ISSN 1870-4069

# Comparación entre métodos de alineamiento de múltiples secuencias para análisis filogenético de secuencias de ADN vaginales en R

Isaí Angulo-Jiménez, Juana Canul-Reich,
Betania Hernández-Ocaña

Universidad Juárez Autónoma de Tabasco,
División Académica de Ciencias y Tecnologías de la Información,
México

{juana.canul, betania.hernandez}@ujat.mx,
182H13001@egresados.ujat.mx

**Resumen.** En la presente investigación se documentan los resultados de aplicar métodos alineamiento de múltiples secuencias (MSA) a un conjunto de datos de secuencias con información perteneciente al microbioma de mujeres caucásicas. El objetivo fue realizar un flujo de trabajo intuitivo para análisis filogenético en R. Se definieron métodos MSA a través de una comparación entre los métodos: DECIPHER, ClustalW, ClustalOmega y MUSCLE, siendo ClustalW el más citado por la literatura. Estos resultados son importantes debido a que dentro de un mismo flujo de trabajo se hace uso de la implementación del algoritmo DADA2 para tareas de preprocesamiento y DECIPHER como herramienta MSA, logrando facilitar y simplificar las tareas para el usuario final, permitiendo realizar este tipo de tareas de manera práctica.

**Palabras clave:** Alineamiento de múltiples secuencias, microbioma, vaginosis bacteriana.

# Multiple Sequence Alignment Comparison in R for Phylogenetic Analysis from Vaginal Sequence Data

**Abstract.** This research documents the results of applying multiple sequence alignment (MSA) methods to a sequence data set with information pertaining to the microbiome of Caucasian women. The objective was to create an intuitive workflow for phylogenetic analysis in R. MSA methods were defined through a comparison between the methods: DECIPHER, ClustalW, ClustalOmega and MUSCLE, with ClustalW being the most cited in the literature. These results are important because the implementation of the DADA2 algorithm for pre-processing tasks and DECIPHER as an MSA tool are used within the same workflow, thus facilitating and simplifying the tasks for the final user, allowing this type of analysis to be carried out in a practical way.

**Keywords:** Multiple sequence alignment, microbiome, bacterial vaginosis.

## 1. Introducción

El análisis computacional de los datos de secuencias a menudo implica el uso de distintos programas, códigos o herramientas que no necesariamente se encuentren relacionados entre sí, o pueden estar escritos en diferentes lenguajes de programación, diseñados para plataformas específicas, o en el peor de lo casos poseer una implementación poco intuitiva.

Las tareas de preprocesamiento, clasificación taxonómica, alineamiento de múltiples secuencias (MSA) y obtención de árboles filogenéticos emplean métodos matemáticos que están implementados en diferentes herramientas, lo que ocasiona que existan diversos formatos para su almacenamiento, haciendo tediosa la tarea del análisis.

R [11], a pesar de ser un entorno para estadística, permite análisis filogenético a través de paquetes e implementación de funciones, lo cual ayuda a que el manejo de secuencias y sus respectivos análisis puedan realizarse dentro de una sola plataforma. Una tarea central para el análisis filogenético es la obtención de un árbol de distancias entre las secuencias, para lo cual se requiere que las secuencias sean alineadas previamente [3].

Los MSA logran su objetivo mediante la programación dinámica, método que requiere tiempo y espacio en memoria de orden $N * M$, en donde $N$ y $M$ son el ancho de las secuencias $a$ y $b$, respectivamente. Debido a la complejidad que implica el alineamiento de secuencias largas, heurísticas son utilizadas para acelerar el alineamiento sin impactar negativamente en la precisión.

Esta precisión varía en función del número de secuencias que son añadidas al alineamiento, pues puede ser que la identidad entre secuencias o similitud cada vez sea menor, lo cual implicaría la obtención de un alineamiento impreciso. Este problema se encuentra frecuentemente en muestras de secuencias donde hay gran diversidad bacteriana, como en el caso de la microbiota vaginal. De acuerdo con Ortiz-Rodriguez [10]:

*"la vaginosis bacteriana, de origen polimicrobiano, es una alteración de la ecología vaginal donde la flora normal se ve prácticamente sustituida por gérmenes anaerobios. Muchos microorganismos han sido propuestos como causa de esta enfermedad, como la Gardnerella, Atopobium, Leptotrichia, Sneathia spp"*.

Han surgido trabajos que ejemplifican el uso de una sola herramienta para lograr un análisis filogenético, como por ejemplo Dadasnake, de Weißbecker et al. [15], el cual es un script en Python que hace uso del Divisive Amplicon Denoising Algorithm (DADA2) [2] para el preprocesamiento de secuencias y ClustalOmega como método de alineamiento, aunque no es su fin realizar un análisis filogenético.

Su uso está orientado a la ejecución en infreaestructuras de cómputo de alto rendimiento, las cuales cuentan con abundantes recursos de hardware, por lo que acceder a ellas en un principio podría implicar un problema al hacer estudios preliminares de este tipo.

De igual manera, Toparslan et al. [14] realizan un flujo de trabajo para secuencias de ADN mitocondriales escrito en su totalidad en R, pero no hay tareas de preprocesamiento debido a la naturaleza de las secuencias. Se hace uso del método de alineamiento ClustalW.

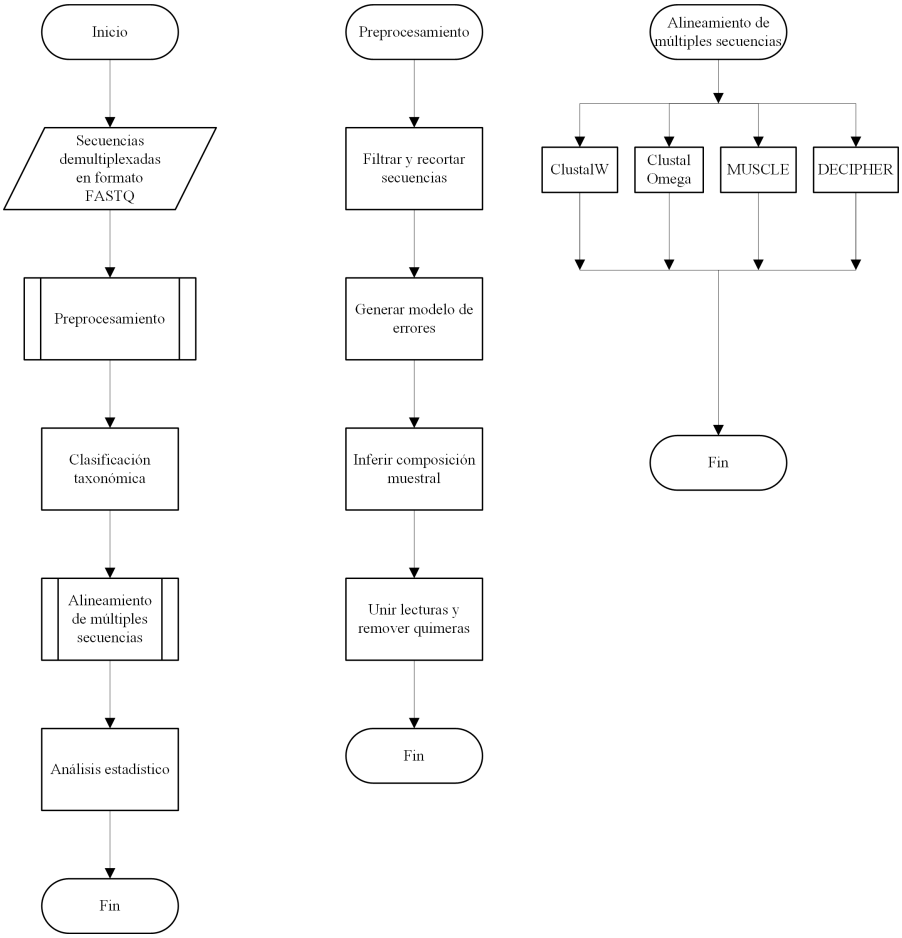


**Fig. 1.** Metodología.

El propósito de esta investigación es proveer un flujo de trabajo unificado en R que incluya implementaciones ya existentes para permitir tareas de análisis filogenético en una única plataforma, así como seleccionar métodos que puedan ser ejecutados en un equipo de cómputo personal. Para ello se hace una comparación entre implementaciones de los siguientes MSA: ClustalW, ClustalOmega, MUSCLE y DECIPHER.

## 2. Materiales y métodos

A continuación se detalla el conjunto de datos utilizado en el presente trabajo, así como conceptos relacionados con este artículo, como los distintos MSA utilizados. De igual manera se describen las medidas de rendimiento utilizadas para la evaluación estadística de los resultados obtenidos de cada alineamiento. Por último, se mencionan la paquetería necesaria para obtener los resultados de este artículo.

**Tabla 1.** Parámetros de entrada de los métodos MSA.

| **Clustal W** | | | |
|---|---|---|---|
| Apertura de Hueco | Extensión de Hueco | Máximo de Iteraciones | Orden |
| 15 | 6.66 | 3 | Entrada |
| **ClustalOmega** | | | |
| Apertura de Hueco | Extensión de Hueco | Máximo de Iteraciones | Orden |
| 6 | 1 | Sin límite | Entrada |
| **MUSCLE** | | | |
| Apertura de Hueco | Extensión de Hueco | Máximo de Iteraciones | Orden |
| 400 | 0 | 16 | Entrada |
| **DECIPHER** | | | |
| Apertura de Hueco | Extensión de Hueco | Máximo de Iteraciones | Orden |
| 16 | 1 | 2 | Entrada |

### 2.1. Conjunto de datos

Las secuencias pertenecen a un grupo de 155 mujeres caucásicas embarazadas. El microbioma de estas mujeres fue caracterizado mediante la secuenciación del gen 16S ARN ribosomal de las regiones V3-V4 en una plataforma MiSeq. La fuente de los datos es del European Nucleotide Archive (ENA), con el número de acceso PRJNA544732 y cargado por el Centro Médico Universitario de Liubliana, Eslovenia. Los datos obtenidos se encuentran comprimidos con la extensión FASTQ [5].

### 2.2. Métodos de alineamiento de múltiples secuencias

El problema de alinear secuencias está clasificado como un problema NP-completo [3]. Los métodos MSA se dividen en 4 tipos: exactos (programación dinámica), progresivos, basados en consistencia, e iterativos [6]. Los métodos ClustalW, ClustalOmega y MUSCLE se encuentran disponibles en R a través del paquete msa [1], mientras DECIPHER se encuentra en el paquete DECIPHER [17]. A continuación, cada método MSA utilizado:

**ClustalW:** Este método MSA basado en alineamiento progresivo inicia alineando pares de secuencias ya sea por el método k-tuple de Wilbur y Lipman o por el método de programación dinámica completo de Needleman-Wunsch, con los cuales obtienen medidas de distancia con las que se construye un árbol guía con el método Neighbour-Joining y por último alinea progesivamente las secuencias más cercanas acorde al árbol guía [13].

**ClustalOmega:** Este método MSA, también progresivo, se basa en un método de agrupamiento mBed para un alineamiento inicial, luego reagrupa por k-means. Utiliza UPGMA como método para construer el árbol guía y produce un alineamiento final alineando dos perfiles usando modelos ocultos de Markov (HMM) [12].

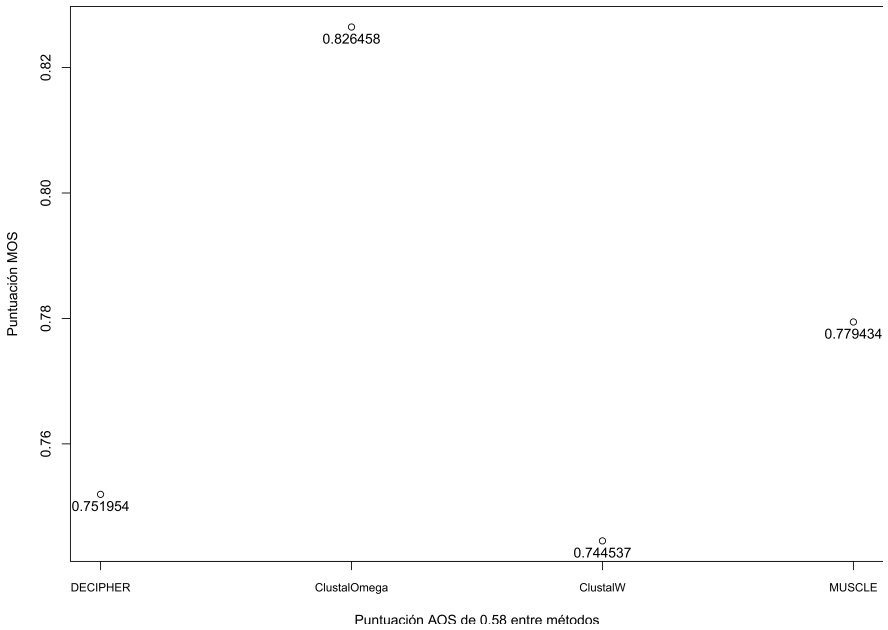


**Fig. 2.** Puntuaciones preliminares MOS y AOS para los cuatro métodos MSA.

**MUSCLE:** A diferencia de los otros tres métodos, este tiene una aproximación iterativa, en lugar de progresiva. Se reevalúan los alineamientos a través de dos distancias distintas: k-mer para pares de secuencias sin alinear y Kimura para las alineadas. Este procedimiento se repite siempre que se encuentre un alineamiento con una mejor puntuación que el anterior [4].

**DECIPHER:** Método MSA que toma en cuenta el contexto de las secuencias a través de la predicción de estructuras secundarias en el contexto de una secuencia local, incrementando la precisión del método. Esto permite la generación escalable de alineamientos de secuencias grandes manteniendo una precisión alta aún en conjuntos diversos de secuencias [16].

### 2.3. Medidas de puntuación

Las tareas tales como las búsquedas de homología entre secuencias, anotación genómica, predicción de la estructura de una proteína, así como áreas de biología evolutiva computacional, redes reguladoras de genes, y genómica funcional dependen del resultado de un método MSA.

El resultado obtenido de estas tareas bioinformáticas antes mencionadas tendrá una mayor significancia biológica a mayor precisión del resultado del MSA [7]. Sin embargo, debido a que no existe una función objetivo para medir verdaderamente la precisión o correctividad biológica de un alineamiento, existen métodos basados en distintas suposiciones.

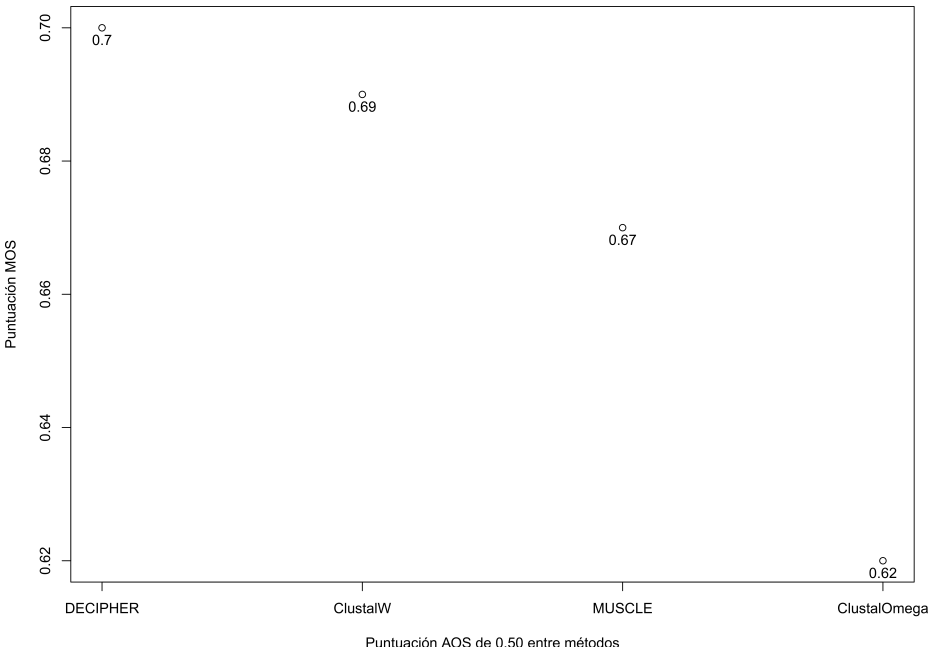


**Fig. 3.** Puntuaciones finales MOS y AOS para los cuatro métodos MSA.

La comparación cuantitativa de dos métodos MSA distintos ayuda a tomar decisiones sobre qué regiones están preservadas o cuáles deben ser removidas para tareas posteriores [8].

**Coincidencia (OS):** La función refleja la similitud entre dos alineamientos $Q_a$ y $Q_b$, y está definida como la relación entre la cardinalidad de la intersección de dos conjuntos de residuos alineados y la cardinalidad promedio de cada conjunto:

$$Q_{ab} = \frac{|Q_a \cap Q_b|}{(|Q_a| + |Q_b|)/2}. \tag{1}$$

**Coincidencia promedio (AOS):** Cada alineamiento se representa mediante el concepto de residuos de pares alineados. Cada uno de estos pares son extraídos de todos los alineamientos $m$ de entrada. La dificultad de un caso de alineamiento está definida por la puntuación de coincidencia promedio entre todos los alineamientos de entrada:

$$AOS = \frac{\sum_i^{m-1} \sum_{j=i-1}^{m} O_{ij}}{m(m-1)/2}. \tag{2}$$

Esta medida representa qué tan dispersos están los alineamientos en el espacio de todas las soluciones y se seleccionó como medida principal para decidir qué alineamiento utilizar.

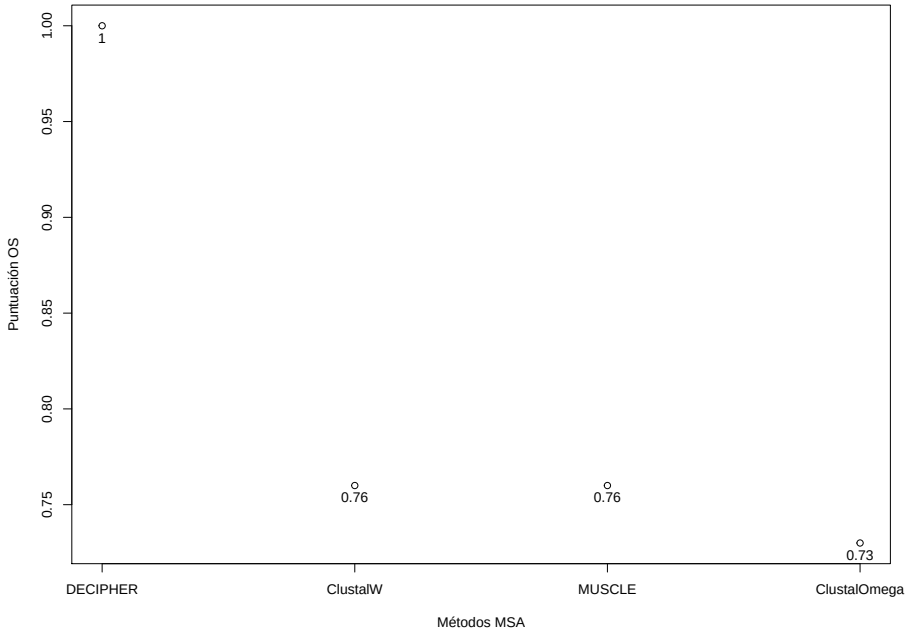


**Fig. 4.** Puntuación de coincidencia OS respecto al método DECIPHER.

Para casos simples, un método MSA dará como resultado alineamientos similares y el valor $AOS$ será muy cercano a $1$, mientras en casos difíciles su valor será cercano a $0$.

**Coincidencia múltiple (MOS):** Se asignan puntuaciones a cada par de residuos alineados reflejando su proliferación en todos los alineamientos. Sea $n(\sigma)$ el número de los $m-1$ alineamientos que contienen $\sigma$.

Un par que ocurra en todos los alineamientos es, en consecuencia, asignado con la puntuación mayor $(m-1)$ mientras que un par que ocurre en un solo alineamiento es asignado con la puntuación menor de cero. Estas puntuaciones son sumadas para el alineamiento $Q_a$ para determinar su puntuación de coincidencia múltiple:

$$MOS\left(Q_A\right) = \frac{\sum n(\sigma) : \sigma \in Q_a}{\left|Q_a\right|\left(m-1\right)}. \tag{3}$$

El numerador suma las puntuaciones de cada par de residuos alineados presentes en el alineamiento $Q_a$ . El denominador refleja la puntuación máxima posible. Los residuos alineados que son encontrados en varios alineamientos son más confiables, y el alineamiento con el mayor número de tales pares se asume como el más significativo biológicamente.

**Tabla 2.** Tiempo de ejecución de funciones principales.

| **Preprocesamiento** | | |
|---|---|---|
| **Proceso** | **Tiempo estimado por tictoc en segundos (s)** | **Multihilo (verdadero o falso)** |
| Filtro y recorte para Cutadapt | 504.795 s | Verdadero |
| Cutadapt | 1098.482 s | Verdadero |
| Filtro y recorte al resultado de Cutadapt | 446.305 s | Verdadero |
| Modelo de errores (forward) | 38.051 s | Verdadero |
| Verdadero | 195.136 s | Verdadero |
| Union de lecturas e inferencia muestral con DADA2 | 2637.742 s | Verdadero |
| **Clasificación taxonómica** | | |
| **Proceso** | **Tiempo estimado por tictoc en segundos (s)** | **Multihilo (verdadero o falso)** |
| Asignación taxonómica | 183.517 s | Verdadero |
| Asignación taxonómica | 651.717 s | Falso |
| **Metodos MSA** | | |
| **Proceso** | **Tiempo estimado por tictoc en segundos (s)** | **Multihilo (verdadero o falso)** |
| ClustalW | 2908.898 s | Falso |
| ClustalOmega | 100.779 s | Falso |
| MUSCLE | 2558.235 s | Falso |
| DECIPHER | 61.368 s | Verdadero |
| **Validacion estadística** | | |
| **Proceso** | **Tiempo estimado por tictoc en segundos (s)** | **Multihilo (verdadero o falso)** |
| MUMSA AOS = 0.58 | 43.188 s | Falso |
| MUMSA AOS = 0.50 | 115.923 s | Falso |
| MUMSA coincidencia OS | 44.139 s | Falso |

### 2.4. Paquetería utilizada

Se usaron los programas externos a R Cutadapt en su versión 2.1 con Python 3.8.5 y MUMSA en su versión 1.0, pero se incluyeron dentro del flujo propuesto con llamadas a los programas por medio de R.

Además, los paquetes en R versión 4.0.5 con las respectivas versiones de los mismas fueron: msa 1.22, DECIPHER 2.18.1, dada2 1.18.0, gridExtra 2.3, phangorn 2.5.5, ShortRead 1.48.0, Biostrings 2.58.0, ggplot2 3.3.3, phyloseq 1.34.0, cluster 2.1.1, dendextend 1.14.0, así como tictoc 1.0, para la medición del tiempo de ejecución de tareas centrales en el flujo de trabajo. Los paquetes fueron instalados junto con las dependencias de cada uno.

## 3. Diseño experimental

En la Figura 1 se describe la metodología, la cual inicia con la obtención de las secuencias a utilizar, en este caso pertenecientes a muestras vaginales. Como parte inicial del preprocesamiento, con Cutadapt se removieron los primers o iniciadores con secuencias ambiguas en los primeros 17 nucleótidos de las lecturas forward y primeros 21 para las reverse. DADA2 fue aplicado a lecturas forward y reverse con una calidad mínima de 20 basada en una puntuación Phred.

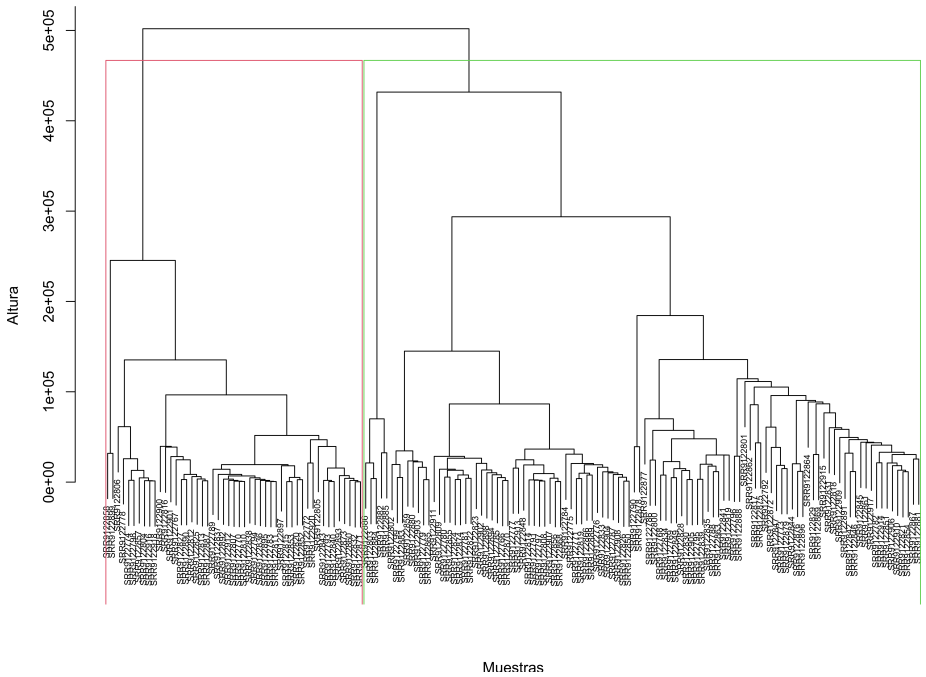


**Fig. 5.** Árbol jerárquico por el método Ward2 con $k = 2$.

La longitud mínima para las lecturas forward fue de 260 bases pareadas (bp) y para las lecturas reverse, 240 bp. Una remoción de secuencias quimeras fue necesaria posterior a la inferencia de variantes de secuencias (ASVs) a partir de la composición muestral, entendiendo por secuencias quimeras aquellas que se pueden construir exactamente mediante la combinación de segmentos izquierdo y derecho de dos secuencias "padre" más abundantes.

Este proceso dio como resultado 2,297 secuencias contenidas en las 155 muestras. Para la asignación taxonómica de las 2,297 secuencias, 1,974 fueron anotadas dentro del reino "Bacteria", 286 dentro de "Eukaryota", 1 dentro de "Archaea" y 36 sin anotar debido a que pertenecen al reino "Fungi", el cual no está presente en la base taxonómica utilizada. Debido al enfoque de la investigación referente a vaginosis bacteriana (VB), se trabajó solo con las 1,974 secuencias del reino "Bacteria".

Como principal parámetro de entrada se utilizaron las 1,974 secuencias obtenidas del preprocesamiento para los 4 MSA, cada método ejecutándose con sus parámetros por defecto, los cuáles se exponen en la Tabla 1.

Del resultado de cada método de alineamiento se obtuvo su respectivo árbol filogenético, el cual no requiere un análisis en esta etapa debido a que en etapas posteriores puede llegar a ser filtrado, aunque sí es requisito para la continuidad de este flujo de trabajo.

Con toda esta información obtenida, se procedió al uso de MUMSA para la validación estadística de los resultados obtenidos por los métodos MSA, con el fin de definir el método MSA a incluir en el flujo de trabajo.

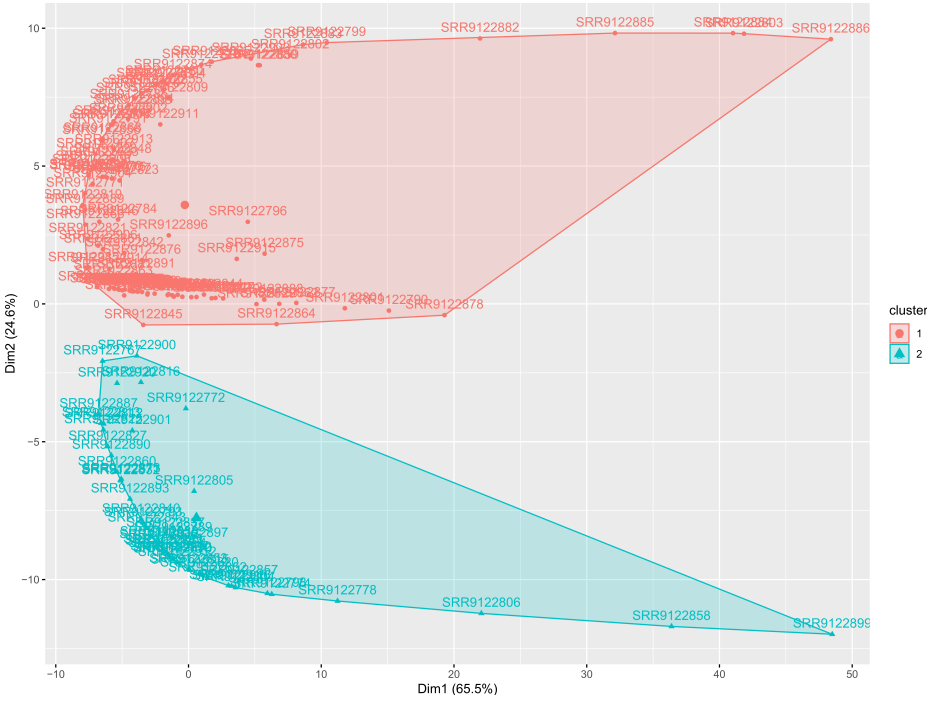


**Fig. 6.** Gráfico de dispersión de las 155 muestras analizadas con $k = 2$.

## 4. Resultados

R fue utilizado en su totalidad para la ejecución de las pruebas, aun sirviendo como interfaz para Cutadapt y MUMSA. Las características generales del equipo portátil tienen como SO Windows 10, procesador Intel Core i7-8750H de 12 núcleos a 2.208 GHz.

Para utilizar el total de núcleos en las tareas que así lo permitían se usó el Subsistema de Linux para Windows en su versión 2 con el SO Ubuntu 20.04 focal con kernel x86_64 Linux 5.4.72-microsoft-standard-WSL2 y un total de memoria RAM disponible de 25,562 GB.

Para todas las tareas realizadas se utilizó una sola semilla de valor 100, incluso para los métodos MSA. Una vez obtenida la salida de los cuatro métodos el primer uso de MUMSA fue predecir la dificultad del alineamiento partiendo de la coincidencia múltiple entre alineamientos y el resultado se observa en la Figura 2.

Esta primer medida no considera los residuos alineados a los huecos ni los pares de residuos alineados. Al tener como resultado un $AOS = 0.580479$, se está frente a un caso de alineamiento de dificultad cercana a la media, recordando el rango de dificultad $[0, 1]$, lo que supondría que el método ClustalOmega pareciese ser el indicado para cuestiones de este flujo de trabajo con un $MOS = 0.826458$, ya que supera al resto de los métodos.

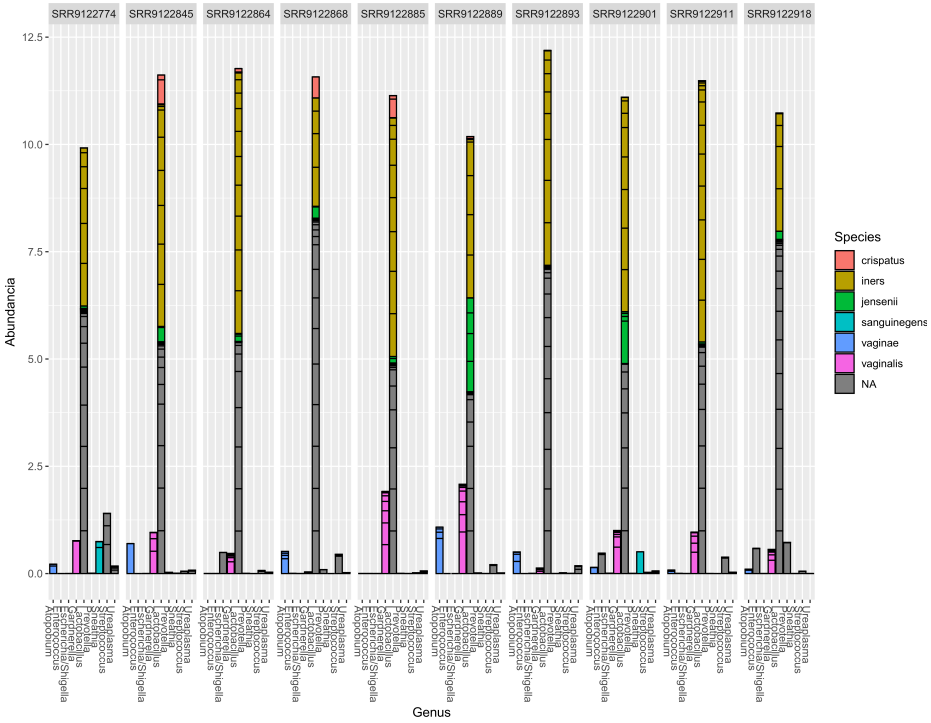


**Fig. 7.** Abundancia bacteriana de 10 muestras en rangos Genus y Species.

Sin embargo, al realizar la evaluación que sí toma las consideraciones de los residuos se obtiene como resultado un $AOS = 0.5$, indicando una dificultad mayor a la anteriormente obtenida, pero es el método DECIPHER con un valor $MOS = 0.7$ el que refleja una mayor confiabilidad estadística en los pares de residuos alineados a diferencia de los otros tres métodos (ver Figura 3).

Partiendo de la premisa de que el alineamiento con el mayor número de estos pares supone el biológicamente más significativo, se calculó la coincidencia OS entre métodos métodos respecto a DECIPHER, como se muestra en la Figura 4, en donde se aprecia que hay una diferencia importante en la intersección entre los demás métodos MSA con DECIPHER, que podría estudiarse a partir del número de columnas: ClustalW = 714, ClustalOmega = 908, MUSCLE = 2379, DECIPHER = 796, pero eso está fuera del enfoque de la presente investigación.

En la Tabla 2 se detalla el tiempo de ejecución de los principales procesos en el flujo de trabajo, así como la disponibilidad de multiprocesamiento de dicho proceso. Si solo se considera el tiempo aquí expuesto, el total sería de aproximadamente 11,588 s, poco más de 3 horas.

Habría que considerar que si se utiliza solo DECIPHER, método MSA más rápido y biológicamente más significativo para este análisis, el flujo de trabajo tomaría un tiempo de ejecución aproximado de 5,817 s, poco más de una hora y media.

Hasta esta etapa no se habían realizado exploraciones de la composición bacteriana a través de las muestras, ya que para esas tareas lo más conveniente es crear un objeto que pueda ser manipulable en R. El estudio de Hočevar et al. [5] posee información clínica de las muestras que son utilizadas para su investigación.

La aportación más grande del presente flujo de trabajo se centra en la posibilidad de llegar a resultados comparables en cuanto a inspección de la diversidad bacteriana de las muestras sin depender de datos clínicos, además de la posibilidad de aplicar un agrupamiento jerárquico a las 155 muestras. Todo esto con el propósito de verificar el alineamiento obtenido por el método DECIPHER.

En [5] se agrupan las 155 muestras en dos grupos: parto temprano con $n = 48$ y parto a término con $n = 107$. Debido a la ausencia de los datos clínicos propios de dicho estudio, se procedió a calcular un agrupamiento jerárquico aglomerativo (HC) mediante la función hclust usando Ward2 como método de agrupamiento, y el cálculo de valores disimilares mediante la función dist usando distancia euclideana.

Esto con el fin de saber si es posible aplicar técnicas de aprendizaje no supervisado a este tipo de secuencias en particular y que los resultados puedan ser analizados por los expertos. La semilla de valor 100 continuó siendo utilizada y al árbol jerárquico resultante se le hizo un recorte $k = 2$ para obtener dos grupos como se muestra en la Figura 5. También es posible su representación de dispersión de las muestras como en la Figura 6.

Los grupos se componen de 106 miembros el primero y 49 el segundo, lo que representa una aproximación muy cercana a lo obtenido por los datos clínicos. También se identificaron las bacterias a través de sus niveles taxonómicos. Como ejemplo, en la Figura 7 se muestran las bacterias más abundantes para 10 muestras obtenidas de un submuestreo del total de 155 con semilla de valor 100.

Los rangos comprendidos son Genus y Species en donde la mayoria de estas muestras están conformadas en gran parte por Lactobacillus, aunque también se aprecian muestras que podrían indicar VB debido a la alta concentración de Gardnerella, Atopobium o Sneathia además de una disminución de Lactobacillus.

## 5. Conclusiones y trabajos futuros

El fin del presente artículo fue comparar métodos MSA para así crear un flujo de trabajo intuitivo que permita un análisis filogenético en R e incluir validaciones estadísticas para tareas del análisis bioinformático.

Para ello se utilizaron herramientas ya desarrolladas para el SO Ubuntu, Cutadapt y MUMSA, y se llamaron mediante R para así complementar las tareas de preprocesamiento y validar estadísticamente los resultados de los alineamientos. El total de lecturas pareadas contenidas en las 155 muestras analizadas fue de 22,154,990.

Después de un primer filtrado con Cutadapt se obtuvieron 21,326,390, representando un 96.25 % del total. Estas fueron preprocesadas hasta llegar a 9,767,545, representando un 44.08 % del total. El número total de ASVs resultantes del preprocesamiento fue de 1,974 bacterias. Aun cuando los métodos MSA utilizados se basan en la programación dinámica, ClustalW, ClustalOmega y DECIPHER además hacen uso de alineamientos progresivos.

MUSCLE se basa en una aproximación iterativa. DECIPHER destacó del resto por la velocidad en la obtención del alineamiento, debido al uso de heurísticas al buscar k-mers de manera ordenada en la creación del árbol guía. Para este estudio se compararon los alineamientos obtenidos entre sí, una primera vez sin considerar residuos y una segunda vez tomándolos en cuenta.

ClustalOmega resultó el de mejor puntuación $MOS$ para la primera, pero en la segunda, DECIPHER resultó el de mayor valor $MOS$. Debido a que se definió utilizar la puntuación $AOS$ como parámetro, la segunda prueba tiene mayor significancia estadística, obteniendo un $AOS = 0.5$ contra el $AOS = 0.580479$ de la primera, implicando que existe una mayor dificultad en el alineamiento.

La segunda medida considerada fue la puntuación $MOS$, siendo DECIPHER el de mayor puntuación con un $MOS = 0.7$, lo cual indica una mayor fiabilidad biológica en el resultado del alineamiento. Así se tomó como referencia el resultado de DECIPHER y comparó contra el resto para obtener la tercer y última medida, $OS$, cuyos valores representan la coincidencia entre métodos respecto a DECIPHER y se visualiza en la Figura 4.

A manera de trabajo a futuro está la realización de una interfaz gráfica en R que permita el uso de este flujo de trabajo y sirva como herramienta práctica para el análisis filogenético de secuencias del gen 16S ARN ribosomal.

## Referencias

1. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., Hochreiter, S.: MSA: An R package for multiple sequence alignment. Bioinformatics, pp. btv494 (2015) doi: 10.1093/bioinformatics/btv494
2. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., Holmes, S. P.: DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, vol. 13, no. 7, pp. 581–583 (2016) doi: 10.1038/nmeth.3869
3. Daugelaite, J., O' Driscoll, A., Sleator, R. D.: An overview of multiple sequence alignments and cloud computing in bioinformatics. ISRN Biomathematics, vol. 2013, pp. 1–14 (2013) doi: 10.1155/2013/615630
4. Edgar, R. C.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, vol. 32, no. 5, pp. 1792–1797 (2004) doi: 10.1093/nar/gkh340
5. Hočevar, K., Maver, A., Vidmar Šimic, M., Hodžić, A., Haslberger, A., Premru Seršen, T., Peterlin, B.: Vaginal microbiome signature is associated with spontaneous preterm delivery. Frontiers in Medicine, vol. 6 (2019) doi: 10.3389/fmed.2019.00201
6. Issa, M., Hassanien, A. E.: Multiple sequence alignment optimization using meta-heuristic techniques. IHandbook of Research on Machine Learning Innovations and Trends, IGI Global, pp. 409–423 (2017) doi: 10.4018/978-1-5225-2229-4.ch018
7. Lecompte, O., Thompson, J. D., Plewniak, F., Thierry, J. C., Poch, O.: Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene, vol. 270, no. 1–2, pp. 17–30 (2001) doi: 10.1016/s0378-1119(01)00461-9
8. Lassmann, T.: Automatic assessment of alignment quality. Nucleic Acids Research, vol. 33, no. 22, pp. 7120–7128 (2005) doi: 10.1093/nar/gki1020
9. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, vol. 17, no. 1, pp. 10 (2011) doi: 10.14806/ej.17.1.200

10. Ortiz-Rodríguez, C., Ley-Ng, M., Llorente-Acebo, C., Almanza-Martínez, C.: Vaginosis bacteriana en mujeres con leucorrea. Revista Cubana de Obstetricia y Ginecología, vol. 26, no. 2, pp. 74–81 (2000)
11. R Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing www.R-project.org/
12. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., Higgins, D. G.: Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Molecular Systems Biology, vol. 7, no. 1, pp. 539 (2011) doi: 10.1038/msb.2011.75
13. Thompson, J. D., Higgins, D. G., Gibson, T. J.: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, vol. 22, no. 22, pp. 4673–4680 (1994) doi: 10.1093/nar/22.22.4673
14. Toparslan, E., Karabag, K., Bilge, U.: A workflow with R: Phylogenetic analyses and visualizations using mitochondrial cytochrome b gene sequences. PLOS ONE, vol. 15, no. 12, pp. e0243927 (2020) doi: 10.1371/journal.pone.0243927
15. Weißbecker, C., Schnabel, B., Heintz-Buschart, A.: Dadasnake, a Snakemake implementation of DADA2 to process amplicon sequencing data for microbial ecology. GigaScience, vol. 9, no. 12 (2020) doi: 10.1093/gigascience/giaa135
16. Wright, E. S.: DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics, vol. 16, no. 1 (2015) doi: 10.1186/s12859-015-0749-z
17. Wright, E. S.: Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. The R Journal, vol. 8, no. 1, pp. 352 (2016) doi: 10.32614/rj-2016-025