13580.636	0.46	n/a	3.88	0.250	0.250	0.250	0.250	0.026	0.026	0.199	0.026	0.026	0.199	0.026	0.199	0.026	0.199	0.026	0.026
GTR	85	29189.403	28437.809	-14133.762	n/a	n/a	3.13	0.350	0.249	0.182	0.219	0.024	0.033	0.152	0.034	0.153	0.021	0.064	0.210	0.021	0.244	0.024	0.018
TN93	82	29214.687	28489.611	-14162.672	n/a	n/a	3.38	0.350	0.249	0.182	0.219	0.028	0.020	0.152	0.039	0.160	0.025	0.039	0.220	0.025	0.244	0.028	0.020
HKY	81	29227.511	28511.273	-14174.507	n/a	n/a	3.38	0.350	0.249	0.182	0.219	0.028	0.021	0.169	0.040	0.140	0.025	0.040	0.193	0.025	0.271	0.028	0.021
T92	79	29248.669	28550.110	-14195.932	n/a	n/a	3.38	0.300	0.300	0.200	0.200	0.033	0.022	0.156	0.033	0.156	0.022	0.033	0.234	0.022	0.234	0.033	0.022
K2	78	29439.763	28750.044	-14296.902	n/a	n/a	3.37	0.250	0.250	0.250	0.250	0.029	0.029	0.193	0.029	0.193	0.029	0.029	0.193	0.029	0.193	0.029	0.029
JC+G+I	79	29937.517	29238.958	-14540.356	0.40	2.66	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G	78	29987.450	29297.731	-14570.745	n/a	0.49	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	78	30019.681	29329.961	-14586.860	0.45	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	77	31334.324	30653.444	-15249.605	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+*G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+*I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 40 nucleotide sequences. There were a total of 1761 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor.

1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

IMAGEN 16. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias H1 mediante MUSCLE en MEGA-X.

7.1.2. Elaboración de los árboles

Los alineamientos para cada uno de los árboles filogenéticos inferidos con cada modelo se realizaron con MUSCLE y con Clustal Omega (tabla 8). El tipo de alineamiento no influyó en el resultado final de los árboles inferidos para los genes de Influenza A: HA y NA; los árboles eran casi idénticos y las pequeñas variaciones obtenidas fueron en nodos con valores de *bootstrap* bajos.

Ya que el tipo de alineamiento no afectó tampoco a la obtención del modelo de evolución, se eligió MUSCLE para los de árboles realizados con MEGA-X que se van a mostrar en este apartado, debido a que este alineamiento es más rápido y preciso que Clustal W. Todos los árboles que se muestran inferidos con este software son árboles consenso. El software MEGA-X, para calcular árboles realizados con ML, no permite introducir un valor de gamma inferior a 2.

Con las secuencias de cada subtipo H1, H3, N1 y N2, se elaboraron diferentes árboles en función de los modelos de evolución obtenidos (tabla 8). El modelo de evolución General Reversible en el tiempo (GTR), que es el más complejo, y el Hasegawa-Kisahano-Yano 1985 (HKY), no se pueden utilizar para realizar árboles con NJ, ni en MEGA ni en R, pero sí se pueden aplicar con el método ML. Para realizar los árboles con NJ, se eligieron los modelos de evolución con mejor puntuación AIC y BIC, de entre los que sí admite este método en MEGA-X y en R: Tamura-Nei 1993 (TN93), Tamura 1992 (T92), Jukes-Cantor 1969 y Kimura 1980.

TABLA 8. Métodos para la elaboración de árboles y modelos de evolución utilizados para inferir los árboles para cada subtipo del gen HA y NA de Influenza A porcino.

Subtipo	Máxima verosimilitud (ML)	Unión de vecinos (NJ)
H1	GTR +G +I	TN93 +G +I
H3	HKY +G	TN93 +G
N1	GTR +G	T92 +G
N2	GTR +G +I	T92 +G +I

En la tabla 9, se muestra el resultado de comparar los árboles inferidos con MEGA-X. Se han comparado los árboles realizados con ML y NJ para cada uno de los subtipos H1, H3, N1 y N2, aplicando el modelo de evolución y parámetros que se han detallado en tabla 8 para cada método. Se han tenido en cuenta diversos aspectos clave, como la agrupación general de las secuencias en los clados principales. También se han valorado las diferencias, si se han formado los mismos grupos monofiléticos, si cada clado se ha formado en la misma posición y con las mismas secuencias. Si se han encontrado diferencias, se tuvo en cuenta si los valores de *bootstrap* del

nodo del antecesor eran bajos en ambos árboles, y podía ser la causa de la variación. Se ha comprobado que los nodos que tenían valores entre 100 y 95 de *bootstrap* aparecían en ambos árboles, y formaban clados con las mismas secuencias y con la misma relación evolutiva. Finalmente se muestra la valoración de que los árboles realizados con ML y NJ, aplicando los modelos de evolución apropiados para cada uno, son equivalentes para H1, H3, N1 y N2.

TABLA 9. Resultados de la comparativa de los árboles filogenéticos inferidos con ML y NJ aplicando los modelos de evolución con mejor puntuación para cada uno, de las secuencias de los subtipos H1, H3, N1 y N2 de Influenza A porcino con MEGA-X.

	H1	**H3**	**N1**	**N2**
Agrupación de las secuencias en grandes clados	Sí, en tres clados que diferencian los tres linajes: humano, pandémico y aviar	Sí, en dos grandes clados	Sí, en tres clados	Sí, en dos clados
Las secuencias que forman cada uno de los clados principales son las mismas	Sí	Sí	Sí	Sí
Los clados coinciden en al menos un 50% de ellos	Sí, en un 50%	Sí, en un 80%	Sí, en un 50%	Sí, en un 50%
Las diferencias se suelen asociar a valores bajos de *bootstrap*	Sí	Sí	Sí	Sí
Aparecen los clados iguales de los nodos con valores altos de *bootstrap* (100-95)	Sí	Sí	Sí	Sí
Son equivalentes	Sí	Sí	Sí	Sí

Para las secuencias del subtipo H1, se obtuvo un árbol tal y como describe la literatura, con tres grandes clados diferenciados (imagen 17), uno con las secuencias provenientes del linaje pandémico (H1pdm), otro con las secuencias del linaje aviar (H1av) y, por último, las secuencias del linaje humano (H1hu), como describió Sosa (2020) [59].

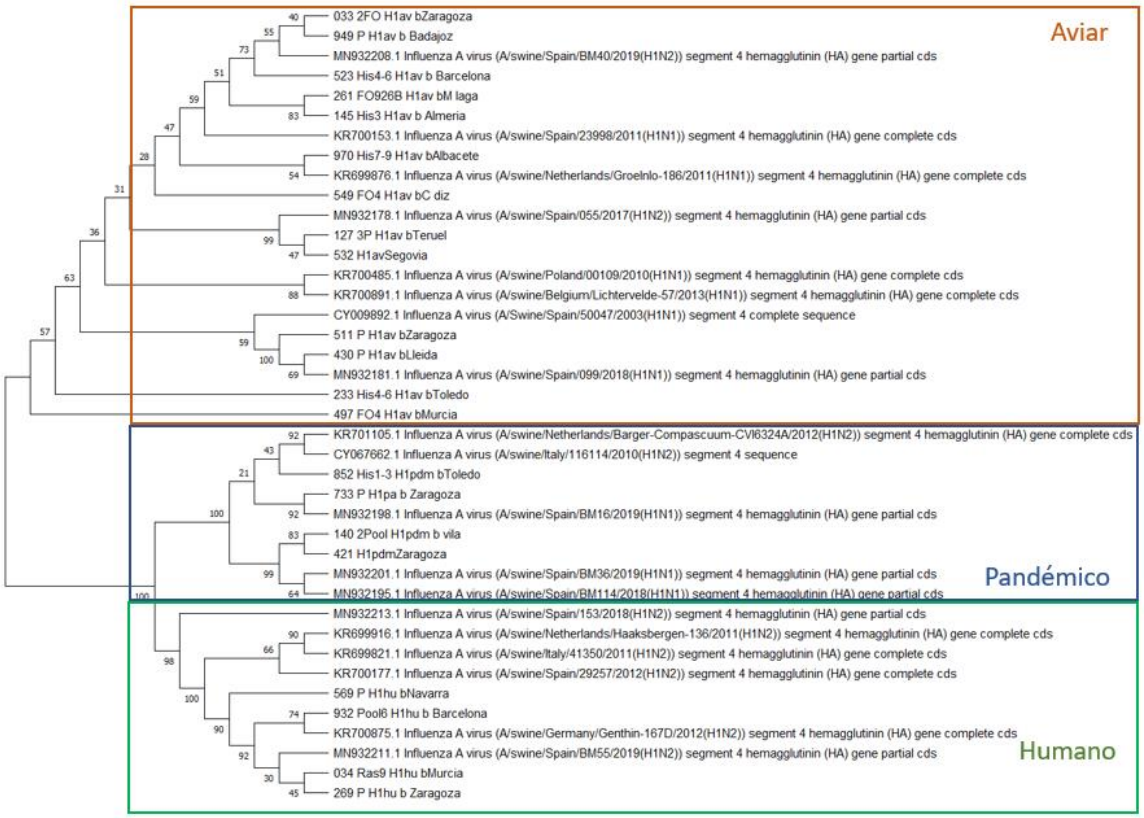


IMAGEN 17. Árbol inferido de las secuencias del subtipo H1 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 2,35 e I, *bootstrap* = 1000. En naranja, se agrupan las secuencias del linaje aviar, en verde, las del linaje humano y, en azul, las del linaje pandémico.

Se compararon los árboles obtenidos para el subtipo H1 con cada modelo de evolución, el obtenido con el modelo de Tamura-Nei 1993 realizado con NJ (imagen 17); y el inferido por el método de ML y el modelo GTR +G +I (imagen 18). Ambos árboles son muy similares. Como se puede ver en las dos imágenes 17 y 18, se distinguen los clados de los tres linajes: aviar, humano y pandémico. Además, la organización de los clados y nodos coincide en ambos árboles, aunque se encontraron algunas diferencias que se señalan más abajo. Los valores de *bootstrap* también fueron similares.

- La cepa 233 del linaje aviar procedente de Toledo con NJ aparece como un *outgroup,* sin embargo, en ML se incluye dentro de un clado con la cepa de referencia CY009892. Los valores de *bootstrap* en los nodos de ambos árboles no son altos, 26 y 17 para GTR, y 57 para NJ, lo que muestra que es normal que aparezca en lugares diferentes.

- La cepa MN932213 aparece como *outgroup* en clado del linaje humano en NJ con valor de *bootstrap* de 98, pero en ML aparece dentro de un clado con otra cepa de referencia MN932211; en este caso, los valores de *bootstrap* son menores en ese nodo y en los nodos ancestrales hasta el nodo del que proviene en NJ. Así que, de nuevo, esta diferencia es aceptable en lo que se refiere a probabilidad.

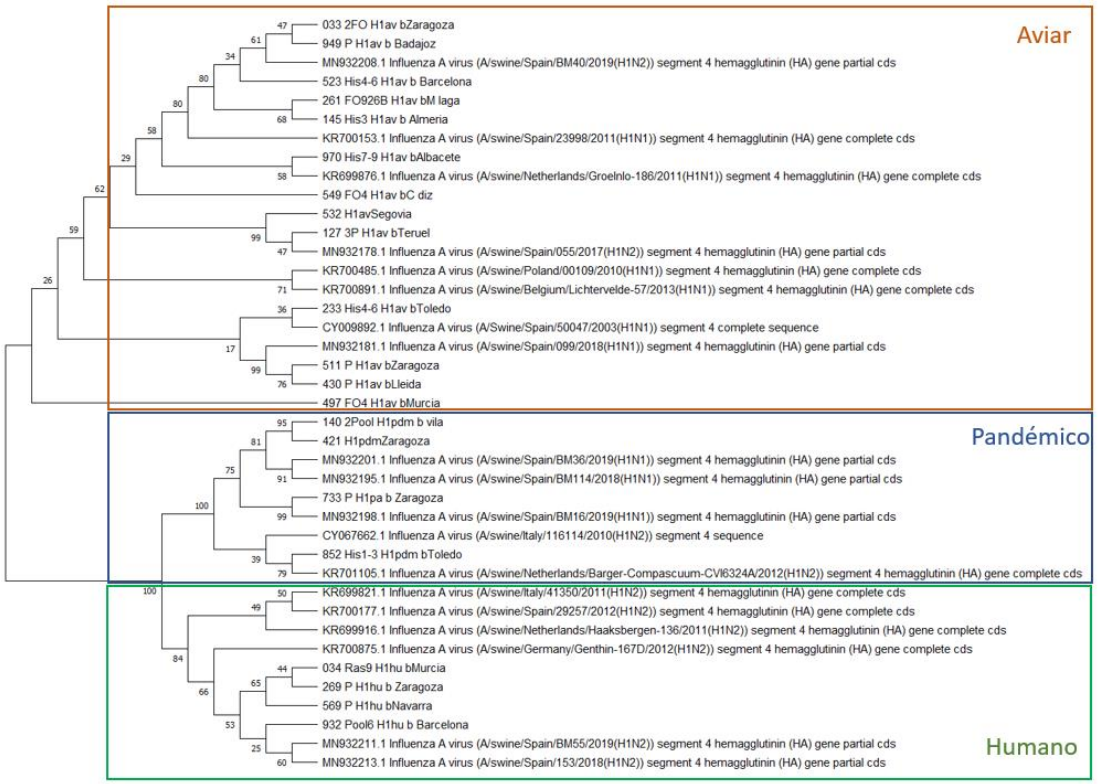


IMAGEN 18. Árbol inferido de las secuencias del subtipo H1 mediante el método ML, y el modelo de evolución General Reversible en el Tiempo 1986 aplicando el parámetro gamma = 2 e I, *bootstrap* = 1000. En naranja, se agrupan las secuencias del linaje aviar, en verde, las del linaje humano y, en azul, las del linaje pandémico.

Para el subtipo H3, en los árboles inferidos tanto con NJ (imagen 19) como con ML (imagen 20) las secuencias se dividen en dos clados, tal y como describe Sosa (2020) [59]. A diferencia de H1, para este subtipo no se han descrito linajes todavía. Uno de los clados contiene las secuencias obtenidas de cepas de muestras clínicas recibidas en Exopol junto con la cepa de referencia MN929035 procedente de España, y el otro con el resto de las cepas de referencia. El clado de las secuencias de referencia agrupa a las secuencias que son parecidas al denominado H3 clásico de 1984 procedente de Bélgica, estas cepas son de hasta 2017 [59]. Por otro lado, está el clado con las cepas más recientes españolas, incluida la cepa de referencia española MN929035 que es de 2019, sobre este clado, Sosa describe que en los últimos tres años han aparecido estos aislamientos que se agrupan con aislados humanos de Malasia en 2004 y de Dinamarca en 2005, poco comunes en cerdos, y por tanto se alejan del H3 clásico, lo que puede suponer su desaparición y sustitución por este nuevo tipo [59].

Se puede observar que, en ambos árboles, las secuencias de la misma procedencia geográfica se suelen agrupar juntas. En el clado de las secuencias de muestras clínicas, no hay diferencias entre ambos árboles. Sin embargo, en el clado de las secuencias de referencia hay diferencias con valores de *bootstrap* altos en los nodos. La secuencia KR701241 y el clado formado por KR701447 y KR701463 están en posiciones distintas. En NJ, la cepa KR701241 se separa

evolutivamente en un nodo de un ancestro más antiguo y, en ML, es el clado de KR701447 y KR701463 el que diverge en ese ancestro.

Ya que el método ML y el modelo de evolución Hasegawa-Kishano-Yano 1985 son los más adecuados, y los valores de *bootstrap* obtenidos para esos nodos son bastante altos, se considera que es más probable que sea correcto el árbol obtenido por ML, pero esta diferencia es perfectamente asumible, teniendo en cuenta el coste computacional que tiene el método ML (el cálculo del árbol superó los 40 minutos, sin embargo, el de NJ utilizó menos de 2 minutos).

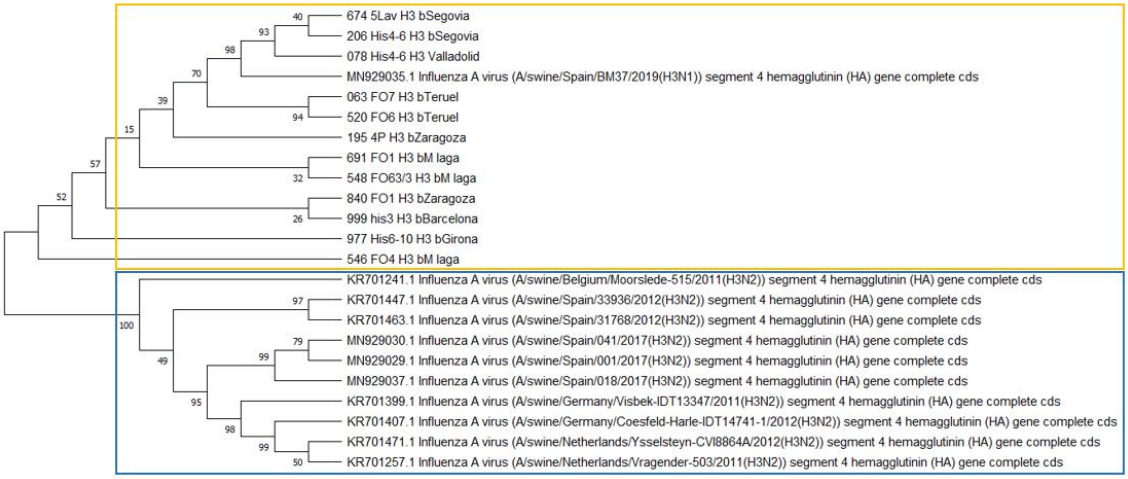


IMAGEN 19. Árbol inferido de las secuencias del subtipo H3 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 0,38 y *bootstrap* = 1000. En amarillo, se agrupan las secuencias obtenidas de muestras clínicas recibidas en Exopol y, en azul, las secuencias de las cepas de referencia.

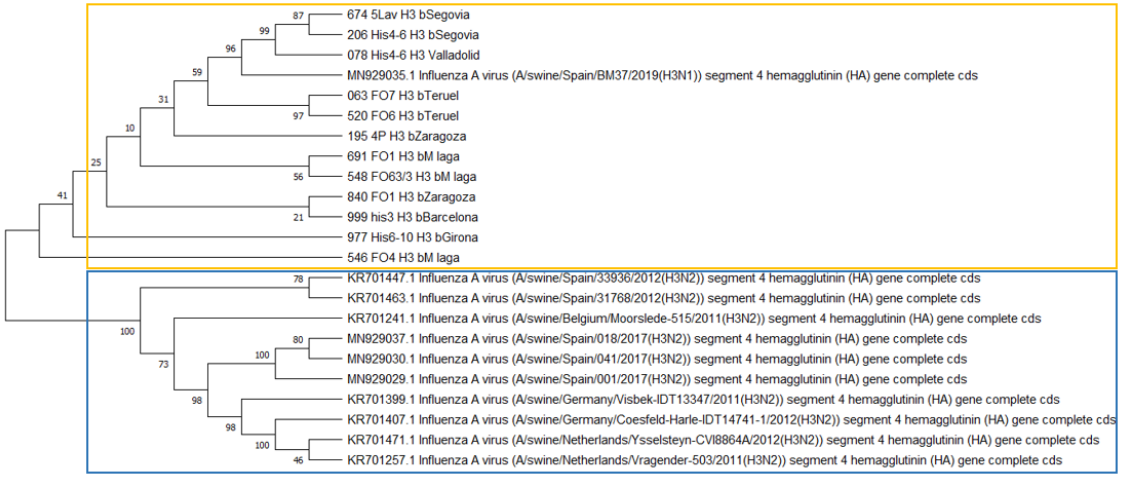


IMAGEN 20. Árbol inferido de las secuencias del subtipo H3 mediante el método ML, y el modelo de evolución Hasegawa-Kishano-Yano 1985 aplicando el parámetro gamma = 2, *bootstrap* = 1000. En amarillo se agrupan las secuencias obtenidas de muestras clínicas recibidas en Exopol, y en azul las secuencias de las cepas de referencia.

En los árboles calculados para el subtipo N1 (imágenes 21 y 22), podemos observar que existen diferencias en algunos nodos y grupos monofiléticos. Estas diferencias suelen coincidir con un bajo número de coincidencias en los pseudorreplicados calculados por *bootstrap*. Por ejemplo, KR69990 y MN932220, en General Reversible en el Tiempo 1986 (GTR) forman un grupo monofilético, sin embargo, en NJ la cepa MN932220 es una rama separada pero que proviene de un nodo que representa a un antepasado común a KR69990. En el caso de las secuencias de muestras clínicas, se puede ver que, en el árbol obtenido con ML, se forman algunos grupos monofiléticos con las secuencias 138 Zaragoza y 321 Huesca. Dada la proximidad geográfica de las dos provincias, es muy probable que ambas cepas tengan un mismo antecesor; pero los valores bajos de *bootstrap* indican que tampoco es tan seguro y, por este motivo, es aceptable que en NJ ese grupo no se haya formado.

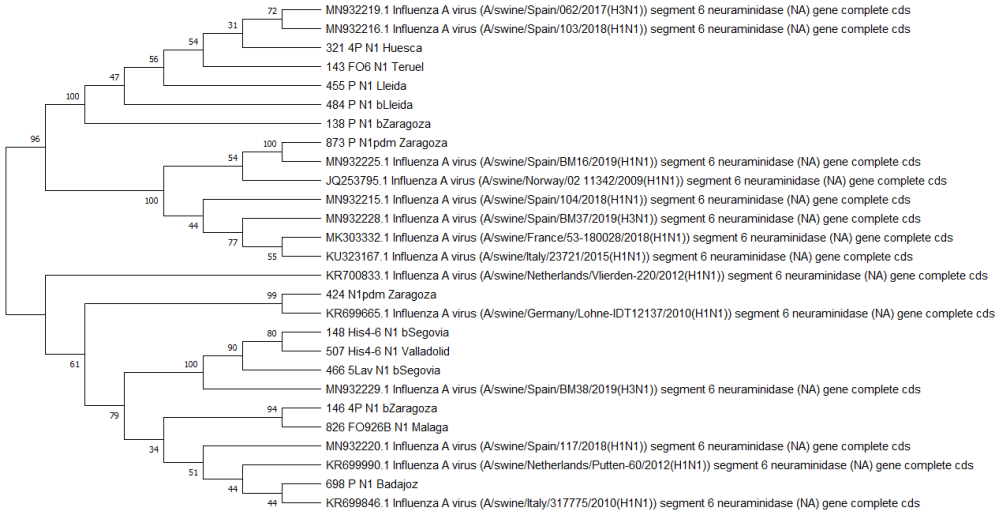


IMAGEN 21. Árbol inferido de las secuencias del subtipo N1 mediante el método NJ, y el modelo de evolución Tamura 1992 aplicando el parámetro gamma = 0,39 y *bootstrap* = 1000.

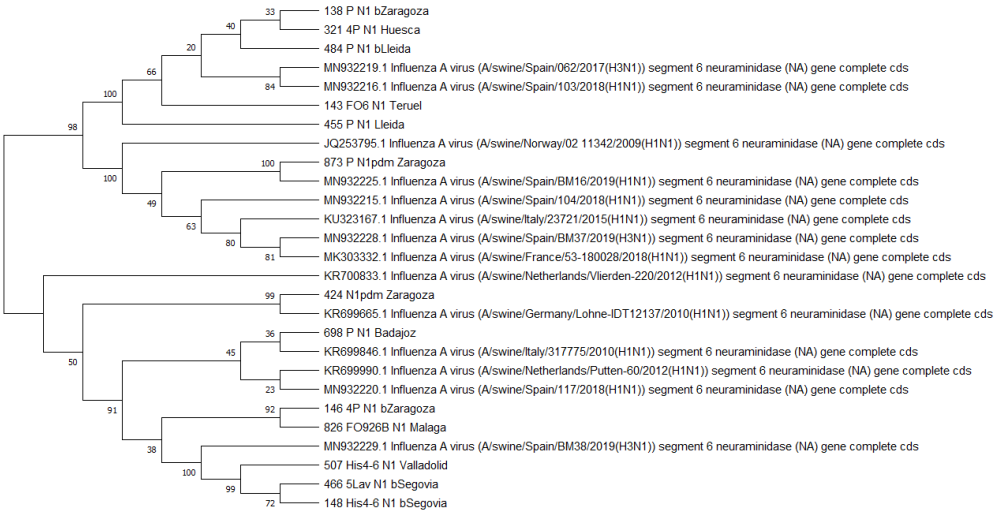


IMAGEN 22. Árbol inferido de las secuencias del subtipo N1 mediante el método ML, y el modelo de evolución General Reversible en el Tiempo 1986 aplicando el parámetro gamma = 2, *bootstrap* = 1000.

Para el subtipo N2, la forma general de los árboles realizados con NJ (imagen 23) y ML (imagen 24) es muy similar, pero se observan diferencias en algunos nodos externos. Por ejemplo, en NJ las cepas 919 y 284 forman un grupo monofilético que en ML no forman. Lo mismo sucede con las secuencias 50 y 82, en ML sí que lo forman mientras que en NJ no. En ML la secuencia 749 y el clado formado con las secuencias 52 y KR700511 tienen un ancestro común y se separan en clados más recientes, sin embargo, en NJ aparece un ancestro común intermedio. En la parte inferior del árbol, los valores de *bootstrap* son muy bajos en NJ y, por este motivo, hay también diferencias con las secuencias 14 y 422. A pesar de ello, las secuencias de regiones geográficas próximas siguen apareciendo en el mismo clado, como 410 y 047 que proceden de Zaragoza, y, en otro clado, la 12 proveniente de Ávila y la 79 de Badajoz. Por ello, ambos árboles son bastante similares y, si además tenemos en cuenta que el tiempo que supuso realizar el árbol con ML fue más de cincuenta minutos, son asumibles las diferencias que se producen con NJ que necesitó menos de dos minutos.

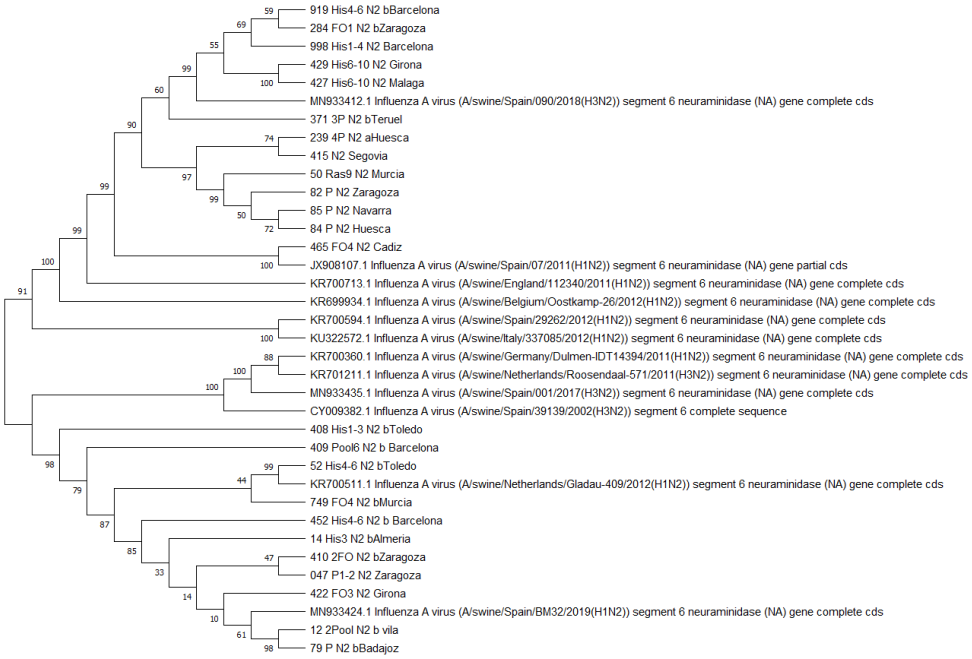


IMAGEN 23. Árbol inferido de las secuencias de N2 mediante el método NJ, y el modelo de evolución Tamura 1992 aplicando el parámetro gamma = 2,51 e I y *bootstrap* = 1000.

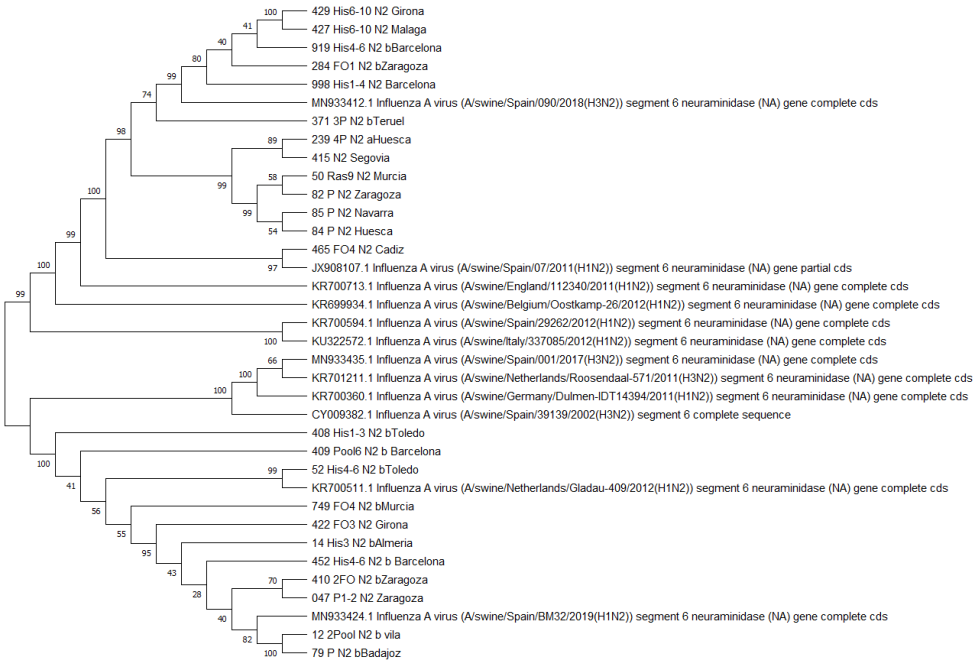


IMAGEN 24. Árbol inferido de las secuencias de N2 mediante el método ML, y el modelo de evolución General Reversible en el Tiempo 1986 aplicando el parámetro gamma = 2 e I, *bootstrap* = 1000.

Debido a que la intención de estudio de la filogenia de las secuencias de los genes HA y NA es implementar un código que permita realizar análisis filogenéticos rápidos, y que pueda servir para ambos genes, se ha estudiado la posibilidad de unificar el estudio de ambos genes con el modelo

de evolución Tamura-Nei 1993. Este modelo es una de las mejores opciones para los subtipos H1 y H3 de HA, y se ha evaluado la posibilidad de que sustituya el modelo Tamura 1992 en los subtipos N1 y N2 del gen NA. El modelo Tamura-Nei 1993 tiene en cuenta más parámetros que el de Tamura 1992, como se puede ver en la tabla 1. Además, en la implementación en R, para el cálculo de las distancias Tamura-Nei 1993 permite introducir el parámetro gamma, sin embargo, Tamura 1992 no.

En la tabla 10, se muestra el resultado de comparar los árboles inferidos con MEGA-X con el método NJ, aplicando el modelo de evolución Tamura 1992 y Tamura-Nei 1993, para cada uno de los subtipos N1 y N2. Se han tenido en cuenta los mismos aspectos clave que en la tabla 9. El resultado fue que los árboles inferidos con ambos modelos de evolución son equivalentes.

TABLA 10. Resultados de la comparativa de los árboles filogenéticos inferidos de las secuencias de los subtipos N1 y N2 de Influenza A porcino con MEGA-X mediante NJ y los modelos de evolución Tamura 1992 y Tamura-Nei 1993.

	N1	N2
Agrupación de las secuencias en grandes clados	Sí, en tres clados	Sí, en dos clados
Las secuencias que forman cada uno de los clados principales son las mismas	Sí	Sí
Los clados coinciden en al menos un 50% de ellos	Sí, en un más de un 90%	Sí, en más de un 90%
Las diferencias se suelen asociar a valores bajos de *bootstrap*	Sí	Sí
Aparecen los clados iguales de los nodos con valores altos de *bootstrap* (100-95)	Sí	Sí
Son equivalentes	Sí	Sí

Si se comparan los árboles del subtipo N1 con las imágenes 21 y 25, son prácticamente iguales, solo se diferencia en el clado formado por las secuencias 146 y 826 que en el árbol calculado con el modelo Tamura-Nei 1993 proviene de un ancestro común a MN932229, 466, 148 y 507; y, sin embargo, en Tamura 1992 aparece como que proviene de un ancestro común al clado que aparece al lado, al formado por MN932220, KR69990, 698, KR699846. En ambos árboles, el nodo del que proviene esta discrepancia tiene valores de *bootstrap* bajos 34 y 39, por tanto, es normal esta diferencia y se pueden considerar árboles equivalentes.

IMAGEN 25. Árbol inferido de las secuencias del subtipo N1 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 0,39 y *bootstrap* = 1000.

En el caso del subtipo N2, los árboles obtenidos mediante Tamura-Nei 1993 (imagen 26) y Tamura 92 (imagen 23) son iguales, y los valores de *bootstrap* son muy similares, por lo tanto, se puede asumir que para este subtipo ambos árboles son equivalentes. Aunque no debemos olvidar las diferencias que se presentaban entre los árboles inferidos, según Tamura 92 y General Reversible en el Tiempo.

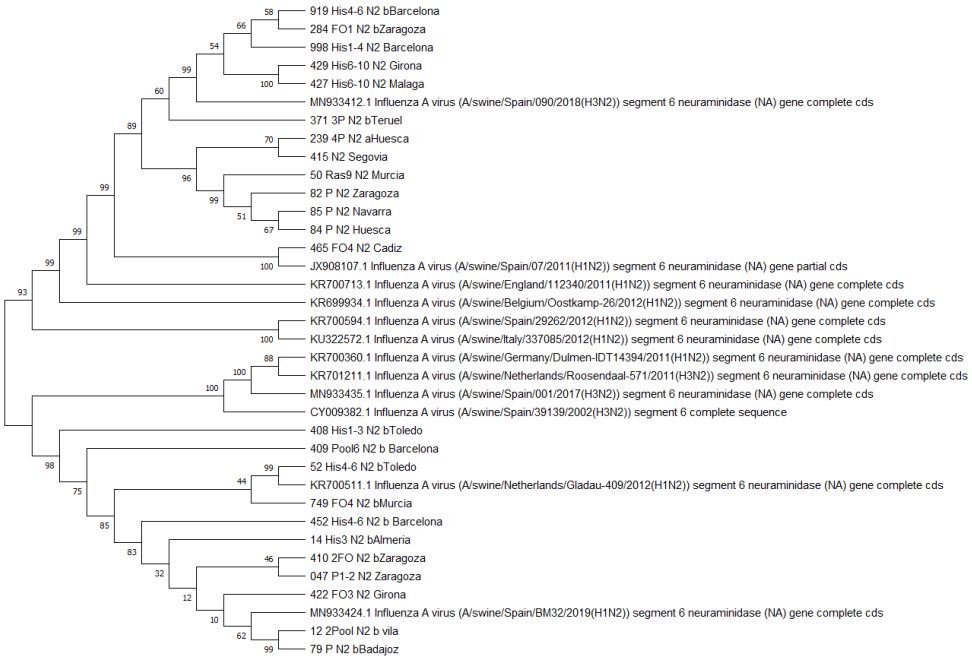


IMAGEN 26. Árbol inferido de las secuencias de N2 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 2,52 e I, y *bootstrap* = 1000.

7.1.3. Implementación en R

La dificultad más importante que se encontró fue que las secuencias no son todas de la misma longitud. Esto dificultó la producción de la matriz de distancias evolutivas, ya que todas las funciones que se encontraron tenían ese requisito.

Finalmente, se optó por realizar, primero, un alineamiento múltiple, y con este alineamiento calcular, después, la matriz de evolución. Aunque para elaborar árboles en R este paso no es necesario, fue la única solución que funcionó. También, se intentó con la función *blocks*, que completa con *gaps* los extremos de las secuencias hasta dejarlas todas con la misma longitud, pero devolvía la variable vacía y, por tanto, se descartó.

En cuanto al tipo de alineamiento múltiple, teniendo en cuenta que para estas secuencias no influía en la inferencia de los árboles, se optó por MUSCLE debido a que ofrece una mayor velocidad y precisión de alineación en comparación con los programas disponibles actualmente, como Clustal Omega [85].

Para la implementación en R, se tuvo en cuenta los listados de modelos de evolución obtenidos del estudio elaborado con MEGA-X, pero también otros aspectos del método para articular los árboles. Finalmente, se eligió el método NJ para inferir los árboles filogenéticos por varios motivos:

- NJ es muy rápido y computacionalmente eficiente, suele tardar milisegundos. Se puede aplicar a estudios evolutivos de entidades cercanas como es este caso, que son secuencias del mismo virus. También es consistente, es decir, que siempre estima de la misma manera el árbol.
- Otra de las alternativas que se descartó fue Máxima Parsimonia (MP). Este método dibuja un árbol en función del menor número de cambios, por lo que no incluye todas las posibles relaciones evolutivas. Además, este método no es consistente lo que podría generar discrepancias en distintas versiones de los informes clínicos. También es relativamente lento y necesita más potencia de cálculo que NJ.
- La principal alternativa que se planteó fue el método de Máxima verosimilitud (ML) ya que es el más eficiente de los tres, es decir, es el que tiene más probabilidades de inferir el árbol correcto, siendo también el más consistente. Sin embargo, es computacionalmente exigente y, por tanto, la velocidad computacional es mucho más lenta que la de NJ. ML puede llegar a tardar más de cuarenta minutos en realizar el árbol de las secuencias estudiadas en este proyecto.
- Además, el análisis comparativo de los árboles realizados con MEGA-X en este proyecto, ha mostrado que las diferencias obtenidas entre los árboles inferidos con ML y NJ no son

demasiado grandes y pueden considerarse equivalentes, por tanto, apoya el uso de NJ en la implementación en R.

Para el cálculo de la matriz de distancias, se ha tenido en cuenta el modelo de evolución obtenido mediante el estudio elaborado con MEGA-X, así como el parámetro gamma. Se intentó implementar el modelo Tamura 1992 con el parámetro gamma, pero la función *dist.dna* no permite introducir ese parámetro para este modelo, así que se unificó el modelo para todos los subtipos con Tamura-Nei 1993. Se aplicó un valor de gamma de 1,40, que es la media de los valores obtenidos para cada subtipo.

Otro de los aspectos que se tuvieron en cuenta fue el enraizamiento. En la implementación original, el enraizamiento se realizaba por el punto medio; tras obtener el árbol mediante NJ, ahora, al incluir el *bootstrap*, suponía un problema. La función *bootstrap* utilizada, comparaba un árbol realizado por NJ con otros 100 realizados por NJ también, con lo que se obtiene el número de veces que ha salido ese mismo nodo. La función que utiliza para realizar los árboles es una función diferente a la que hace el *bootstrap*; por lo que se incluyó el enraizamiento por la «mitad» dentro de la función que hace el árbol NJ. De ese modo, todos los árboles que comparaba estaban enraizados igual y no habría discrepancias a la hora de colocar en los nodos correspondientes la frecuencia de recurrencia de ese nodo.
Otra de las opciones de enraizamiento es la elección de un *outgroup* como raíz, pero se elige manualmente tras ver el resultado del árbol, por lo que, al ser un código que se va a ejecutar desde *FileMaker* y que no se va a manipular, se descartó esta opción.
Se decidió realizar el *bootstrap* 100 veces para ahorrar coste computacional y poder obtener el árbol en menos tiempo. Aunque no hay que olvidar, para futuras mejoras, que realizando 1000 árboles pseudorreplicados, se puede valorar la confianza del árbol con más información.

La intención era poder realizar un árbol consenso, pero no se encontró una función que uniese los formatos que se estaban trabajando y el que se necesitaba para devolver a *FileMaker* y poder representar los árboles. Además, para este árbol consenso no correspondían las posiciones de los nodos con las frecuencias que se obtenían del *bootstrap*, ya que eran funciones de distintas librerías que se usaban para desarrollos distintos, y no estaban correlacionadas. Finalmente, se decidió dejar la opción que funcionaba, es decir, representar el árbol frente al que se comparaban los otros 100, y mostrar en los nodos los valores obtenidos de *bootstrap*.

7.1.4. Comparativa de los árboles obtenidos con R y con MEGA-X

En la tabla 11, se muestra el resultado de comparar los árboles inferidos con MEGA-X y R con NJ y aplicando el modelo de evolución Tamura-Nei 1993, para cada uno de los subtipos H1, H3, N1 y N2. Se han tenido en cuenta diversos aspectos clave, como la agrupación general de las

secuencias en los clados principales. También se han valorado las diferencias, si se han formado los mismos grupos monofiléticos, si cada clado se ha formado en la misma posición y con las mismas secuencias. Si se han encontrado diferencias, se valoraron si los valores de *bootstrap* del nodo del antecesor eran bajos en ambos árboles, y podía ser la causa de la variación. Se ha comprobado que los nodos que tenían valores entre 100 y 95 de *bootstrap* aparecían en ambos árboles, y formaban clados con las mismas secuencias y con la misma relación evolutiva. Finalmente se evaluó que los árboles eran equivalentes.

TABLA 11. Resultados de la comparativa de los árboles filogenéticos inferidos de las secuencias de los subtipos H1, H3, N1 y N2 de Influenza A porcino con MEGA-X y R mediante el método Unión de vecinos y el modelo de evolución Tamura-Nei 1993.

	H1	H3	N1	N2
Agrupación de las secuencias en grandes clados	Sí, en tres clados que diferencian los tres linajes: humano, pandémico y aviar	Sí, en dos grandes clados	Sí, en tres clados	Sí, en dos clados
Las secuencias que forman cada uno de los clados principales son las mismas	Sí	Sí	Sí	Sí
Los clados coinciden en al menos un 50% de ellos	Sí, en un 50%	Sí, en un 80%	Sí, en un 50%	Sí, en un 50%
Las diferencias se suelen asociar a valores bajos de *bootstrap*	Sí	Sí	Sí	Sí
Aparecen los clados iguales de los nodos con valores altos de *bootstrap* (100-95)	Sí	Sí	Sí	Sí
Son equivalentes	Sí	Sí	Sí	Sí

Con más detalle se puede ver que, en el árbol inferido con R de las secuencias del subtipo H1 (imagen 27), aparecen los tres clados correspondientes a los tres linajes: pandémico, aviar y humano. El clado del linaje pandémico es muy similar al obtenido con MEGA-X (imagen 17), se presentan los grupos monofiléticos formados por las secuencias 421 de Zaragoza y la 140 de Ávila, también el de MN932198 y 733 de Zaragoza del linaje pandémico. En el clado del linaje aviar, es donde se genera más variación y donde los valores de *bootstrap* son más bajos. En clado del linaje humano sucede lo mismo, está representado el grupo monofilético formado por

KR699916 y KR699821, pero el resto muestra variaciones y con valores de *bootstrap* bajos. Sin embargo, al menos el 50% de los clados coinciden y los más probables aparecen en ambos árboles.

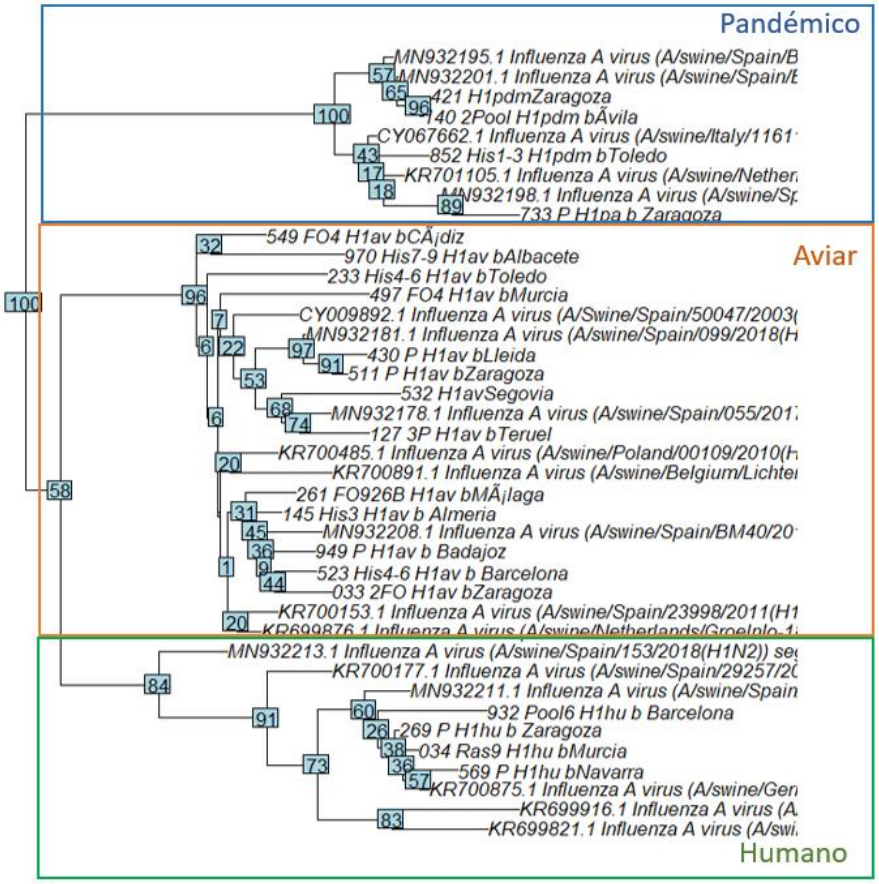


IMAGEN 27. Árbol inferido con R de las secuencias de H1 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 1,40 y *bootstrap* = 100. En azul, se ha marcado el clado del linaje pandémico, en naranja, el linaje aviar y, en verde, el linaje humano.

En los árboles inferidos con MEGA-X (imagen 19) y con R (imagen 28) para el subtipo H3, en ambos se obtienen los dos clados bien diferenciados. Por un lado, el que contiene las secuencias de muestras clínicas recibidas en Exopol y la secuencia de referencia MN929035 y, en el otro clado, el resto de las secuencias de referencia. Los nodos externos son prácticamente iguales en ambos árboles. Así que ambos son equivalentes.

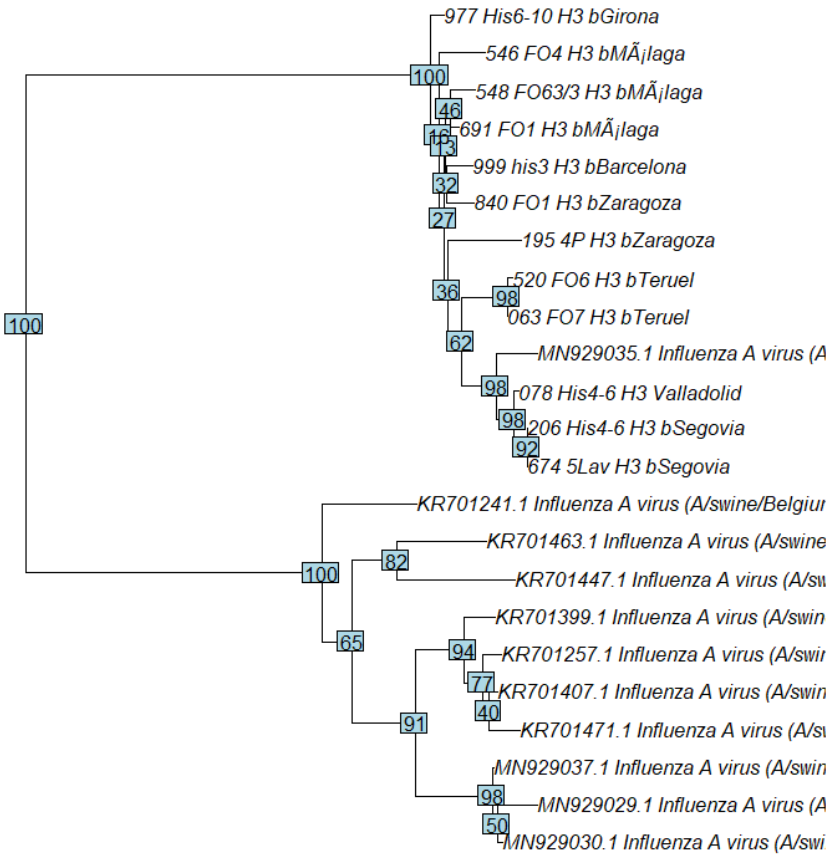


IMAGEN 28. Árbol inferido con R de las secuencias de H3 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 1,40 y *bootstrap* = 100.

Para el subtipo N1 –si se compara el árbol de la imagen 29 con el de la imagen 25, son muy similares–, se aprecian los dos clados más grandes. El clado más grande de la parte superior del árbol realizado con R, está rotado respecto del elaborado con MEGA-X. Del nodo que aparece en la parte superior en la imagen 29, divergen dos clados uno con la secuencia del linaje pandémico 873 procedente de Zaragoza que forma un clado externo en ambos con MN932225, como el inferido con MEGA-X. Hay diferencias en las topologías más externas, pero en ambos árboles los valores de *bootstrap* son bajos y esto puede producir esta variabilidad. En el clado que aparece en la parte central del árbol de la imagen 29 elaborado con R, se ha formado un grupo monofilético con las secuencias 138 y 484 que en la imagen 25 no aparece. En el clado inferior, la única diferencia es el clado formado por las secuencias 146 y 826 que ha cambiado a la rama de al lado, pero que proviene del mismo antecesor. En términos generales, el árbol obtenido con R es bastante similar al calculado con MEGA-X; las diferencias son asumibles debido a los bajos valores de *bootstrap* obtenidos con ambos en algunos de los nodos en los que se han producidos los cambios.

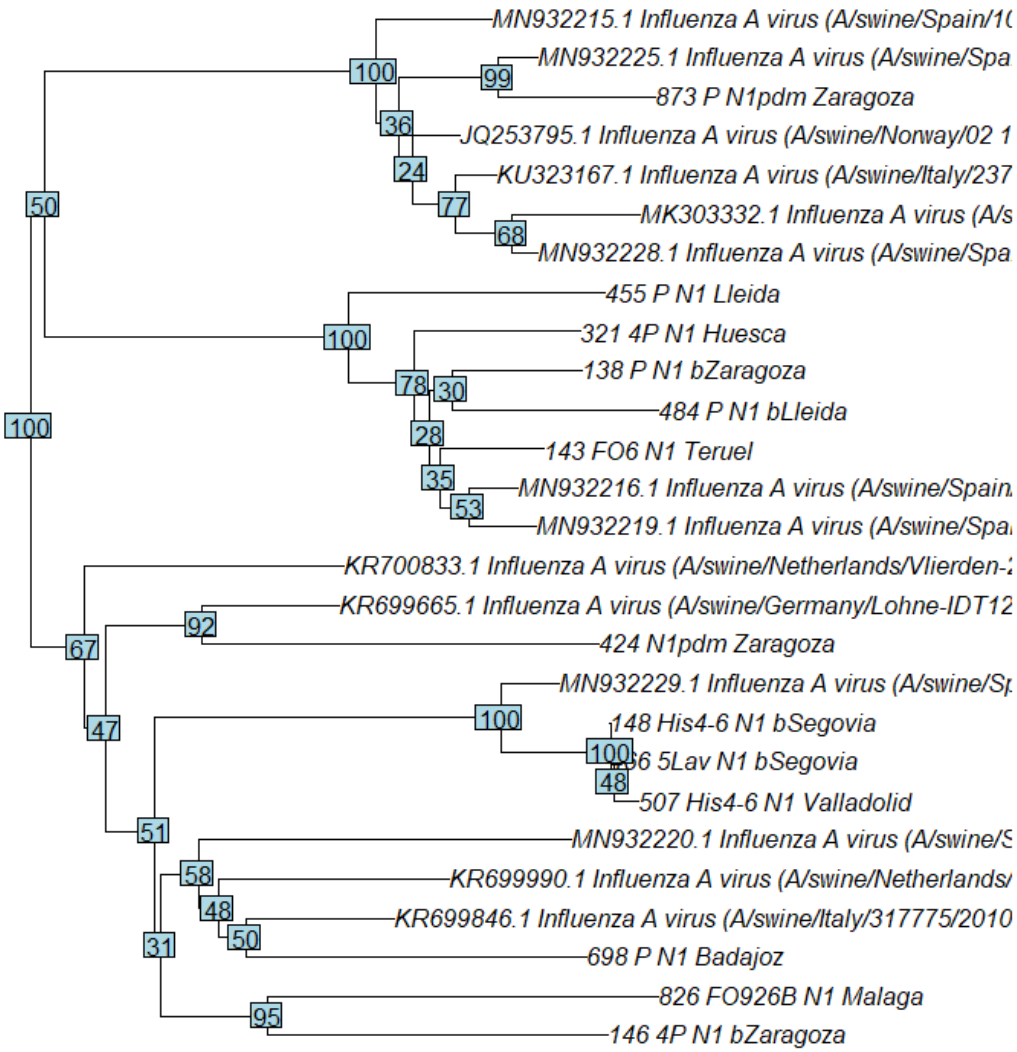


IMAGEN 29. Árbol inferido con R de las secuencias de N1 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 1,40 y *bootstrap* = 100.

En el árbol inferido con R (imagen 30) de secuencias del subtipo N2, se ven dos clados diferenciados como en el inferido con MEGA-X (imagen 26). Si se observa el clado de la parte superior del árbol de la imagen 26, aparece como en el clado de la parte de abajo de la imagen 30; estas rotaciones son normales y aceptables. Los grupos monofiléticos formados por KR700511 y 52, y el clado formado con la misma topología para las secuencias CY009382, MN933435, KR701211, KR700594 aparecen igual en los dos árboles. En el resto del clado superior, los valores de *bootstrap* son bajos en ambos y, por ello, hay variabilidad.

En el clado inferior del árbol obtenido con R, se pueden ver los grupos monofiléticos formados por: las secuencias JX908107 y 465 de Cádiz; también, el formado por las secuencias 415 de

Segovia y 239 de Huesca; 429 de Málaga y 428 de Gerona. De igual modo, se observa la misma bifurcación en el nodo con 20 de *bootstrap* en el árbol hecho con R, por un lado, las secuencias de 84, 82, 85, 50, 239, 415 y, por el otro, las de 427, 429, 284,919, MN933412, 988; ocurre de la misma manera en el elaborado con MEGA-X solo diferenciándose en la posición de la secuencia 371 de Teruel.

Por tanto, aunque existen diferencias entre ambos árboles, las similitudes son muchas y, teniendo en cuenta que ni el modelo de evolución ni el método de construir los árboles son los más óptimos para estas secuencias, el resultado es razonable.

También, hemos de observar que, en el día a día para realizar el diagnóstico en Exopol, no se van a procesar árboles con tantas secuencias: se suelen comparar secuencias provenientes de cepas obtenidas de un mismo cliente con alguna cepa de referencia, por lo que los clados van a ser más claros y probablemente tengan valores de *bootstrap* más elevados, por tanto habrá menos variabilidad e incertidumbre en los árboles resultantes con esta implementación en R.

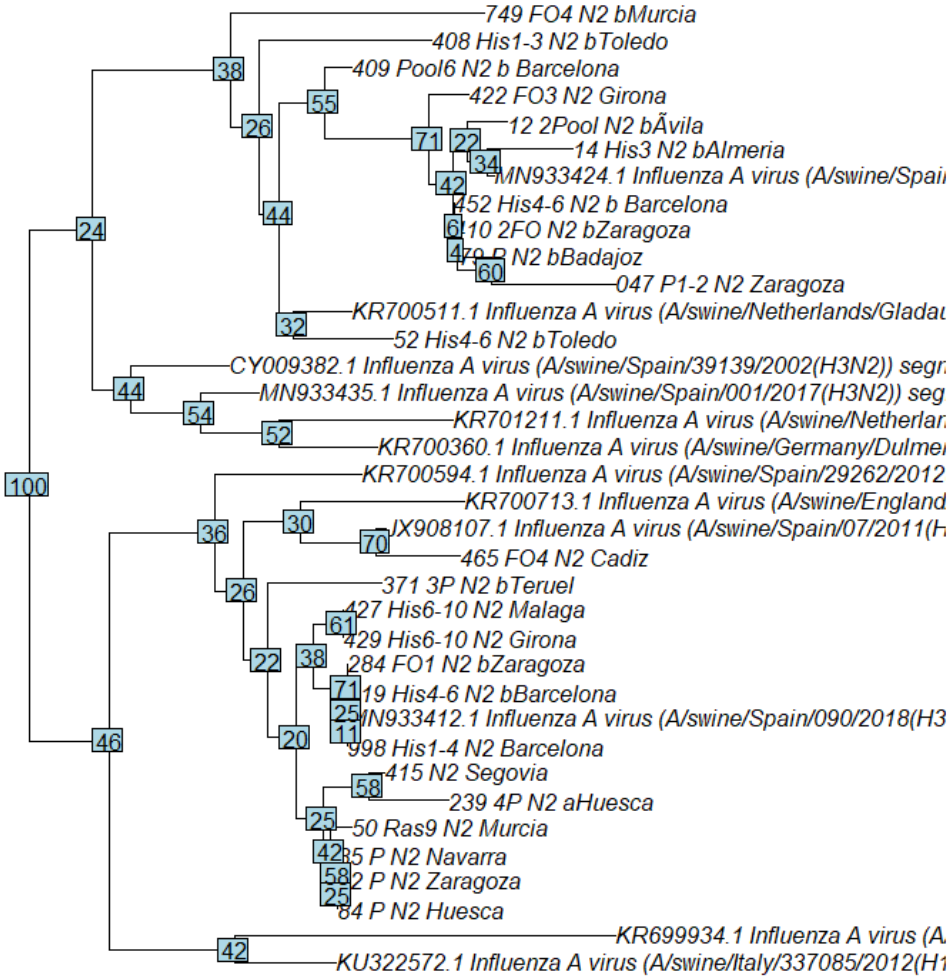


IMAGEN 30. Árbol inferido con R de las secuencias de N2 mediante el método NJ, y el modelo de evolución Tamura-Nei 1993 aplicando el parámetro gamma = 1,40 y *bootstrap* = 100.

En general, la evaluación de los árboles resultantes es óptima, ya que los clados principales siempre han aparecido, así como los nodos externos importantes y con valores de *bootstrap* altos. Para facilitar el trabajo diario, es aceptable perder precisión al intentar unificar, pero no se puede olvidar que se han encontrado discrepancias en la comparación de los árboles inferidos con R y los inferidos con MEGA-X. Estas diferencias pueden atribuirse al hecho de no haber podido representar un árbol consenso con R.

De momento, con esta implementación se ha conseguido elaborar un árbol aplicando un modelo de evolución apropiado, además de ver la probabilidad de que el árbol que se muestra sea así realmente.

Una de las posibles mejoras sería efectuar una paralelización para realizar el *bootstrap* o el alineamiento múltiple y que sea el proceso más rápido, incluso, poder plantear ML como método para inferir árboles. Si se realizasen los árboles con ML, también sería necesario realizar *bootstrap,* lo que supone un proceso muy costoso, así que, paralelizarlo resultaría una buena opción para poder llevarlo a cabo en un tiempo razonable. Con ML, se podría seguir estudiando si es posible implementar que se infiera un árbol consenso, porque las librerías que se utilizan son distintas.

7.2. PRRS

En la tabla 12, se muestran los resultados obtenidos de porcentaje de identidad en un caso de PRRS, en cerdos lactantes de agosto de 2021. Si se comparan los resultados obtenidos mediante las mejoras realizadas en R y los obtenidos mediante LALIGN, tanto para secuencias de nucleótidos como para aminoácidos, los resultados son prácticamente iguales. Los resultados se diferencian, solamente en algunos de los casos, entre 0,1 - 0,3%, siendo esta diferencia insignificante. En el anexo 4, se pueden ver capturas de pantalla de los alineamientos realizados con LALIGN de la secuencia del nuevo caso con la número 12, tanto de las secuencias de nucleótidos como de su traducción a aminoácidos. Por tanto, con estos resultados se validaron los cambios realizados en el código y se procedió a instaurar el código en el sistema, con el propósito de ofrecer unos resultados más precisos a los clientes de Exopol.

Aunque se podría haber elegido una matriz de sustitución para secuencias de aminoácidos menos divergentes, como la BLOSUM90, o la BLOSUM62 que se aplica a divergencias intermedias, se decidió dejar la BLOSUM50, porque este código R también se utiliza para otro tipo de secuencias que se manejan en el laboratorio, y podría generar errores.

TABLA 12: Resultados de porcentaje de identidad obtenidos mediante el alineamiento de la secuencia del caso recibido en Exopol en agosto de 2021, con la secuencia de la cepa referencia de PRRS europeo (Lelystad), las principales cepas vacunales en España y cepas de otros casos remitidos por el cliente. Los resultados son de alineamientos calculados con las mejoras realizadas en R y LALIGN de secuencias de nucleótidos y su traducción a aminoácidos.

Comparación de secuencias	**Identidad (%)**			
	Mejoras realizadas en R		**LALIGN**	
Secuencias de referencia	**Nucleótidos**	**Aminoácidos**	**Nucleótidos**	**Aminoácidos**
Lelystad (M96262.2)	98,80	98,50	98,80	98,50
Pyrsvac (DQ345726.1)	94,10	91,60	94,10	91,50
Unistrain (DQ345725.1)	94,20	92,60	94,20	92,50
Porcillis-DV (KJ127878.1)	99,80	99,50	99,80	99,50
Reprocyc B.I. (KT988004.1)	89,60	87,10	89,60	87,10
Suvaxyn (MK876228.1)	88,80	88,60	88,80	88,60
Cepas de otros casos remitidos por el cliente	**Nucleótidos**	**Aminoácidos**	**Nucleótidos**	**Aminoácidos**
1	86,60	84,90	86,60	84,90
2	86,20	85,60	86,10	85,60
3	85,30	83,70	85,30	84,00
4	88,40	87,60	88,40	87,60
5	86,30	86,10	86,30	86,10
6	88,00	87,10	88,00	87,10
7	85,70	85,60	85,60	85,60
8	94,10	93,60	94,10	93,50
9	94,20	93,60	94,20	93,50
10	94,20	93,60	94,20	93,50
11	85,80	86,60	85,80	86,60
12	94,20	94,10	94,20	94,00

8. Conclusiones

En base a los resultados obtenidos en este proyecto, y teniendo en cuenta los objetivos planteados, se puede concluir lo siguiente:

8.1. Conclusiones del estudio filogenético de Influenza A porcino e implementación en R:

- Para la realización de los árboles filogenéticos de las secuencias de los subtipos H1 y H3 del gen HA, y las secuencias de los subtipos N1 y N2 del gen NA del virus influenza A porcino, no influye en los árboles resultantes el método de alineamiento múltiple MUSCLE o Clustal Omega. Por tanto, se procedió a la utilización de MUSCLE en favor de un menor coste computacional y una mejor precisión en comparación con Clustal Omega.
- Para las secuencias del subtipo H1 se obtuvo que el modelo más apropiado es General Reversible en el Tiempo + G + I utilizando el método de Máxima verosimilitud. Pero se puede ahorrar coste computacional utilizando el método de Unión de vecinos y el modelo de evolución Tamura-Nei 1993 +G +I, ya que, los árboles obtenidos de ambas formas son muy similares y el tiempo de cálculo es más reducido.
- Para las secuencias del subtipo H3 se obtuvo que el modelo más apropiado es Hasegawa-Kishano-Yano 1985 + G utilizando el método de Máxima verosimilitud. Pero se puede ahorrar coste computacional utilizando el método de Unión de vecinos y el modelo de evolución Tamura-Nei 1993 +G, ya que, los árboles obtenidos de ambas formas son muy similares.
- Para las secuencias del subtipo N1 el modelo más apropiado es General Reversible en el Tiempo + G utilizando el método de Máxima verosimilitud. Pero se puede ahorrar coste computacional utilizando el método de Unión de vecinos y el modelo de evolución Tamura 1992 +G, ya que, los árboles obtenidos de ambas formas son muy similares.
- Para las secuencias del subtipo N2 el modelo más apropiado es General Reversible en el Tiempo + G + I utilizando el método de Máxima verosimilitud. Pero se puede ahorrar coste computacional utilizando el método de Unión de vecinos y el modelo de evolución Tamura 1992 +G +I, ya que, los árboles obtenidos de ambas formas son muy similares.
- Los árboles obtenidos para los subtipos N1 y N2 con el modelo de evolución Tamura-Nei 1993 son equivalentes a los obtenidos con el modelo Tamura 1992, por lo tanto, es posible utilizar para todos los subtipos H1, H3, N1 y N2 el modelo de evolución, Tamura-Nei 1993.
- Para la implementación en R se aplicó, se unificó en un solo código el análisis de los cuatro subtipos de influenza, el modelo de evolución Tamura-Nei 1993 con el parámetro gamma = 1,40 y el método de elaboración de árboles de Unión de vecinos. El resultado del cálculo de los árboles, se programó para que fuese en formato *newick* y que pudiera

leerse con *FileMaker* para graficarlo. En este formato se incluyeron los valores de *bootstrap* de cada nodo del árbol.

- Los árboles obtenidos con el código implementado en R son muy similares a los obtenidos con el software MEGA-X y, por tanto, se valida el código implementado, pero con algunas limitaciones:
    - No se consiguió implementar en R el cálculo del árbol consenso y, por tanto, es necesario seguir trabajando en esta mejora.
    - Debido al coste computacional no se pudo implementar el método de Máxima verosimilitud y los modelos de evolución aplicables a este método que son los más precisos. Para mejorar el coste computacional es necesario investigar en una ampliación de este proyecto la vía de la paralelización.

8.2. Conclusiones de la selección de los parámetros para el estudio de la identidad de secuencias de PRRS y su implementación en R:

- Se seleccionó el alineamiento local mediante el algoritmo de Smith-Waterman al ser el más apropiado para comparar secuencias de diferente longitud.
- Los valores óptimos para el cálculo de la identidad mediante los alineamientos de pares de secuencias de ADN fueron: *match*: +5; *mismatch*: -4; *gap opening*: -12; *gap extension*: 0 y para secuencias de aminoácidos fueron: matriz BLOSUM50, *gap opening:* -12; *gap extension:* -2.
- En la implementación en R además de incluir los parámetros del punto anterior, fue necesario implementar la matriz de sustitución para las secuencias de ADN teniendo en cuenta los nucleótidos ambiguos recogidos por la IUPAC (*International Union of Pure and Applied Chemistry*).
- Es necesario estudiar en un futuro si es necesario cambiar la matriz BLOSUM50 por otra para secuencias menos divergentes.
- Los resultados obtenidos con la implementación en R de los alineamientos de pares de secuencias de ADN y de aminoácidos comparados con los de LALIGN son iguales y, por tanto, la implementación es válida. Debido a ello, se decidió utilizar este código para calcular los resultados de identidad que se incluyen en los informes clínicos elaborados por Exopol.

9. Bibliografía

[1] Pevsner J. Bioinformatics and Functional Genomics. vol. 3. 3ª Edición. Oxford: John Wiley & Sons; 2015. https://doi.org/10.1093/bfgp/3.2.187.

[2] Reeck G. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. Cell 1987; 50:667–667. https://doi.org/10.1016/0092-8674(87)90322-9.

[3] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. The EMBO Journal, 1986; 5:823–6. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.

[4] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology, 1990; 215: 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

[5] Bleidorn C. Phylogenomics: An introduction, Springer, 2017. ISBN: 978-3-319-54062-7. https://doi.org/10.1007/978-3-319-54064-1.

[6] Dayhoff M. Atlas of protein sequence and structure. Supplement. Washington D.C.: National Biomedical Research Foundation, 1978.

[7] Henikoff S. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America, 1992; 89.

[8] Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, et al. Glocal alignment: Finding rearrangements during alignment. Bioinformatics, 2003; 19. https://doi.org/10.1093/bioinformatics/btg1005.

[9] Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 1970; 48.

[10] Smith T, Waterman M. Identification of common molecular subsequences. Journal of Molecular Biology 1981;147.

[11] Morgenstern B. Local versus global alignments. Sequence Alignment: Methods, Models, Concepts, and Strategies, 2009:39–54. https://doi.org/10.1525/CALIFORNIA/9780520256972.003.0003.

[12] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America, 1992; 89:10915–9. https://doi.org/10.1073/PNAS.89.22.10915.

[13] Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nature Biotechnology, 2004; 22:1035–6. https://doi.org/10.1038/NBT0804-1035.

[14] Huang X, Miller W. A time-efficient, linear-space local similarity algorithm. Advances in Applied Mathematics, 1991; 12:337–57. https://doi.org/10.1016/0196-8858(91)90017-D.

[15] LALIGN Pairwise Sequence Alignment EMBL-EBI, https://www.ebi.ac.uk/Tools/psa/lalign/; Acceso: 8 Septiembre 2021.

[16] Godini R, Fallahi H. A brief overview of the concepts, methods and computational tools used in phylogenetic tree construction and gene prediction. Meta Gene, 2019; 21:100586. https://doi.org/10.1016/j.mgene.2019.100586.

[17] Mount DW. Bioinformatics: Sequence and Structural Analysis. 2ª Edición. Cold Spring Harbor Laboratory Press; 2004.

[18] Gupta SK, Kececioglu JD, Schäffer AA. Improving the Practical Space and Time Efficiency of the Shortest-Paths Approach to Sum-of-Pairs Multiple Sequence Alignment. Journal of Computational Biology, 1995; 2:459–72. https://doi.org/10.1089/cmb.1995.2.459.

[19] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W (improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice). Nucleic Acids Research, 1994; 22:4673–80. https://doi.org/10.1007/978-1-4020-6754-9_3188.

[20] Notredame C, Higgins DG, Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology, 2000; 302:205–17. https://doi.org/10.1006/jmbi.2000.4042.

[21] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 2013; 30:772–80. https://doi.org/10.1093/molbev/mst010.

[22] Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004; 32:1792–7. https://doi.org/10.1093/nar/gkh340.

[23] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology, 1990; 215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

[24] Altschul SF, Madden TL, Shäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997; 273:3389–402. https://doi.org/10.1093/oxfordjournals.molbev.a026201.

[25] Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. BMC Biology, 2006; 4:1–14. https://doi.org/10.1186/1741-7007-4-41.

[26] Hall B, Hallgrimsson B. Strickberger's Evolution. 4ª Edición. Jones and Bartlett Publishers; 2008.

[27] Darwin C. El Origen de las especies. 1998th ed. Barcelona: S.L.U. ESPASA Libros; 1859.

[28] Huxley J. Evolution. The modern synthesis. London: George Alien & Unwin Ltd.; 1942.

[29] Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 2012; 13:303–14. https://doi.org/10.1038/NRG3186.

[30] Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983.

[31] Page RDM, Holmes EC. Molecular evolution: a phylogenetic approach, 1998, 346.

[32] Rodríguez F, Oliver JL, Marín A, Medina JR. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology, 1990; 142:485–501. https://doi.org/10.1016/S0022-5193(05)80104-3.

[33] Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequence. In: Miura RM, editor. Lectures on mathematics in the life science., vol. 17, American Mathematical Society; 1986, p. 57–86.

[34] Yang Z. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution, 1996; 11:367–72. https://doi.org/10.1016/0169-5347(96)10041-0.

[35] Yang F, Waldbieser GC, Lobb CJ. The Nucleotide Targets of Somatic Mutation and the Role of Selection in Immunoglobulin Heavy Chains of a Teleost Fish. The Journal of Immunology, 2006; 176:1655–67. https://doi.org/10.4049/JIMMUNOL.176.3.1655.

[36] Fitch WM, Margoliash E. Construction of phylogenetic trees. Science (New York, NY) 1967;155:279–84. https://doi.org/10.1126/SCIENCE.155.3760.279.

[37] Gu X, Fu YX, Li WH. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Molecular Biology and Evolution, 1995; 12:546–57. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040235.

[38] Sullivan J, Swofford DL, Naylor GJP. The Effect of Taxon Sampling on Estimating Rate Heterogeneity Parameters of Maximum-Likelihood Models. Mol Biol Evol, 1999; 16:1347–56.

[39] Jia F, Lo N, Ho SYW. The Impact of Modelling Rate Heterogeneity among Sites on Phylogenetic Estimates of Intraspecific Evolutionary Rates and Timescales. PLOS ONE, 2014; 9:e95722. https://doi.org/10.1371/JOURNAL.PONE.0095722.

[40] Kück P, Mayer C, Wägele J-W, Misof B. Long Branch Effects Distort Maximum Likelihood Phylogenies in Simulations Despite Selection of the Correct Model. PLOS ONE, 2012; 7:e36593. https://doi.org/10.1371/JOURNAL.PONE.0036593.

[41] Jones D, Taylor W, Thornton J. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences: CABIOS, 1992; 8:275–82. https://doi.org/10.1093/BIOINFORMATICS/8.3.275.

[42] Whelan S, Goldman N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and Evolution, 2001; 18:691–9. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A003851.

[43] Le SQ, Gascuel O. An Improved General Amino Acid Replacement Matrix. Molecular Biology and Evolution, 2008; 25:1307–20. https://doi.org/10.1093/MOLBEV/MSN067.

[44] Miyazawa S. Superiority of a mechanistic codon substitution model even for protein sequences in Phylogenetic analysis. BMC Evolutionary Biology, 2013; 13:1–10. https://doi.org/10.1186/1471-2148-13-257.

[45] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution, 1994; 11:725–36. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040153.

[46] Ren W, Vanden-Eijnden E, Maragakis P, E W. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. The Journal of Chemical Physics ,2005; 123:134109. https://doi.org/10.1063/1.2013256.

[47] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 1987; 4:406–25. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040454.

[48] Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H. An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics. Systematic Biology, 2005; 54:743–57. https://doi.org/10.1080/10635150500234609.

[49] Steel M, Penny D. Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. Molecular Biology and Evolution, 2000; 17:839–50. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A026364.

[50] Felsenstein J. Statistical Inference of Phylogenies. Journal of the Royal Statistical Society: Series A (General), 1983; 146:246–62. https://doi.org/10.2307/2981654.

[51] Huelsenbeck JP. Performance of Phylogenetic Methods in Simulation. Systematic Biology, 1995; 44:17–48. https://doi.org/10.1093/SYSBIO/44.1.17.

[52] Larget B, Weiss RE, Simon DL. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees Cite this paper Related papers Bayesian Select ion of Cont inuous-T ime Markov Chain Evolutionary Models Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of . Mol Biol Evol, 1999; 16:750–9.

[53] Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews Genetics, 2003; 4:275–84. https://doi.org/10.1038/nrg1044.

[54] Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Systematic Biology, 2012; 61:539–42. https://doi.org/10.1093/SYSBIO/SYS029.

[55] Efron B. The. Jacknife, the Bootstrap and Other Resampling Plans. Philadelphia: Society for Industrial and Applied Mathematics (SIAM); 1982.

[56] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 1985; 39:783–91. https://doi.org/10.1111/J.1558-5646.1985.TB00420.X.

[57] Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. How Many Bootstrap Replicates Are Necessary? Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2009; 5541 LNBI:184–200. https://doi.org/10.1007/978-3-642-02008-7_13.

[58] Anderson T, Macken C, Lewis N, Al. E. A Phylogeny-Based Global Nomenclature System and Automated Annotation Tool for H1 Hemagglutinin Genes from Swine Influenza A Viruses. MSphere, 2016; 1. https://doi.org/10.1128/MSPHERE.00275-16.

[59] Sosa Portugal S, Cortey M, Tello M, Casanovas C, Mesonero-Escuredo S, Barrabés S, et al. Diversity of influenza A viruses retrieved from respiratory disease outbreaks and subclinically infected herds in Spain (2017–2019). Transboundary and Emerging Disease,s 2021; 68:519–30. https://doi.org/10.1111/tbed.13709.

[60] Simon G, Larsen LE, Dürrwald R, Foni E, Harder T, van Reeth K, et al. European surveillance network for influenza in pigs: Surveillance programs, diagnostic tools and swine influenza virus subtypes identified in 14 European countries from 2010 to 2013. PLoS ONE, 2014; 9:1–21. https://doi.org/10.1371/journal.pone.0115815.

[61] Rambo-Martin BL, Keller MW, Wilson MM, Nolting JM, Anderson TK, Vincent AL, et al. Influenza A Virus Field Surveillance at a Swine-Human Interface. MSphere, 2020; 5:1–12. https://doi.org/10.1128/msphere.00822-19.

[62] Sebastian MR, Lodha R, Kabra SK. Swine origin influenza (swine flu). Indian Journal of Pediatrics, 2009; 76:833–41. https://doi.org/10.1007/s12098-009-0170-6.

[63] Simon-Grifé M, Martín-Valls GE, Vilar MJ, Busquets N, Mora-Salvatierra M, Bestebroer TM, et al. Swine influenza virus infection dynamics in two pig farms; results of a longitudinal assessment. Veterinary Research, 2012; 43:24. https://doi.org/10.1186/1297-9716-43-24.

[64] Influenza Research Database, https://www.fludb.org/brc/staticContent.spg?decorator=influenza&type=FluInfo&subtype=Mission Acceso: 20 Agosto 2021.

[65] Amarilla S, Avalos A, Suarez M. Síndrome reproductivo y respiratorio porcino: epidemiología, síntomas y lesiones. Compendio de Ciencias Veterinarias, 2015; 5(2):38–46. https://doi.org/10.18004/compend.cienc.vet.2015.05.02.38-46.

[66] Torrents D, Miranda J, Gauger P, Ramirez A, Linhares D. Effect of PRRSV stability on productive parameters in breeding herds of a swine large integrated group in Spain. Porcine Health Management, 2021; 7:1–7. https://doi.org/10.1186/s40813-021-00203-4.

[67] Mardassi H, Mounir S, Dea S. Identification of major differences in the nucleocapsid protein genes of a Québec strain and European strains of porcine reproductive and respiratory syndrome virus. Journal of General Virology, 1994; 75:681–5. https://doi.org/10.1099/0022-1317-75-3-681.

[68] Indik S, Valíček L, Klein D, Klánová J. Variations in the major envelope glycoprotein GP5 of Czech strains of porcine reproductive and respiratory syndrome virus. Journal of General Virology, 2000; 81:2497–502. https://doi.org/10.1099/0022-1317-81-10-2497.

[69] Prieto C, Vázquez A, Núñez JI, Álvarez E, Simarro I, Castro JM. Influence of time on the genetic heterogeneity of Spanish porcine reproductive and respiratory syndrome virus isolates. Veterinary Journal, 2009; 180:363–70. https://doi.org/10.1016/j.tvjl.2008.01.005.

[70] Mateu E, Martín M, Vidal D. Genetic diversity and phylogenetic analysis of glycoprotein 5 of European-type porcine reproductive and respiratory virus strains in Spain. Journal of General Virology, 2003; 84:529–34. https://doi.org/10.1099/vir.0.18478-0.

[71] GenBank n.d. https://www.ncbi.nlm.nih.gov/genbank/ (accessed September 8, 2021).

[72] Kumar S, Stecher G, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution, 2018; 35:1547–9. https://doi.org/10.1093/MOLBEV/MSY096.

[73] Paradis E, Blomberg B, Brown J. Package "ape" Title Analyses of Phylogenetics and Evolution Depends R 2021.

[74] Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. Bioinformatics, 2015; 31:3997–9. https://doi.org/10.1093/BIOINFORMATICS/BTV494.

[75] "stats" package – Rdocumentation, RCran and worldwide community, https://www.rdocumentation.org/packages/stats/versions/3.6.2 Acceso: 8 septiembre 2021.

[76] Bengtsson H. Various Programming Utilities, R package R.utils version 2.10.1, 2020.

[77] SD B, RA C, S B, MC L, J M-O, CJ V, et al. Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources, 2012; 12:562–5. https://doi.org/10.1111/J.1755-0998.2011.03108.X.

[78] Charif D, Lobry JR. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis 2007:207–32. https://doi.org/10.1007/978-3-540-35306-5_10.

[79] Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 2012; 3:217–23. https://doi.org/10.1111/J.2041-210X.2011.00169.X.

[80] ExPASy - Translate tool n.d. https://web.expasy.org/translate/ Acceso: 8 Septiembre 2021.

[81] Pagès H, Aboyoun P, Gentleman R, DebRoy S. Bioconductor - Biostrings: Efficient manipulation of biological strings 2021.

[82] Huber W, Carey J v, Gentleman R. Bioconductor - BiocGenerics: "Orchestrating high-throughput genomic analysis with Bioconductor." 2015.

[83] Ripley B, Tierney L, Urbanek S. Package "parallel", RDocumentation, 2011.

[84] Acer Store n.d. https://store.acer.com/es-es/acer-aspire-c-24-todo-en-uno-c24-963-zilver-1 (accessed August 20, 2021).

[85] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 2004; 5:1–19. https://doi.org/10.1186/1471-2105-5-113.

10. Anexos

10.1. Anexo 1: Secuencias del artículo de Sosa, S. 2020.

Secuencias del subtipo H1 con el código de acceso del GenBank:

- MN932201.1 Influenza A virus (A/swine/Spain/BM36/2019(H1N1)) segment 4 hemagglutinin (HA) gene, partial cds
- MN932195.1 Influenza A virus (A/swine/Spain/BM114/2018(H1N1)) segment 4 hemagglutinin (HA) gene, partial cds
- KR701105.1 Influenza A virus (A/swine/Netherlands/Barger-Compascuum-CVI6324A/2012(H1N2)) segment 4 hemagglutinin (HA) gene, complete cds
- CY067662.1 Influenza A virus (A/swine/Italy/116114/2010(H1N2)) segment 4 sequence
- MN932198.1 Influenza A virus (A/swine/Spain/BM16/2019(H1N1)) segment 4 hemagglutinin (HA) gene, partial cds
- MN932211.1 Influenza A virus (A/swine/Spain/BM55/2019(H1N2)) segment 4 hemagglutinin (HA) gene, partial cds
- KR700875.1 Influenza A virus (A/swine/Germany/Genthin-167D/2012(H1N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR699916.1 Influenza A virus (A/swine/Netherlands/Haaksbergen-136/2011(H1N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR699821.1 Influenza A virus (A/swine/Italy/41350/2011(H1N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR700177.1 Influenza A virus (A/swine/Spain/29257/2012(H1N2)) segment 4 hemagglutinin (HA) gene, complete cds
- MN932178.1 Influenza A virus (A/swine/Spain/055/2017(H1N2)) segment 4 hemagglutinin (HA) gene, partial cds
- MN932181.1 Influenza A virus (A/swine/Spain/099/2018(H1N1)) segment 4 hemagglutinin (HA) gene, partial cds
- CY009892.1 Influenza A virus (A/Swine/Spain/50047/2003(H1N1)) segment 4, complete sequence
- MN932213.1 Influenza A virus (A/swine/Spain/153/2018(H1N2)) segment 4 hemagglutinin (HA) gene, partial cds
- KR700485.1 Influenza A virus (A/swine/Poland/00109/2010(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds
- KR699876.1 Influenza A virus (A/swine/Netherlands/Groelnlo-186/2011(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds
- KR700153.1 Influenza A virus (A/swine/Spain/23998/2011(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds
- MN932208.1 Influenza A virus (A/swine/Spain/BM40/2019(H1N2)) segment 4 hemagglutinin (HA) gene, partial cds
- KR700891.1 Influenza A virus (A/swine/Belgium/Lichtervelde-57/2013(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds

Secuencias del subtipo H3 con el código de acceso del GenBank:

- MN929037.1 Influenza A virus (A/swine/Spain/018/2017(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- MN929030.1 Influenza A virus (A/swine/Spain/041/2017(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- MN929029.1 Influenza A virus (A/swine/Spain/001/2017(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR701471.1 Influenza A virus (A/swine/Netherlands/Ysselsteyn-CVI8864A/2012(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR701407.1 Influenza A virus (A/swine/Germany/Coesfeld-Harle-IDT14741-1/2012(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR701257.1 Influenza A virus (A/swine/Netherlands/Vragender-503/2011(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR701399.1 Influenza A virus (A/swine/Germany/Visbek-IDT13347/2011(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR701447.1 Influenza A virus (A/swine/Spain/33936/2012(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR701463.1 Influenza A virus (A/swine/Spain/31768/2012(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- KR701241.1 Influenza A virus (A/swine/Belgium/Moorslede-515/2011(H3N2)) segment 4 hemagglutinin (HA) gene, complete cds
- MN929035.1 Influenza A virus (A/swine/Spain/BM37/2019(H3N1)) segment 4 hemagglutinin (HA) gene, complete cds

Secuencias del subtipo N1 con el código de acceso del GenBank:

- MN932225.1 Influenza A virus (A/swine/Spain/BM16/2019(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- MN932228.1 Influenza A virus (A/swine/Spain/BM37/2019(H3N1)) segment 6 neuraminidase (NA) gene, complete cds
- MK303332.1 Influenza A virus (A/swine/France/53-180028/2018(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- KU323167.1 Influenza A virus (A/swine/Italy/23721/2015(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- JQ253795.1 Influenza A virus (A/swine/Norway/02_11342/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- MN932215.1 Influenza A virus (A/swine/Spain/104/2018(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- MN932219.1 Influenza A virus (A/swine/Spain/062/2017(H3N1)) segment 6 neuraminidase (NA) gene, complete cds
- MN932216.1 Influenza A virus (A/swine/Spain/103/2018(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- KR700833.1 Influenza A virus (A/swine/Netherlands/Vlierden-220/2012(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- KR699665.1 Influenza A virus (A/swine/Germany/Lohne-IDT12137/2010(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- KR699990.1 Influenza A virus (A/swine/Netherlands/Putten-60/2012(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- KR699846.1 Influenza A virus (A/swine/Italy/317775/2010(H1N1)) segment 6 neuraminidase (NA) gene, complete cds

- MN932220.1 Influenza A virus (A/swine/Spain/117/2018(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
- MN932229.1 Influenza A virus (A/swine/Spain/BM38/2019(H3N1)) segment 6 neuraminidase (NA) gene, complete cds

Secuencias del subtipo N2 con el código de acceso del GenBank:

- CY009382.1 Influenza A virus (A/swine/Spain/39139/2002(H3N2)) segment 6, complete sequence
- MN933435.1 Influenza A virus (A/swine/Spain/001/2017(H3N2)) segment 6 neuraminidase (NA) gene, complete cds
- KR700360.1 Influenza A virus (A/swine/Germany/Dulmen-IDT14394/2011(H1N2)) segment 6 neuraminidase (NA) gene, complete cds
- KR701211.1 Influenza A virus (A/swine/Netherlands/Roosendaal-571/2011(H3N2)) segment 6 neuraminidase (NA) gene, complete cds
- MN933424.1 Influenza A virus (A/swine/Spain/BM32/2019(H1N2)) segment 6 neuraminidase (NA) gene, complete cds
- KR700511.1 Influenza A virus (A/swine/Netherlands/Gladau-409/2012(H1N2)) segment 6 neuraminidase (NA) gene, complete cds
- MN933412.1 Influenza A virus (A/swine/Spain/090/2018(H3N2)) segment 6 neuraminidase (NA) gene, complete cds
- JX908107.1 Influenza A virus (A/swine/Spain/07/2011(H1N2)) segment 6 neuraminidase (NA) gene, partial cds
- KR700713.1 Influenza A virus (A/swine/England/112340/2011(H1N2)) segment 6 neuraminidase (NA) gene, complete cds
- KR699934.1 Influenza A virus (A/swine/Belgium/Oostkamp-26/2012(H1N2)) segment 6 neuraminidase (NA) gene, complete cds
- KR700594.1 Influenza A virus (A/swine/Spain/29262/2012(H1N2)) segment 6 neuraminidase (NA) gene, complete cds
- KU322572.1 Influenza A virus (A/swine/Italy/337085/2012(H1N2)) segment 6 neuraminidase (NA) gene, complete cds

10.2. Anexo 2: Pasos para realizar un análisis filogenético

1.º Elegir el marcador molecular

ADN: permite comparar genes altamente conservados entre especies

o

Proteínas: estudio de la evolución entre dos organismos muy diferentes.

2.º Realizar el alineamiento de secuencias múltiple

Paso más critico, solo un correcto alineamiento hará posible inferir un correcto árbol filogenético.
Se puede realizar con los softwares: Clustal Omega, MUSCLE, MAFFT, T-COFFEE

3.º Elección del modelo de evolución

Modelos de sustitución de nucleótidos

Modelo	Suposición	Parámetros del modelo
Jukes - Cantor 1969	• Todos los nucleótidos tienen la misma probabilidad de ser sustituidos.	μ
Kimura 1980	• Transiciones ($\beta = \varepsilon$) y transversiones ocurren con diferente probabilidad ($\alpha = \delta = \gamma = \eta$). • Todos los nucleótidos pueden ser sustituidos con la misma probabilidad ($\pi_A = \pi_C = \pi_G = \pi_T = 0{,}25$)	Modelo original α era la tasa de transiciones y β la única transversión, (κ y μ).
Felsestein 1981	Es una extension de Jukes-Cantor 1969: • todos los nucleótidos pueden ser susitutidos con diferentes probabilidades (π_A ! = π_C ! = π_G ! = π_T , con valores entre 0 y 1) • Transiciones y transversiones ocurren con la misma probabilidad ($\alpha = \beta = \delta = \gamma = \varepsilon = \eta$ =1).	μ, π_A, π_C, π_G, π_T
Tamura 1992	Es una extension de Kimura 1980: • Las transiciones ($\beta = \varepsilon$) y transversiones ocurren con diferente probabilidad ($\alpha = \delta = \gamma = \eta$). • Los nucleótidos son sustituidos con una probabilidad ajustada al contenido de GC en la secuencia de DNA ($\pi_C = \pi_G = \pi_{GC} / 2$; $\pi_A = \pi_T = (1 - \pi_{GC})/2$).	μ, π_{GC}
Hasegawa-Kishano-Yano 1985	Es una combinación de las extensions de Kimura 1980 y Felsestein 1981: • Transiciones ($\beta = \varepsilon$)y transversiones ocurren con diferente probabilidad ($\alpha = \delta = \gamma = \eta$). • Los nucleótidos pueden ser sustituidos con diferente probabilidad (π_A ! = π_C ! = π_G ! = π_T, con valores entre 0 y 1).	μ, κ, π_A, π_C, π_G, π_T
Tamura - Nei 1993	• La probabilidad de transición A<->G (β) es diferente de la probabilidad de transición C<->T (ε), y también la transversión ocurre con diferente probabilidad, pero todos los tipos de transversiones ocurren con la misma ($\alpha = \delta = \gamma = \eta$). • Los nucleótidos pueden ser sustituidos con diferente probabilidad (π_A ! = π_C ! = π_G ! = π_T, con valores entre 0 y 1).	μ, κ_1, κ_2, π_A, π_C, π_G, π_T
Generalised time-reversible 1986	• Es el modelo más complejo. • Los nucleótidos pueden ser sustituidos con diferente probabilidad (π_A ! = π_C ! = π_G ! = π_T, con valores entre 0 y 1). • La probabilidad de las transiciones y transversiones es distinta en cada combinación (α ! = β ! = δ ! = γ != ε != η).	α, β, δ, γ, ε, η, π_A, π_C, π_G, π_T

Modelos de sustitución de aminoácidos:
- Modelos empíricos: matrices de sustitución calculadas a partir de la comparación de secuencias. Modelos: Jones-Taylor-Thornton 1992 (JTT) y Wheland and Goldman 2001 (WAG), Le Gascuel 2008 (LG).

Modelos de sustitución de codones:
- Modelos mecanísticos: Nei y Gojobori 1986
  1. Cuenta sitios sinónimos y no sinónimos
  2. Cuenta sitios sinónimos y no sinónimos diferentes.
  3. Calcula la proporción de los diferentes y los correctos

4.º Determinar el método de construcción de árboles

Basados en la distancia: se mide cuánta diferencia hay entre pares de secuencias después de la alineación.
- Unweighted pair-group method using arithmetic averages (UPGMA)
- Neighbour joining (NJ)

Basados en los caracteres: cómo encajan los datos en las diferentes posibilidades de árboles:
- Maximum parsimony (MP)
- Maximum likelihood (ML)
- Bayesian method (BM)

5.º Evaluación de la fiabilidad del árbol

Bootstrapping: análisis estadístico computacional que consiste en volver a muestrear las muestras originales para crear nuevos subconjuntos. En estos nuevos subconjuntos se aplica la misma metodología de análisis. El proceso de análisis se repite cientos de veces.

IMAGEN 1. Pasos para realizar un análisis filogenético

10.3. Anexo 3: Resultados de la búsqueda del mejor modelo de evolución para cada variante H3, N1 y N2

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
HKY+G	48	11565.799	11167.232	-5535.537	n/a	0.37	4.30	0.347	0.237	0.190	0.226	0.023	0.018	0.183	0.033	0.154	0.021	0.033	0.192	0.021	0.281	0.023	0.018
TN93+G	49	11570.707	11163.840	-5532.838	n/a	0.38	4.21	0.347	0.237	0.190	0.226	0.023	0.018	0.160	0.033	0.179	0.021	0.033	0.223	0.021	0.247	0.023	0.018
HKY+G+I	49	11575.949	11169.082	-5535.459	0.29	0.77	4.30	0.347	0.237	0.190	0.226	0.023	0.018	0.183	0.033	0.154	0.021	0.033	0.192	0.021	0.281	0.023	0.018
TN93+G+I	50	11580.643	11165.477	-5532.653	0.38	1.15	4.21	0.347	0.237	0.190	0.226	0.023	0.018	0.160	0.033	0.180	0.021	0.033	0.224	0.021	0.246	0.023	0.018
T92+G	46	11584.237	11202.271	-5555.063	n/a	0.38	4.19	0.292	0.292	0.208	0.208	0.027	0.020	0.169	0.027	0.169	0.020	0.027	0.237	0.020	0.237	0.027	0.020
GTR+G	52	11585.564	11153.797	-5524.806	n/a	0.40	3.87	0.347	0.237	0.190	0.226	0.014	0.030	0.162	0.021	0.170	0.027	0.054	0.213	0.018	0.249	0.028	0.015
GTR+G+I	53	11589.638	11149.572	-5521.690	0.36	1.09	4.16	0.347	0.237	0.190	0.226	0.014	0.031	0.159	0.021	0.179	0.028	0.056	0.224	0.007	0.245	0.029	0.006
T92+G+I	47	11594.205	11203.938	-5554.894	0.35	0.98	4.19	0.292	0.292	0.208	0.208	0.027	0.020	0.169	0.027	0.169	0.020	0.027	0.237	0.020	0.237	0.027	0.020
HKY+I	48	11620.466	11221.900	-5562.871	0.36	n/a	3.78	0.347	0.237	0.190	0.226	0.025	0.020	0.178	0.037	0.150	0.024	0.037	0.187	0.024	0.274	0.025	0.020
TN93+I	49	11622.368	11215.501	-5558.669	0.36	n/a	3.74	0.347	0.237	0.190	0.226	0.025	0.020	0.155	0.036	0.177	0.024	0.036	0.221	0.024	0.238	0.025	0.020
GTR+I	52	11635.701	11203.934	-5549.875	0.36	n/a	3.43	0.347	0.237	0.190	0.226	0.016	0.032	0.156	0.023	0.168	0.029	0.059	0.209	0.021	0.240	0.030	0.017
T92+I	46	11636.478	11254.512	-5581.184	0.36	n/a	3.73	0.292	0.292	0.208	0.208	0.030	0.021	0.165	0.030	0.165	0.021	0.030	0.232	0.021	0.232	0.030	0.021
K2+G	45	11649.254	11275.589	-5592.725	n/a	0.36	4.23	0.250	0.250	0.250	0.250	0.024	0.024	0.202	0.024	0.202	0.024	0.024	0.202	0.024	0.202	0.024	0.024
K2+G+I	46	11659.392	11277.426	-5592.641	0.20	0.57	4.22	0.250	0.250	0.250	0.250	0.024	0.024	0.202	0.024	0.202	0.024	0.024	0.202	0.024	0.202	0.024	0.024
K2+I	45	11707.354	11333.689	-5621.775	0.36	n/a	3.72	0.250	0.250	0.250	0.250	0.027	0.027	0.197	0.027	0.197	0.027	0.027	0.197	0.027	0.197	0.027	0.027
HKY	47	11730.917	11340.650	-5623.250	n/a	n/a	3.49	0.347	0.237	0.190	0.226	0.027	0.021	0.175	0.039	0.147	0.025	0.039	0.184	0.025	0.269	0.027	0.021
TN93	48	11731.497	11332.930	-5618.387	n/a	n/a	3.48	0.347	0.237	0.190	0.226	0.026	0.021	0.153	0.038	0.173	0.025	0.038	0.216	0.025	0.236	0.026	0.021
GTR	51	11742.803	11319.336	-5608.579	n/a	n/a	3.20	0.347	0.237	0.190	0.226	0.017	0.034	0.154	0.025	0.165	0.030	0.062	0.206	0.021	0.236	0.031	0.018
T92	45	11745.256	11371.591	-5640.726	n/a	n/a	3.47	0.292	0.292	0.208	0.208	0.032	0.023	0.162	0.032	0.162	0.023	0.032	0.228	0.023	0.228	0.032	0.023
K2	44	11817.944	11452.579	-5682.223	n/a	n/a	3.46	0.250	0.250	0.250	0.250	0.028	0.028	0.194	0.028	0.194	0.028	0.028	0.194	0.028	0.194	0.028	0.028
JC+G	44	12171.058	11805.693	-5858.780	n/a	0.41	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G+I	45	12181.267	11807.601	-5858.731	0.19	0.62	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	44	12217.512	11852.147	-5882.007	0.36	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	43	12319.227	11962.163	-5938.018	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (*+G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (*+I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 23 nucleotide sequences. There were a total of 1719 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor./div>

1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

IMAGEN 1. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias H3 mediante Clustal Omega.

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
HKY+G	48	11565.799	11167.232	-5535.537	n/a	0.37	4.30	0.347	0.237	0.190	0.226	0.023	0.018	0.183	0.033	0.154	0.021	0.033	0.192	0.021	0.281	0.023	0.018
TN93+G	49	11570.707	11163.840	-5532.838	n/a	0.38	4.21	0.347	0.237	0.190	0.226	0.023	0.018	0.160	0.033	0.179	0.021	0.033	0.223	0.021	0.247	0.023	0.018
HKY+G+I	49	11575.949	11169.082	-5535.459	0.29	0.77	4.30	0.347	0.237	0.190	0.226	0.023	0.018	0.183	0.033	0.154	0.021	0.033	0.192	0.021	0.281	0.023	0.018
TN93+G+I	50	11580.643	11165.477	-5532.653	0.38	1.15	4.21	0.347	0.237	0.190	0.226	0.023	0.018	0.160	0.033	0.180	0.021	0.033	0.224	0.021	0.246	0.023	0.018
T92+G	46	11584.237	11202.271	-5555.063	n/a	0.38	4.19	0.292	0.292	0.208	0.208	0.027	0.020	0.169	0.027	0.169	0.020	0.027	0.237	0.020	0.237	0.027	0.020
GTR+G	52	11585.564	11153.797	-5524.806	n/a	0.40	3.87	0.347	0.237	0.190	0.226	0.014	0.030	0.162	0.021	0.170	0.027	0.054	0.213	0.018	0.249	0.028	0.015
GTR+G+I	53	11589.638	11149.572	-5521.690	0.36	1.09	4.16	0.347	0.237	0.190	0.226	0.014	0.031	0.159	0.021	0.179	0.028	0.056	0.224	0.007	0.245	0.029	0.006
T92+G+I	47	11594.205	11203.938	-5554.894	0.35	0.98	4.19	0.292	0.292	0.208	0.208	0.027	0.020	0.169	0.027	0.169	0.020	0.027	0.237	0.020	0.237	0.027	0.020
HKY+I	48	11620.466	11221.900	-5562.871	0.36	n/a	3.78	0.347	0.237	0.190	0.226	0.025	0.020	0.178	0.037	0.150	0.024	0.037	0.187	0.024	0.274	0.025	0.020
TN93+I	49	11622.368	11215.501	-5558.669	0.36	n/a	3.74	0.347	0.237	0.190	0.226	0.025	0.020	0.155	0.036	0.177	0.024	0.036	0.221	0.024	0.238	0.025	0.020
GTR+I	52	11635.701	11203.934	-5549.875	0.36	n/a	3.43	0.347	0.237	0.190	0.226	0.016	0.032	0.156	0.023	0.168	0.029	0.059	0.209	0.021	0.240	0.030	0.017
T92+I	46	11636.478	11254.512	-5581.184	0.36	n/a	3.73	0.292	0.292	0.208	0.208	0.030	0.021	0.165	0.030	0.165	0.021	0.030	0.232	0.021	0.232	0.030	0.021
K2+G	45	11649.254	11275.589	-5592.725	n/a	0.36	4.23	0.250	0.250	0.250	0.250	0.024	0.024	0.202	0.024	0.202	0.024	0.024	0.202	0.024	0.202	0.024	0.024
K2+G+I	46	11659.392	11277.426	-5592.641	0.20	0.57	4.22	0.250	0.250	0.250	0.250	0.024	0.024	0.202	0.024	0.202	0.024	0.024	0.202	0.024	0.202	0.024	0.024
K2+I	45	11707.354	11333.689	-5621.775	0.36	n/a	3.72	0.250	0.250	0.250	0.250	0.027	0.027	0.197	0.027	0.197	0.027	0.027	0.197	0.027	0.197	0.027	0.027
HKY	47	11730.917	11340.650	-5623.250	n/a	n/a	3.49	0.347	0.237	0.190	0.226	0.027	0.021	0.175	0.039	0.147	0.025	0.039	0.184	0.025	0.269	0.027	0.021
TN93	48	11731.497	11332.930	-5618.387	n/a	n/a	3.48	0.347	0.237	0.190	0.226	0.026	0.021	0.153	0.038	0.173	0.025	0.038	0.216	0.025	0.236	0.026	0.021
GTR	51	11742.803	11319.336	-5608.579	n/a	n/a	3.20	0.347	0.237	0.190	0.226	0.017	0.034	0.154	0.025	0.165	0.030	0.062	0.206	0.021	0.236	0.031	0.018
T92	45	11745.256	11371.591	-5640.726	n/a	n/a	3.47	0.292	0.292	0.208	0.208	0.032	0.023	0.162	0.032	0.162	0.023	0.032	0.228	0.023	0.228	0.032	0.023
K2	44	11817.944	11452.579	-5682.223	n/a	n/a	3.46	0.250	0.250	0.250	0.250	0.028	0.028	0.194	0.028	0.194	0.028	0.028	0.194	0.028	0.194	0.028	0.028
JC+G	44	12171.058	11805.693	-5858.780	n/a	0.41	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G+I	45	12181.267	11807.601	-5858.731	0.19	0.62	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	44	12217.512	11852.147	-5882.007	0.36	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	43	12319.227	11962.163	-5938.018	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximu parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (*+G*) with 5 rate categories and by assuming tha invariable (*+I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each mo frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. F automatically computed. This analysis involved 23 nucleotide sequences. There were a total of 1719 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor./div>

1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

IMAGEN 2. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias H3 mediante MUSCLE.

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G	60	15596.432	15089.432	-7484.610	n/a	0.39	5.28	0.318	0.263	0.181	0.238	0.019	0.029	0.178	0.023	0.174	0.015	0.050	0.253	0.002	0.239	0.017	0.002
GTR+G+I	61	15605.037	15089.591	-7483.686	0.26	0.73	5.29	0.318	0.263	0.181	0.238	0.019	0.029	0.178	0.023	0.174	0.015	0.050	0.253	0.002	0.239	0.017	0.001
T92+G	54	15617.841	15161.523	-7526.676	n/a	0.39	5.26	0.291	0.291	0.209	0.209	0.023	0.016	0.177	0.023	0.177	0.016	0.023	0.245	0.016	0.245	0.023	0.016
HKY+G	56	15625.064	15151.851	-7519.833	n/a	0.38	5.40	0.318	0.263	0.181	0.238	0.021	0.014	0.201	0.025	0.153	0.019	0.025	0.222	0.019	0.268	0.021	0.014
T92+G+I	55	15626.668	15161.902	-7525.862	0.30	0.83	5.27	0.291	0.291	0.209	0.209	0.023	0.016	0.177	0.023	0.177	0.016	0.023	0.245	0.016	0.245	0.023	0.016
TN93+G	57	15628.695	15147.036	-7516.422	n/a	0.39	5.29	0.318	0.263	0.181	0.238	0.021	0.014	0.179	0.025	0.173	0.019	0.025	0.252	0.019	0.239	0.021	0.014
HKY+G+I	57	15633.528	15151.868	-7518.839	0.32	0.89	5.41	0.318	0.263	0.181	0.238	0.020	0.014	0.201	0.025	0.153	0.019	0.025	0.222	0.019	0.269	0.020	0.014
TN93+G+I	58	15637.512	15147.405	-7515.604	0.28	0.78	5.30	0.318	0.263	0.181	0.238	0.021	0.014	0.179	0.025	0.173	0.019	0.025	0.252	0.019	0.239	0.021	0.014
GTR+I	60	15661.942	15154.942	-7517.365	0.54	n/a	5.21	0.318	0.263	0.181	0.238	0.019	0.028	0.187	0.023	0.165	0.017	0.048	0.239	0.002	0.251	0.019	0.002
T92+I	54	15677.175	15220.857	-7556.343	0.54	n/a	5.09	0.291	0.291	0.209	0.209	0.023	0.017	0.176	0.023	0.176	0.017	0.023	0.244	0.017	0.244	0.023	0.017
HKY+I	56	15682.523	15209.311	-7548.563	0.54	n/a	5.23	0.318	0.263	0.181	0.238	0.021	0.015	0.200	0.026	0.152	0.019	0.026	0.221	0.019	0.267	0.021	0.015
TN93+I	57	15689.594	15207.935	-7546.872	0.54	n/a	5.16	0.318	0.263	0.181	0.238	0.021	0.015	0.187	0.026	0.164	0.019	0.026	0.238	0.019	0.250	0.021	0.015
K2+G	53	15700.316	15252.445	-7573.140	n/a	0.38	5.25	0.250	0.250	0.250	0.250	0.020	0.020	0.210	0.020	0.210	0.020	0.020	0.210	0.020	0.210	0.020	0.020
K2+G+I	54	15709.611	15253.293	-7572.560	0.25	0.69	5.27	0.250	0.250	0.250	0.250	0.020	0.020	0.210	0.020	0.210	0.020	0.020	0.210	0.020	0.210	0.020	0.020
K2+I	53	15764.919	15317.048	-7605.441	0.54	n/a	5.08	0.250	0.250	0.250	0.250	0.021	0.021	0.209	0.021	0.209	0.021	0.021	0.209	0.021	0.209	0.021	0.021
GTR	59	16200.376	15701.824	-7791.810	n/a	n/a	4.74	0.318	0.263	0.181	0.238	0.021	0.030	0.165	0.025	0.181	0.017	0.053	0.263	0.004	0.221	0.018	0.003
T92	53	16230.179	15782.308	-7838.071	n/a	n/a	4.75	0.291	0.291	0.209	0.209	0.025	0.018	0.174	0.025	0.174	0.018	0.025	0.241	0.018	0.241	0.025	0.018
TN93	56	16233.637	15760.425	-7824.120	n/a	n/a	4.75	0.318	0.263	0.181	0.238	0.022	0.015	0.165	0.027	0.181	0.020	0.027	0.263	0.020	0.221	0.022	0.015
HKY	55	16253.181	15788.416	-7839.119	n/a	n/a	4.75	0.318	0.263	0.181	0.238	0.023	0.016	0.196	0.028	0.150	0.021	0.028	0.217	0.021	0.263	0.023	0.016
K2	52	16321.189	15881.765	-7888.803	n/a	n/a	4.74	0.250	0.250	0.250	0.250	0.022	0.022	0.206	0.022	0.206	0.022	0.022	0.206	0.022	0.206	0.022	0.022
JC+G	52	16906.156	16466.733	-8181.287	n/a	0.41	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G+I	53	16916.247	16468.376	-8181.105	0.29	0.89	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	52	16956.013	16516.590	-8206.215	0.54	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	51	17478.410	17047.434	-8472.640	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+*G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+*I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 27 nucleotide sequences. There were a total of 1411 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor.

1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

IMAGEN 3. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias N1 mediante Clustal Omega.

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G	60	15596.432	15089.432	-7484.610	n/a	0.39	5.28	0.318	0.263	0.181	0.238	0.019	0.029	0.178	0.023	0.174	0.015	0.050	0.253	0.002	0.239	0.017	0.002
GTR+G+I	61	15605.037	15089.591	-7483.686	0.26	0.73	5.29	0.318	0.263	0.181	0.238	0.019	0.029	0.178	0.023	0.174	0.015	0.050	0.253	0.002	0.239	0.017	0.001
T92+G	54	15617.841	15161.523	-7526.676	n/a	0.39	5.26	0.291	0.291	0.209	0.209	0.023	0.016	0.177	0.023	0.177	0.016	0.023	0.245	0.016	0.245	0.023	0.016
HKY+G	56	15625.064	15151.851	-7519.833	n/a	0.38	5.40	0.318	0.263	0.181	0.238	0.021	0.014	0.201	0.025	0.153	0.019	0.025	0.222	0.019	0.268	0.021	0.014
T92+G+I	55	15626.668	15161.902	-7525.862	0.30	0.83	5.27	0.291	0.291	0.209	0.209	0.023	0.016	0.177	0.023	0.177	0.016	0.023	0.245	0.016	0.245	0.023	0.016
TN93+G	57	15628.695	15147.036	-7516.422	n/a	0.39	5.29	0.318	0.263	0.181	0.238	0.021	0.014	0.179	0.025	0.173	0.019	0.025	0.252	0.019	0.239	0.021	0.014
HKY+G+I	57	15633.528	15151.868	-7518.839	0.32	0.89	5.41	0.318	0.263	0.181	0.238	0.020	0.014	0.201	0.025	0.153	0.019	0.025	0.222	0.019	0.269	0.020	0.014
TN93+G+I	58	15637.512	15147.405	-7515.604	0.28	0.78	5.30	0.318	0.263	0.181	0.238	0.021	0.014	0.179	0.025	0.173	0.019	0.025	0.252	0.019	0.239	0.021	0.014
GTR+I	60	15661.942	15154.942	-7517.365	0.54	n/a	5.21	0.318	0.263	0.181	0.238	0.019	0.028	0.187	0.023	0.165	0.017	0.048	0.239	0.002	0.251	0.019	0.002
T92+I	54	15677.175	15220.857	-7556.343	0.54	n/a	5.09	0.291	0.291	0.209	0.209	0.023	0.017	0.176	0.023	0.176	0.017	0.023	0.244	0.017	0.244	0.023	0.017
HKY+I	56	15682.523	15209.311	-7548.563	0.54	n/a	5.23	0.318	0.263	0.181	0.238	0.021	0.015	0.200	0.026	0.152	0.019	0.026	0.221	0.019	0.267	0.021	0.015
TN93+I	57	15689.594	15207.935	-7546.872	0.54	n/a	5.16	0.318	0.263	0.181	0.238	0.021	0.015	0.187	0.026	0.164	0.019	0.026	0.238	0.019	0.250	0.021	0.015
K2+G	53	15700.316	15252.445	-7573.140	n/a	0.38	5.25	0.250	0.250	0.250	0.250	0.020	0.020	0.210	0.020	0.210	0.020	0.020	0.210	0.020	0.210	0.020	0.020
K2+G+I	54	15709.611	15253.293	-7572.560	0.25	0.69	5.27	0.250	0.250	0.250	0.250	0.020	0.020	0.210	0.020	0.210	0.020	0.020	0.210	0.020	0.210	0.020	0.020
K2+I	53	15764.919	15317.048	-7605.441	0.54	n/a	5.08	0.250	0.250	0.250	0.250	0.021	0.021	0.209	0.021	0.209	0.021	0.021	0.209	0.021	0.209	0.021	0.021
GTR	59	16200.376	15701.824	-7791.810	n/a	n/a	4.74	0.318	0.263	0.181	0.238	0.021	0.030	0.165	0.025	0.181	0.017	0.053	0.263	0.004	0.221	0.018	0.003
T92	53	16230.179	15782.308	-7838.071	n/a	n/a	4.75	0.291	0.291	0.209	0.209	0.025	0.018	0.174	0.025	0.174	0.018	0.025	0.241	0.018	0.241	0.025	0.018
TN93	56	16233.637	15760.425	-7824.120	n/a	n/a	4.75	0.318	0.263	0.181	0.238	0.022	0.015	0.165	0.027	0.181	0.020	0.027	0.263	0.020	0.221	0.022	0.015
HKY	55	16253.181	15788.416	-7839.119	n/a	n/a	4.75	0.318	0.263	0.181	0.238	0.023	0.016	0.196	0.028	0.150	0.021	0.028	0.217	0.021	0.263	0.023	0.016
K2	52	16321.189	15881.765	-7888.803	n/a	n/a	4.74	0.250	0.250	0.250	0.250	0.022	0.022	0.206	0.022	0.206	0.022	0.022	0.206	0.022	0.206	0.022	0.022
JC+G	52	16906.156	16466.733	-8181.287	n/a	0.41	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G+I	53	16916.247	16468.376	-8181.105	0.29	0.89	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	52	16956.013	16516.590	-8206.215	0.54	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	51	17478.410	17047.434	-8472.640	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and
uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+*G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (+*I*). Whenever applicable, estima
shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous
equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 27 nucleotide sequences. There were a total of 1411 positions in the final dataset. Evolutionary analyses were conducted in ME

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor.

1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

Disclaimer: Although utmost care has been taken to ensure the correctness of the caption, the caption text is provided "as is" without any warranty of any kind. Authors advise the user to carefully check the caption prior to its use for any purpose and report any errors or problems to the authors immediately (www
but not limited to special, consequential, or other damages. Authors specifically disclaim all other warranties expressed or implied, including but not limited to the determination of suitability of this caption text for a specific purpose, use, or application.

IMAGEN 4. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias N1 mediante MUSCLE.

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G+I	79	19209.229	18525.486	-9183.594	0.44	2.47	4.91	0.317	0.264	0.181	0.238	0.024	0.027	0.182	0.029	0.168	0.015	0.047	0.244	0.004	0.242	0.017	0.003
T92+G+I	73	19219.901	18588.067	-9220.906	0.44	2.55	4.82	0.290	0.290	0.210	0.210	0.024	0.018	0.174	0.024	0.174	0.018	0.024	0.241	0.018	0.241	0.024	0.018
GTR+G	78	19222.828	18547.736	-9195.723	n/a	0.44	4.93	0.317	0.264	0.181	0.238	0.024	0.027	0.182	0.029	0.167	0.015	0.047	0.244	0.004	0.242	0.016	0.003
HKY+G+I	75	19232.246	18583.109	-9216.420	0.44	2.53	4.91	0.317	0.264	0.181	0.238	0.022	0.015	0.198	0.027	0.150	0.020	0.027	0.219	0.020	0.263	0.022	0.015
T92+G	72	19235.999	18612.817	-9234.285	n/a	0.43	4.86	0.290	0.290	0.210	0.210	0.024	0.018	0.175	0.024	0.175	0.018	0.024	0.242	0.018	0.242	0.024	0.018
TN93+G+I	76	19237.046	18579.258	-9213.491	0.44	2.57	4.84	0.317	0.264	0.181	0.238	0.022	0.015	0.181	0.027	0.166	0.020	0.027	0.242	0.020	0.241	0.022	0.015
GTR+I	78	19242.523	18567.431	-9205.570	0.50	n/a	4.79	0.317	0.264	0.181	0.238	0.024	0.027	0.182	0.029	0.166	0.016	0.048	0.242	0.004	0.242	0.017	0.003
HKY+G	74	19247.722	18607.236	-9229.488	n/a	0.43	4.96	0.317	0.264	0.181	0.238	0.022	0.015	0.198	0.027	0.151	0.020	0.027	0.220	0.020	0.264	0.022	0.015
T92+I	72	19251.866	18628.684	-9242.218	0.50	n/a	4.68	0.290	0.290	0.210	0.210	0.025	0.018	0.174	0.025	0.174	0.018	0.025	0.240	0.018	0.240	0.025	0.018
TN93+G	75	19253.236	18604.099	-9226.915	n/a	0.44	4.88	0.317	0.264	0.181	0.238	0.022	0.015	0.182	0.027	0.166	0.020	0.027	0.242	0.020	0.241	0.022	0.015
HKY+I	74	19265.588	18625.103	-9238.421	0.50	n/a	4.76	0.317	0.264	0.181	0.238	0.023	0.016	0.197	0.027	0.150	0.021	0.027	0.218	0.021	0.262	0.023	0.016
TN93+I	75	19268.698	18619.561	-9234.646	0.50	n/a	4.71	0.317	0.264	0.181	0.238	0.023	0.016	0.181	0.027	0.165	0.021	0.027	0.240	0.021	0.240	0.023	0.016
K2+G+I	72	19309.545	18686.363	-9271.058	0.43	2.21	4.76	0.250	0.250	0.250	0.250	0.022	0.022	0.207	0.022	0.207	0.022	0.022	0.207	0.022	0.207	0.022	0.022
K2+G	71	19319.854	18705.324	-9281.541	n/a	0.43	4.80	0.250	0.250	0.250	0.250	0.022	0.022	0.207	0.022	0.207	0.022	0.022	0.207	0.022	0.207	0.022	0.022
K2+I	71	19347.812	18733.282	-9295.521	0.51	n/a	4.61	0.250	0.250	0.250	0.250	0.022	0.022	0.205	0.022	0.205	0.022	0.022	0.205	0.022	0.205	0.022	0.022
GTR	77	20011.109	19344.669	-9595.193	n/a	n/a	4.28	0.317	0.264	0.181	0.238	0.027	0.030	0.167	0.032	0.174	0.015	0.053	0.254	0.006	0.221	0.017	0.004
T92	71	20041.205	19426.675	-9642.217	n/a	n/a	4.28	0.290	0.290	0.210	0.210	0.027	0.019	0.171	0.027	0.171	0.019	0.027	0.237	0.019	0.237	0.027	0.019
TN93	74	20049.586	19409.101	-9630.420	n/a	n/a	4.28	0.317	0.264	0.181	0.238	0.025	0.017	0.166	0.030	0.174	0.022	0.030	0.253	0.022	0.221	0.025	0.017
HKY	73	20068.417	19436.583	-9645.164	n/a	n/a	4.28	0.317	0.264	0.181	0.238	0.025	0.017	0.193	0.030	0.147	0.023	0.030	0.214	0.023	0.257	0.025	0.017
K2	70	20118.038	19512.160	-9685.963	n/a	n/a	4.27	0.250	0.250	0.250	0.250	0.024	0.024	0.203	0.024	0.203	0.024	0.024	0.203	0.024	0.203	0.024	0.024
JC+G+I	71	20782.779	20168.249	-10013.004	0.44	3.08	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G	70	20798.977	20193.099	-10026.432	n/a	0.48	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	70	20804.773	20198.895	-10029.331	0.50	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	69	21521.208	20923.982	-10392.877	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (*+G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (*+I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 36 nucleotide sequences. There were a total of 1437 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor./div>

1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

IMAGEN 5. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias N2 mediante Clustal Omega.

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G+I	79	19207.207	18523.464	-9182.583	0.44	2.42	4.91	0.317	0.264	0.181	0.238	0.024	0.027	0.182	0.029	0.167	0.015	0.047	0.244	0.004	0.242	0.017	0.003
T92+G+I	73	19218.030	18586.196	-9219.971	0.44	2.51	4.82	0.290	0.290	0.210	0.210	0.024	0.018	0.174	0.024	0.174	0.018	0.024	0.242	0.018	0.242	0.024	0.018
GTR+G	78	19221.026	18545.935	-9194.822	n/a	0.44	4.93	0.317	0.264	0.181	0.238	0.024	0.027	0.183	0.029	0.167	0.015	0.047	0.243	0.004	0.243	0.016	0.003
HKY+G+I	75	19229.987	18580.850	-9215.291	0.44	2.50	4.91	0.317	0.264	0.181	0.238	0.022	0.015	0.198	0.027	0.150	0.020	0.027	0.219	0.020	0.263	0.022	0.015
T92+G	72	19232.504	18609.323	-9232.538	n/a	0.43	4.86	0.290	0.290	0.210	0.210	0.024	0.018	0.175	0.024	0.175	0.018	0.024	0.242	0.018	0.242	0.024	0.018
TN93+G+I	76	19235.067	18577.278	-9212.501	0.44	2.52	4.84	0.317	0.264	0.181	0.238	0.022	0.015	0.181	0.027	0.166	0.020	0.027	0.242	0.020	0.241	0.022	0.015
GTR+I	78	19241.708	18566.616	-9205.163	0.50	n/a	4.79	0.317	0.264	0.181	0.238	0.024	0.027	0.182	0.029	0.165	0.016	0.048	0.241	0.004	0.242	0.017	0.003
HKY+G	74	19245.416	18604.930	-9228.335	n/a	0.43	4.96	0.317	0.264	0.181	0.238	0.022	0.015	0.198	0.027	0.151	0.020	0.027	0.220	0.020	0.264	0.022	0.015
TN93+G	75	19251.156	18602.019	-9225.875	n/a	0.44	4.88	0.317	0.264	0.181	0.238	0.022	0.015	0.182	0.027	0.166	0.020	0.027	0.242	0.020	0.242	0.022	0.015
T92+I	72	19251.167	18627.985	-9241.869	0.50	n/a	4.68	0.290	0.290	0.210	0.210	0.025	0.018	0.174	0.025	0.174	0.018	0.025	0.240	0.018	0.240	0.025	0.018
HKY+I	74	19264.424	18623.938	-9237.838	0.50	n/a	4.76	0.317	0.264	0.181	0.238	0.023	0.016	0.197	0.027	0.150	0.021	0.027	0.218	0.021	0.262	0.023	0.016
TN93+I	75	19267.888	18618.751	-9234.241	0.50	n/a	4.71	0.317	0.264	0.181	0.238	0.023	0.016	0.181	0.027	0.165	0.021	0.027	0.240	0.021	0.241	0.023	0.016
K2+G+I	72	19308.175	18684.994	-9270.373	0.43	2.17	4.76	0.250	0.250	0.250	0.250	0.022	0.022	0.207	0.022	0.207	0.022	0.022	0.207	0.022	0.207	0.022	0.022
K2+G	71	19316.934	18702.404	-9280.082	n/a	0.43	4.80	0.250	0.250	0.250	0.250	0.022	0.022	0.207	0.022	0.207	0.022	0.022	0.207	0.022	0.207	0.022	0.022
K2+I	71	19347.744	18733.214	-9295.487	0.51	n/a	4.61	0.250	0.250	0.250	0.250	0.022	0.022	0.205	0.022	0.205	0.022	0.022	0.205	0.022	0.205	0.022	0.022
GTR	77	20010.730	19344.290	-9595.004	n/a	n/a	4.28	0.317	0.264	0.181	0.238	0.027	0.030	0.167	0.032	0.174	0.015	0.053	0.254	0.006	0.222	0.017	0.004
T92	71	20042.096	19427.566	-9642.663	n/a	n/a	4.28	0.290	0.290	0.210	0.210	0.027	0.019	0.171	0.027	0.171	0.019	0.027	0.237	0.019	0.237	0.027	0.019
TN93	74	20049.206	19408.720	-9630.229	n/a	n/a	4.28	0.317	0.264	0.181	0.238	0.025	0.017	0.166	0.030	0.173	0.022	0.030	0.253	0.022	0.221	0.025	0.017
HKY	73	20067.400	19435.566	-9644.656	n/a	n/a	4.28	0.317	0.264	0.181	0.238	0.025	0.017	0.193	0.030	0.147	0.023	0.030	0.214	0.023	0.257	0.025	0.017
K2	70	20116.582	19510.705	-9685.235	n/a	n/a	4.27	0.250	0.250	0.250	0.250	0.024	0.024	0.203	0.024	0.203	0.024	0.024	0.203	0.024	0.203	0.024	0.024
JC+G+I	71	20781.603	20167.073	-10012.416	0.44	3.01	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G	70	20796.041	20190.163	-10024.965	n/a	0.48	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	70	20804.720	20198.842	-10029.304	0.50	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC	69	21522.777	20925.551	-10393.662	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

NOTE.-- Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (*lnL*), and the number of parameters (including branch lengths) are also presented [1]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (*+G*) with 5 rate categories and by assuming that a certain fraction of sites are evolutionarily invariable (*+I*). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown.Assumed or estimated values of transition/transversion bias (*R*) are shown for each model, as well. They are followed by nucleotide frequencies (*f*) and rates of base substitutions (*r*) for each nucleotide pair.Relative values of instantaneous *r* should be considered when evaluating them. For simplicity, sum of *r* values is made equal to 1 for each model. For estimating ML values, a tree topology was automatically computed. This analysis involved 36 nucleotide sequences. There were a total of 1437 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2]

Abbreviations: TR: General Time Reversible; HKY: Hasegawa-Kishino-Yano; TN93: Tamura-Nei; T92: Tamura 3-parameter; K2: Kimura 2-parameter; JC: Jukes-Cantor./div>

1. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* 35:1547-1549.

IMAGEN 6. Captura de pantalla de la tabla obtenida para la búsqueda del modelo de evolución tras realizar el alineamiento múltiple de las secuencias N2 mediante MUSCLE.

10.4. Anexo 4. Resultados de los alineamientos realizados con LALIGN

```
>>EMBOSS_001                                              (606 aa)
 Waterman-Eggert score: 4398;  134.5 bits; E(1) <  1.2e-35
94.2% identity (96.7% similar) in 606 aa overlap (1-606:1-606)

               10        20        30        40        50        60
EMBOSS ATGAGATGTTCTCACAAATTGGGGCGTTTCTTGACTCCGCACTCTTGCTTCTGGTGGCTT
       :::::::::::::::::::::::.:::::::::::::::: ::::::::::::::::::::
EMBOSS ATGAGATGTTCTCACAAATTGGAGCGTTTCTTGACTCCTCACTCTTGCTTCTGGTGGCTT
               10        20        30        40        50        60

               70        80        90       100       110       120
EMBOSS TTTTTGCTGTGTACCGGTTTGTCCTGGTCCTTTGCCGATGGCAACGGCAACAGCTCGACA
       :::::::: :::::::: ::::: :::::::::: :::::::::::::::::::::::::
EMBOSS TTTTTGCTTTGTACCGGCTTGTCTTGGTCCTTTGTCGATGGCAACGGCAACAGCTCGACA
               70        80        90       100       110       120

              130       140       150       160       170       180
EMBOSS TACCAATACATATATAACTTGACGATATGCGAGCTGAATGGGACCGACTGGTTGTCCAGC
       ::::::::::::::::::::::::::::::::::::::::::.: ::::::::::::
EMBOSS TACCAATACATATATAACTTGACGATATGCGAGCTGAATGGGACCAAATGGTTGTCCAGC
              130       140       150       160       170       180

              190       200       210       220       230       240
EMBOSS CATTTTGGTTGGGCAGTCGAGACCTTTGTGTTTTACCCGGTTGCCACTCATATCCTCTCA
       :::::::. :::::::::::::::::::::: :::::::::::::::::::::::::: :::
EMBOSS CATTTTGACTGGGCAGTCGAGACCTTTGTGCTTTACCCGGTTGCCACTCATATCCTTTCA
              190       200       210       220       230       240


              250       260       270       280       290       300
EMBOSS CTGGGTTTTCTCACAACAAGCCATTTTTTTGACGCGCTCGGTCTCGGCGCTGTATCCACT
       :::::::::::::::::::::::::::::::: :::::::::::::::::::::.::::::
EMBOSS CTGGGTTTTCTCACAACAAGCCATTTTTTTGATGCGCTCGGTCTCGGCGCTGTGTCCACT
              250       260       270       280       290       300

              310       320       330       340       350       360
EMBOSS GCAGGATTTGTTGGCGGGCGGTATGTACTCTGCAGCGTCTACGGCGCTTGTGCTTTCGCA
       .:::::::::::::::.:::::::::::::.:::::::: ::::::::::::::::::::
EMBOSS ACAGGATTTGTTGGCAGGCGGTATGTACTCAGCAGCGTGTACGGCGCTTGTGCTTTCGCA
              310       320       330       340       350       360

              370       380       390       400       410       420
EMBOSS GCGTTCGTATGTTTTGTCATCCGTGCTGCTAAAAATTGCATGGCCTGCCGCTATGCCCGT
       ::: ::::::::::::::::::: :::::::::::::: ::::: ::::: ::::::::::
EMBOSS GCGCTCGTATGTTTTGTCATCCGCGCTGCTAAAAATTGTATGGCTTGCCGTTATGCCCGT
              370       380       390       400       410       420

              430       440       450       460       470       480
EMBOSS ACCCGGTTTACCAACTTCATTGTAGACAACCGGGGGAGAGTTCATCGATGGAAGTCTCCA
       :::::::::::::::::::::::.:: .::::::::::..: ::::::::::::::::::
EMBOSS ACCCGGTTTACCAACTTCATTGTGGATGACCGGGGGAGGATCCATCGATGGAAGTCTCCA
              430       440       450       460       470       480

              490       500       510       520       530       540
EMBOSS ATAGTGGTAGAAAAATTGGGCAAAGCCGAAGTCGACGGCAACCTCGTCACCATCAAACAT
       ::::::::::.:::::::::::::: ::.::::: ::::.:::::::::::::::::::::
EMBOSS ATAGTGGTAGAGAAATTGGGCAAAGCTGAGGTCGATGGCGACCTCGTCACCATCAAACAT
              490       500       510       520       530       540

              550       560       570       580       590       600
EMBOSS GTCGTCCTCGAAGGGGTTAAAGCTCAACCCTTGACGAGGACTTCGGCTGAGCAATGGGAG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::.
EMBOSS GTCGTCCTCGAAGGGGTTAAAGCTCAACCCTTGACGAGGACTTCGGCTGAGCAATGGGAA
              550       560       570       580       590       600


EMBOSS GCCTAG
       ::::::
EMBOSS GCCTAG
```

IMAGEN 1. Captura de pantalla del resultado del alineamiento realizado con LALIGN de la secuencia de ADN problema y la secuencia n.º 12.

```
>>EMBOSS_001                                              (201 aa)
 Waterman-Eggert score: 1310;  160.0 bits; E(1) <  2.8e-44
94.0% identity (97.5% similar) in 201 aa overlap (1-201:1-201)

               10        20        30        40        50        60
EMBOSS MRCSHKLGRFLTPHSCFWWLFLLCTGLSWSFADGNGNSSTYQYIYNLTICELNGTDWLSS
       ::::::: :::::::::::::::::::::::::.:::::::::::::::::::::::: ::::
EMBOSS MRCSHKLERFLTPHSCFWWLFLLCTGLSWSFVDGNGNSSTYQYIYNLTICELNGTKWLSS
               10        20        30        40        50        60

               70        80        90       100       110       120
EMBOSS HFGWAVETFVFYPVATHILSLGFLTTSHFFDALGLGAVSTAGFVGGRYVLCSVYGACAFA
       :: :::::::.:::::::::::::::::::::::::::::.:::: :::: :::::::::
EMBOSS HFDWAVETFVLYPVATHILSLGFLTTSHFFDALGLGAVSTTGFVGRRYVLSSVYGACAFA
               70        80        90       100       110       120

              130       140       150       160       170       180
EMBOSS AFVCFVIRAAKNCMACRYARTRFTNFIVDNRGRVHRWKSPIVVEKLGKAEVDGNLVTIKH
       :.:::::::::::::::::::::::::.:::.::::::::::::::::::.::::::
EMBOSS ALVCFVIRAAKNCMACRYARTRFTNFIVDDRGRIHRWKSPIVVEKLGKAEVDGDLVTIKH
              130       140       150       160       170       180

              190       200
EMBOSS VVLEGVKAQPLTRTSAEQWEA
       ::::::::::::::::::::::
EMBOSS VVLEGVKAQPLTRTSAEQWEA
              190       200
```

IMAGEN 2. Captura de pantalla del resultado del alineamiento realizado con LALIGN de la secuencia de aminoácidos problema y la secuencia n.º 12.