ISSN 1870-4069

Minería de secuencias de ADN para identificación de bacterias asociadas con Vaginosis Bacteriana

Freddy Garcia-Fuentes, Juana Canul-Reich,
Erick De-la-Cruz-Hernández, Betania Hernández-Ocaña,
Oscar Chávez-Bosquez

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

ffreddy.mx@gmail.com, {juana.canul, erick.delacruz,
betania.hernandez, oscar.chavez}@ujat.mx

Resumen. El presente trabajo clasifica documentos con secuencias de ADN para identificar microorganismos presentes en la vaginosis bacteriana. Se aplica el método probabilístico de Latent Dirichlet Allocation (LDA) para llevar a cabo un análisis de secuencia sin realizar ninguna técnica de alineación de secuencias. El proceso consiste en fragmentar secuencias de ADN en subsecuencias cortas llamadas k-mer. Con la colección de k-mers contenidas en los documentos se crea el corpus de documentos y se importa en LDA para generar la matriz de términos y tópicos. De la matriz de términos se mide la similitud coseno de los temas resultantes con el gen codificante ARNr 16S de los microorganismos: *(Gardnerella vaginalis, Atopobium vaginae, Prevotella sp, Lactobacillus crispatus y Lactobacillus inners).* Si el resultado de similitud coseno es mayor al 40 % se etiquetan los temas desconocidos con el nombre de la bacteria de mayor puntuación. De la matriz de tópicos, se observan las probabilidades con una puntuación de 99 % y se toman los documentos con las secuencias completas de ADN para validar manualmente en el banco de secuencias RDP disponible en la web. Los resultados hallaron los microorganismos presentes en la vaginosis bacteriana del conjunto de secuencias de ADN mediante las técnicas de minería de texto.

Palabras clave: ADN, ARNr 16S, LDA, k-mer, vaginosis bacteriana.

DNA Sequence Mining for Identification of Bacteria Associated with Bacterial Vaginosis

Abstract. The present work classifies documents with DNA sequences to identify microorganisms present in bacterial vaginosis. The probabilistic method of Latent Dirichlet Allocation (LDA) is applied to conduct sequence analysis without performing any sequence alignment technique. The process consists of

fragmenting DNA sequences into short subsequences called *(k-mer)*. With the collection of k-mers contained in the documents, the document corpus is created and imported into LDA to generate the matrix of terms and topics. From the matrix of terms, the cosine similarity of the resulting subjects with the 16S rRNA coding gene of the microorganisms is measured: *(Gardnerella vaginalis, Atopobium vaginae, Prevotella sp, Lactobacillus crispatus and Lactobacillus inners)*. If the cosine similarity result is greater than 40%, the unknown subjects are labeled with the name of the bacteria with the highest score. From the matrix of topics, probabilities with a score of 99% are identified and the documents with the entire DNA sequences are taken for manual validation in the RDP sequence bank available on the web. Results show that microorganisms present in bacterial vaginosis were found in the set of DNA sequences using text mining techniques.

Keywords: DNA, 16S rRNA, LDA, k-mer, bacterial vaginosis.

1. Introducción

La capacidad computacional en los últimos años ha incrementado exponencialmente, lo que permite procesar enormes cantidades de información en poco tiempo. Áreas como la biología se han beneficiado de estos avances, lo que ha permitido generar una gran cantidad de datos de secuencias de ADN.

Estas secuencias son piezas fundamentales en el estudio de la filogenética, que es una de las principales áreas de investigación de la ciencia biológica. Hay muchas investigaciones en el área de las ciencias de la vida para las que la Informática propone emplear técnicas de inteligencia artificial, específicamente del dominio de la minería de datos y comparar datos de secuencias de ADN sin necesidad de emplear el alineamiento.

Otras bacterias se alojan en la vagina humana y actuan como la primera línea de defensa contra las infecciones vaginales [10] y es importante en la salud integral de la mujer [5].

La vaginosis bacteriana (VB) se diagnostica mayoritariamente en mujeres en edad fértil, hasta ahora no se conoce con certeza la causa de su aparición, aunque se cree que implica la pérdida de lactobacilos vaginales y la adquisición de comunidades bacterianas que incluyen muchas bacterias asociadas a la VB [16].

La VB es una causa común de vaginitis y aumenta el riesgo de enfermedades inflamatorias pélvicas, resultados adversos del embarazo, riesgo de infección por inmunodeficiencia humana y problemas de infertilidad [16].

La clasificación de las secuencias genómicas para su identificación ha desempeñado un papel muy importante en el campo de la medicina y en el análisis de la biodiversidad [24]. La de clasificación de las secuencias presentes en un meta-genoma se lleva a cabo de la siguiente manera:

1. Comparando las secuencias contra genomas conocidos,
2. Extrayendo las secuencias de genes ribosomales (16S y 18S) y comparándo contra una base de datos [2].

La bioinformática emplea técnicas de análisis basadas en el alineamiento de secuencias, que buscan la similitud comparando una secuencia desconocida con otra conocida para identificar si existe una relación entre ambas secuencias. Estas técnicas de alineamiento son muy fiables. Sin embargo involucran operaciones con matrices y éstas tienden a ser demasiado grandes, lo que ocasiona que los algoritmos de alineamiento tomen mucho tiempo para presentar resultados.

Otro descubrimiento importante es la descomposición de cadenas de ADN en subcadenas cortas llamadas k-mer. Estos avances han permitido explotar las técnicas de minería de datos reduciendo la complejidad computacional de los problemas del tipo NP *(tiempo polinomial no determinista)* [1] para el tratamiento de secuencias genómicas con métodos de minería de textos [13].

2. Trabajos relacionados

En el campo de la bioinformática, el alineamiento de las secuencias genómicas [34] es el método utilizado para identificar regiones similares en las secuencias [19]. Actualmente, existen aproximaciones en el alineamiento empleando algoritmos de programación dinámica que garantizan un alineamiento óptimo pero requiere mucho tiempo computacional [33, 1].

Los algoritmos para el alineamiento de múltiples secuencias se consideran problemas NP-duros [12]. Por otro lado, se han propuesto algoritmos que se basan en la teoría de que no es necesario el alineamiento [34].

Los algoritmos sin alineamiento son conocidos por el uso de métodos de comparación para cuantificar la similitud o disimilitud entre una o más secuencias biológicas en lugar del proceso de alineamiento [1].

La distancia genética puede considerarse como uno de los mejores criterios para comparar diferentes especies, teniendo en cuenta sus características [12]. Los modelos probabilísticos se aplican principalmente en el campo de la minería de texto, para organizar un corpus de documentos de acuerdo a un conjunto de temas que representa la ocurrencia de temas identificados a partir de esos documentos [8, 36].

El primer estudio computacional con ADN basado en las similitudes de la secuencia por comparación, fue a través de las distancias evolutivas de un conjunto de datos genómicos públicos [25].

Los modelos de temas se propusieron originalmente para el procesamiento de palabras, luego se aplicaron al procesamiento de imágenes y de audio [27, 32], así como al procesamiento de música [31].

Recientemente, algunos investigadores aplicaron el modelado de temas para el procesamiento de datos biológicos, como la extracción de relaciones proteína-proteína a partir de resúmenes científicos de la literatura de MEDLINE [35].

3. Materiales y métodos

3.1. Secuencias de ADN a analizar

Las secuencias de ADN se descargaron en formato fastq del banco de secuencias (ENA) [30].

Estas secuencias de ADN a analizar están compuestas de 155 pares de secuencias crudas para el análisis de vaginosis bacteriana. Las secuencias se tomaron de 48 mujeres embarazadas, de raza caucásica, con inicio espontáneo de parto prematuro con o sin RPM ($\leq 36^{6/7}$ semanas)(casos) y 107 mujeres embarazadas con inicio espontáneo de parto con o sin ruptura prematura (RPM) o cesáreas planificadas a término ($\geq 38^{0/7}$) (controles) [20].

Del banco de secuencias GenBank [6] se tomaron las secuencias válidas del gen ARNr 16S de la región vaginal V2 y V6: *Gardnerella vaginalis* (GenBank, NR118377), *Atopobium vaginae* (GenBank, AF325325), *Prevotella sp* (GenBank, KF007172), *Lactobacillus crispatus* (GenBank, AF243150) y *Lactobacillus inners* (GenBank, AY526083).

Para la revisión manual y comparar los documentos de las secuencias completas de ADN se usa la herramienta RDP [28], como la base de datos principal. RDP es una base de datos en linea que contiene herramientas para el análisis de la secuencia del gen 16S ARN ribosomal [15].

3.2. Gen ARNr 16S

El gen ARNr 16S es considerado una buena opción para la clasificación de las bacterias [26], es como un estándar para identificar microorganismos y se considera la diana universal para la identificación bacteriana a partir del ADN [3, 23].

En este estudio, se consideran las secuencias del gen ARNr 16S de ADN de cinco bacterias presentes en la vaginosis bacteriana y se obtienen del GenBank: *Gardnerella vaginalis, Atopobium vaginae, Prevotella sp, Lactobacillus inner y Lactobacillus crispatus*.

3.3. Modelado de temas (LDA)

El método LDA está basado en el modelado probabilístico [8] Asignación Latente de Dirichlet (LDA) [9] de aprendizaje no supervisado. Es un modelo estadístico aplicado inicialmente a documentos de texto con el objetivo de descubrir los temas de una colección de documentos denominado corpus. Es un modelo que obtiene información de los diferentes temas tratados con la distribución de probabilidad sobre el conjunto de las palabras.

La distribución de probabilidad [24] se expresa como:

$$P\left(w_i\right) = \sum_{j=1}^{T} P\left(w_i | z = z_j\right) P\left(z = z_j\right), \tag{1}$$

donde $P\left(w_i\right)$ es la probabilidad de que la palabra w_i este en un documento dado; $P\left(z = z_j\right)$ es la probabilidad de elegir una palabra del tema z_j para el documento actual; $P\left(w_i | z = z_j\right)$ es la probabilidad de muestreo de la palabra w_i, dado el tema zj y T es el numero de temas.

1 La distribución φ para cada tópico, representa la probabilidad de las ocurrencias de palabras en cada tópico dado, como el conjunto de:

$$\varphi \approx Dirichlet\left(\delta\right), \tag{2}$$

donde $\approx$ quiere decir "está distribuido como".

2 La proporción de Θ es la distribución del tópico para el documento d como el conjunto de:

$$\Theta \approx Dirichlet\,(\alpha)\,. \tag{3}$$

Los modelos de temas LDA, tienen la característica de seleccionar la cantidad de temas que se desean generar para el estudio de la investigación.

3.4. Muestreo de Gibbs

Gibbs es un algoritmo que genera una muestra aleatoria a partir de distribuciones de probabilidades de datos completos o incompletos [21]. Se trata de un algoritmo Metropolis-Hastings y es un método MCMC (Monte Carlo Markov Chain) y gracias a los avances computacionales en la actualidad se aplica en áreas como la biología [22].

La ecuación para encontrar la distribución de probabilidad de la asignación de una sola palabra w en un documento d de pertenecer al tema k esta dada por [18]:

$$p(Z_{d,n} = k|\overrightarrow{Z}_{-d,n}\,, \overrightarrow{W}, \alpha, \lambda) = \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{^V k, W_{d,n} + ^\lambda W_{d,n}}{\sum_i V_{K,i} + \lambda_i}, \tag{4}$$

donde:

$n_{d,k}$: Veces que el documento d usa el tema k

$V_{k,W}$: Veces que el tema k usa la palabra dada w

αk: Parámetro de Dirichlet para la distribución de documentos a temas

λw: Parámetro de Dirichlet para la distribución de tema a palabra

El muestreo Gibbs se configura con valor α = 0.1, puesto que da mejores resultados [24].

3.5. Similitud coseno

Es una medida de similitud y describe el grado de semejanza o disimilitud de los objetos comparados. Su formula esta dada en (5):

$$cos(\theta) = \frac{A.B}{||\ A\ ||||\ B\ ||}. \tag{5}$$

Los objetos se consideran vectores de frecuencia de términos de los documentos y su índice de similitud se mide a partir de la multiplicación de dos vectores (A, B) dividida por el producto de sus longitudes. Este resultado representa el ángulo coseno de los dos vectores. Su valor cae en el rango [0, 1] donde: si el ángulo es cero, su similitud es uno, y cuanto mayor sea el ángulo menor será su similitud [11].

Del resultado LDA, se mide la similitud coseno de cada tema conformado por las subsecuencias k-mers para encontrar una relación entre el conjunto de temas y las secuencias ARNr 16S.

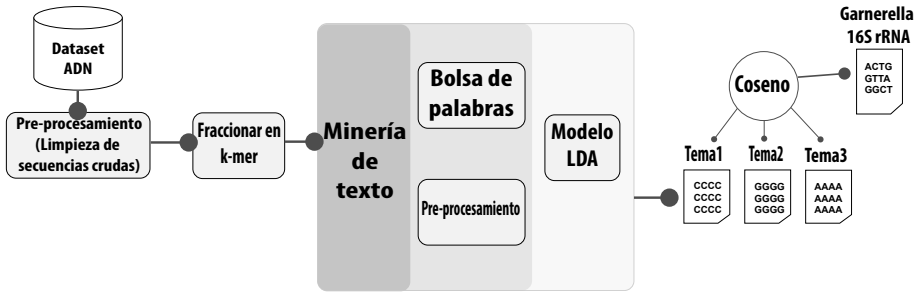

Fig. 1. Flujo de trabajo para el análisis de secuencias de ADN con técnicas de minería de texto.

4. Estudio experimental

En esta investigación se plantea estudiar secuencias de ADN sin métodos tradicionales de alineación para la identificación de vaginosis bacteriana; en su lugar se aplican técnicas de minería de texto. Cada una de las secuencias limpias de ADN representan los documentos y los k-mers las palabras. El desarrollo del trabajo experimental se ilustra en la Figura 1.

4.1. Preparación de las secuencias de ADN

Los procesos de secuenciación, a través de los cuales se generan las secuencias de ADN introducen errores de mala calidad en los datos secuenciados [7]. El conjunto de secuencias de ADN utilizado en este estudio descrito en la sección 3 son secuencias crudas que requirieron un pre procesamiento para identificar y eliminar los posibles errores de secuenciación.

Este pre procesamiento se realizó con el algoritmo DADA2 [7] en el lenguaje de programación estadístico R [20]. DADA2 es un paquete disponible en R y proporciona un conjunto de herramientas para medir la calidad de las secuencias duplicadas y de mala calidad DADA2.

El algoritmo DADA2 [7] recibe como entrada una o muchas muestras de secuencias crudas. Estas muestras están compuestas de dos lecturas separadas $forward$ y $reverse$ realizadas por el secuenciador. DADA2 realiza el pre procesamiento de las secuencias crudas y genera una salida compuesta de un conjunto de secuencias únicas.

La cantidad de secuencias únicas puede variar dependiendo de la configuración de los parámetros en DADA2.

Los valores de los parámetros en este estudio para la función DADA2 son: truncLen=c(260,240) se eliminan los extremos de las secuencias de baja calidad, trimRight = 5 número de nucleótidos que se eliminarán al final de cada lectura de secuencias después de truncLen, maxEE=c(2,3) elimina las lecturas de secuencias con los “errores esperados”, truncQ=2 para leer la primera instancia con el puntaje dado.

DADA2 fusiona las secuencias únicas y se obtienen secuencias completas sin ruido. Estas secuencias fusionadas se almacenan por separado en archivos de texto plano en formato txt. Cada una de las secuencias contienen las lecturas de nucleótidos (A, C, G, T).

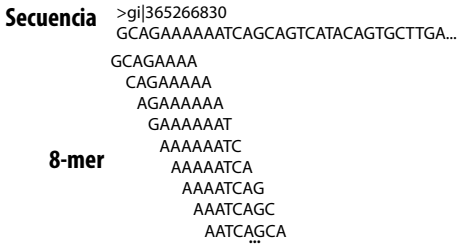

Fig. 2. K-mer de longitud k = 8 nucleotidos.

Cada una de las secuencias limpias representa un documento, por lo tanto el conjunto de secuencias almacenadas en archivos txt puede considerarse como el corpus de los documentos.

Un k-mer es una subsecuencia de ADN, un pequeño fragmento de secuencia definida por un tamaño k, donde k representa la cantidad de nucleótidos que conformara el k-mer. Ver Figura 2.

Las secuencias limpias de ADN obtenidas del pre procesamiento de nucleótidos descrita en esta sección, esta compuesta de una sola cadena de texto definida por los nucleótidos (A, C, G, T). Para extraer las palabras de las secuencias de ADN, se toma cada una de la secuencia almacenada en los documentos y se fragmenta en subcadenas de nucleótidos con la función substring() en R y cada secuencia fragmentada se almacena en un nuevo documento de texto plano txt.

Para este análisis, la longitud de la cadena k-mer es k = 8 nucleótidos. Para decidir el valor de k nucleótidos, se realizaron tres pruebas con k = 8, k = 9 y k = 10, dando mejor resultado k = 8 con mayor puntuación en la similitud del tema generado y la secuencia 16S. En la práctica, la longitud k-mer (k) puede establecerse con seguridad entre 8 y 10 para el gen ARNr 16S, como se describe en [4].

Con los nuevos documentos que contienen la colección de k-mers se crea el corpus de documentos.

Este mismo procedimiento de fragmentado se aplica a las secuencias ARNr 16S de la sección 3.2, utilizadas para medir la similitud coseno con los temas generado por LDA.

4.2. Latent Dirichlet Allocation (LDA)

Con LDA se procesa el corpus de documentos compuesto por la colección de k-mer y genera como resultado una matriz de términos y tópicos como se observa en la Figura 3.

Los términos son la distribución de los k-mers con las probabilidad de pertenecer a uno de los temas generados y los tópicos son los documentos marcados con la probabilidad de pertenecer a un tema.

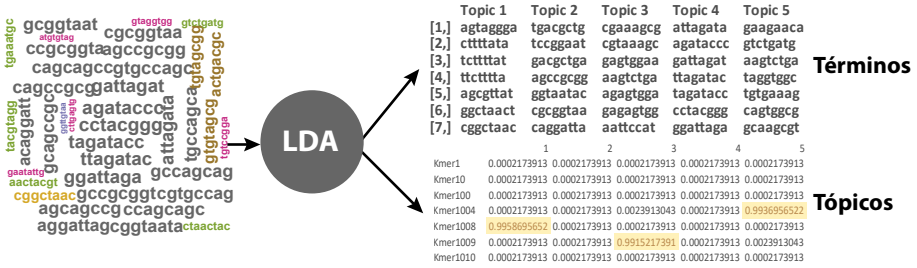

Fig. 3. Términos y Tópicos.

4.3. Interpretación de los temas

A partir del resultado LDA de la Figura 3 descrita en la sección 4.2, LDA solo genera los temas (tema1, tema2, tema3, ...) y no tiene la cualidad de identificar el significado de cada tema generado.

Para identificar y clasificar cada uno de los temas generados por LDA se toman los primeros k-mers con las probabilidades más altas de pertenecer al tema, cuanto más cerca este la probabilidad al valor 1 el k-mer pertenecerá al tópico asignado.

Para asociar los temas desconocidos, se les asigna una etiqueta de la bacteria midiendo la similitud coseno contra el conjunto de bacterias del gen ARNr 16S que conforman la vaginosis bacteriana descrito en la sección 3. Si los temas logran alcanzar una similitud coseno con una puntuación mayor a 40 %, se etiqueta el tema con el nombre del gen comparado, tal como se sugiere en [14]. Ver las tablas de resultados del apartado 5.

El valor del 40 % de la similitud coseno se tomo como umbral debido a que todas las secuencias que cumplieron con este valor como mínimo, se ingresaron al sitio RDP para realizar su búsqueda en la bases de datos de secuencias conocidas correspondiente al gen codificante ARNr 16S. En todos los casos estas secuencias dieron un resultado positivo, es decir la secuencia introducida coincide con la bacteria localizada en el gen ARNr 16S.

Por ejemplo al introducir una secuencia de la bacteria *Atopobium vaginae* se encontró que coincide con la bacteria *Atopobium vaginae* ya conocida en el gen ARNr 16S. Por otro lado, se realizó este mismo procedimiento con secuencias cuyos valores de similitud eran menores a 40 %, y en todos los casos el resultado obtenido de la búsqueda en el sitio RDP fue negativo.

De la matriz de tópicos, a los archivos que obtienen una probabilidad de 99 %, se toma la secuencia completa libre de impureza y se valida manualmente en el banco de secuencias ribosómica RDP 16S disponible en la web [28]. Así, se confirma si la similitud coseno del tema generado por LDA y el gen ARNr 16S etiquetan correctamente a la bacteria del conjunto de vaginosis bacteriana.

Si los resultados en RDP no confirman la existencia de la bacteria etiquetada con la similitud coseno, es un indicativo para cambiar la cantidad de temas a generar en LDA o también los parámetros de limpieza en DADA2.

Tabla 1. Muestra 1.

Temas	**Similitud**	**ARNr 16S**
Tema 1	0.41742100	Prevotella sp
Tema 2	0.51097878	Gardnerella vaginalis
Tema 3	0.19383758	Atopobium vaginae
Tema 5	0.41727181	Lactobacillus Crispatus
Tema 5	0.41524543	Lactobacillus Inner

5. Resultados

Con la finalidad de observar la eficiencia y ajustar los temas idóneos a generar en el algoritmo LDA. Se analizan las variaciones de los resultados de LDA en diferentes situaciones, aumentando la cantidad de documentos de secuencias en el corpus y los temas generados en LDA.

Se toma una secuencia de ADN de las 155 tomadas a las pacientes y posteriormente se realiza una limpieza con DADA2. Se generan 174 documentos de secuencias de ADN limpias de impurezas. Las 174 secuencias de ADN, se fragmentan por separado en secuencias cortas k-mer, generando 174 nuevos documentos que almacenan la colección de k-mer. En total se obtienen $6,913$-mer.

Se crea el corpus de documentos con los documentos k-mer y se generan 5 temas en LDA, la misma cantidad de microorganismo que se desean identificar de la vaginosis bacteriana.

En la Tabla 1 están registrados los valores obtenidos al medir la similitud coseno de los temas generados por LDA y las bacterias del gen ARNr 16S correspondientes a vaginosis bacteriana.

Con excepción del Tema 3 cuya similitud es inferior al 40 %, el resto de los temas superaron este porcentaje. Vea sección 4.3 para la interpretación de este umbral del 40 %.

El Tema 2 con una similitud coseno = 0.51097878 es un indicativo de la presencia del microorganismo *Gardnerella vaginalis*, mientras que el Tema 1 indica la presencia de *Prevotella sp* y en el Tema 5 están presente dos bacterias *Lactobacillus Crispatus y Lactobacillus Inner*.

En la siguiente ejecución se toman 25 pares de secuencia de las 155 pacientes y se obtiene en total 1,651 documentos de secuencias de ADN limpias con un corpus de documentos de 21,889-mers. Se generan 5 temas en LDA. El Tema 2 clasifica positivo para *Gardnerella vaginalis* con una similitud coseno = 0.52801140.

Sin embargo, se hallaron en el mismo tema, documentos que coinciden con el mismo umbral de probabilidad del 99 % y al comprobar en la bases de datos online RDP clasificó para *Gardnerella vaginalis* y otras bacterias tales como: *Actinobaculum massiliense*, *Mobiluncus*, *Corynebacterium* todas pertenecientes a la clase *Actinobacteria*. Con este hallazgo, se modifica la cantidad de temas de 5 a 10 y se ejecuta nuevamente el algoritmo LDA.

Los resultados en esta ejecución para 10 temas, el tema clasificado como *Gardnerella* continua con una similitud coseno = 0.52801140 y ahora comparte características similares a *Bifidobacterium* puesto que aun se tienen documentos clasificados con probabilidades del 99 % de pertenecer al mismo tema.

Tabla 2. 25 pares de secuencias con 15 temas.

Temas	Similitud	ARNr 16S
Tema 7	0.44190317	Prevotella sp
Tema 2	0.52801140	Gardnerella vaginalis
Tema 11	0.36626978	Atopobium vaginae
Tema 12	0.53305327	Lactobacillus Crispatus
Tema 13	0.53621329	Lactobacillus Inner

Tabla 3. 25 pares de secuencias con 20 temas.

Temas	Similitud	ARNr 16S
Tema 6	0.51253985	Atopobium vaginae

Tabla 4. Similitud con 155 muestras y 20 temas.

Temas	Similitud	ARNr 16S
Tema 4	0.45783722	Gardnerella vaginalis
Tema 8	0.52961422	Atopobium vaginae
Tema 10	0.54504403	Prevotella sp
Tema 14	0.45047193	Lactobacillus Crispatus
Tema 16	0.52807122	Lactobacillus Inner

Se incrementa la cantidad de temas a 15 y se ejecuta LDA. *Gardnerella vaginalis* continua con una similitud coseno = 0.52801140 y al revisar los documentos con las probabilidades de 99 % de pertenecer al Tema 2 como se muestra en la Tabla 2, se observo que los documentos marcados con esta probabilidad solo pertenecían a *Gardnerella vaginalis*.

Con 15 tópicos, la similitud coseno=0.36626978 para el Tema 11 de la Tabla 2 es muy baja para *Atopobium* y no se puede determinar si existe o no documentos de secuencias con presencia de la bacteria. Se ejecuta nuevamente LDA generando 20 temas y la similitud para *Atopobium vaginae* incrementa como se muestra en la Tabla 3.

Si la cantidad de temas generados por LDA son insuficientes para un corpus de documentos demasiado grande, los términos contenidos en los documentos se mezclan en un mismo tema cuando los términos comparten características similares.

Después de este análisis previo en los diferentes resultados del algoritmo LDA variando la cantidad de temas a generar y documentos en el corpus, se toman las 155 pares de secuencias de ADN de las muestras tomadas a las pacientes, del resultado de la preparación y limpieza de las secuencias se obtiene un total de 5,933 documentos de secuencias limpias y se generan 35,446-mers en total.

En esta ejecución se generan 20 temas, siendo un indicador razonable considerar [29]. En esta ejecución se atribuyen los mayores pesos probabilísticos a los documentos donde las palabras están principalmente correlacionadas y clasifican para los microorganismos presentes en la vaginosis bacteriana, estos resultados se muestran en la Tabla 4.

Con el aumento en la cantidad de temas, la distribución de los k-mers se organizan correctamente en los temas correspondientes. Al observar los resultados de la Tabla 4, la similitud coseno de 0.45783722 calculada entre el Tema 4 y el gen codificante ARNr 16S de las bacterias que componen la vaginosis bacteriana, este valor es suficiente tal

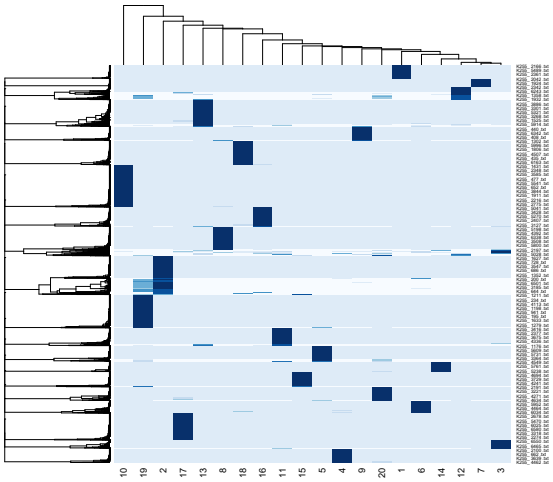

Fig. 4. LDA, extracción de los temas contenidos de la colección de k-mer. Cada rectángulo de color azul representa un tema.

como se describe en la sección 4.3 y en [14] para etiquetar el Tema 4 con la bacteria correspondiente a *Gardnerella vaginalis*. De la misma forma se etiquetan los Temas 8, 10, 14 y 16 con sus bacterias correspondiente.

En la Figura 4 se visualizan los 20 temas generados de los 155 pares de secuencias de ADN. El eje x representa la diversidad cualitativa de los microorganismos presentes en el microbioma vaginal.

En el eje y están los documentos con las probabilidades del 99 % y representan la diversidad cuantitativa en orden según la relevancia del tema. Cada bloque azul marino concentra los documentos con mayor probabilidad de estar asociados a un tema.

6. Conclusiones

Gracias a la información disponible en los bancos de secuencias del GenBank es posible descargar secuencias parciales específicas del gen ARNr 16S.

Los métodos probabilísticos son eficaces en el tratamiento de los datos. Aunque clasifican muy bien, es importante generar la cantidad óptima de temas que pueden estar presentes en un corpus de documentos.

El desarrollo de esta investigación como primer paso se tomó en cuenta la calidad de las secuencias de ADN, se buscó la calidad óptima en el filtrado y corte de error de las secuencias crudas con el pre procesamiento de secuencias de la herramienta DADA2.

El segundo paso consistió en transformar las secuencias limpias en secuencias cortas llamadas k-mer. Estas secuencias cortas tomaron el papel de palabras y se almacenaron en documentos para ser importados en los modelados de temas probabilísticos.

LDA es una técnica de minería de texto basada en modelos probabilísticos que identifica el contenido de cada documento y los clasifica en temas.

Un punto importante del algoritmo LDA, es que no identifica cada uno de los temas generados, se puede inferir de acuerdo al contenido de las palabra asociadas a cada tema.

La interpretación de los tópicos, puede llegar a ser compleja y gracias al gen ARNr 16S en nuestro análisis la interpretación de la composición de cada tópico se puedo identificar midiendo la distancia coseno. Estos nuevos métodos de comparación de secuencias sin alineación, mejoran el rendimiento computacional [1]. Los métodos tradicionales de alineamiento de secuencias, tienen la cualidad de procesar la información con mayor tiempo y recurso computacional.

Como línea de investigación futura, se busca complementar todos los microorganismos que conforman la vaginosis bacteriana, en conjunto con las líneas de investigación clínica para un pronóstico confiable y su diagnostico oportuno.

Referencias

1. Zielezinski, A., Vinga, S., Almeida, J., Karlowski, W. M.: Alignment-free sequence comparison: Benefits, applications, and tools. Genome biology, vol. 18, no. 1, pp. 1–17 (2017) doi: 10.1186/s13059-017-1319-7
2. CIAD: Ecología microbiana, secuenciación masiva y bioinformática (2016)
3. Moya, A. S.: Microbioma y secuenciación masiva. Revista Espsñola de Quimioterapia, vol. 30, no. 5, pp. 305–390 (2017)
4. Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M., Ragan, M. A.: Inferring phylogenies of evolving sequences without multiple sequence alignment. Scientific Reports, vol. 4, no. 1, pp. 1–9 (2014) doi: 10.1038/srep06504
5. Lledó-Bosch, B.: Efecto del microbioma vaginal en la tasa de embarazo en pacientes que se someten a técnicas de reproducción asistida (2018)
6. Ouellette, B. F., Rapp, B. A., Wheeler, D. L., Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J.: Genblack. Nucleic Acids Research, vol. 27, no. 1, pp. 12–17 (1999) doi: 10.1093/nar/27.1.12
7. 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
8. Blei, D. M.: Probabilistic topic models. Communications of the ACM, vol. 55, no. 4, pp. 77–84 (2012) doi: 10.1145/2133806.2133826
9. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research, vol. 3, pp. 993–1022 (2003)
10. Financiera, D.: El microbioma vaginal y su relación con el comportamiento, la salud sexual y las enfermedades de transmisión sexual. Obstetrics & Gynecology, vol. 129, pp. 643–54 (2017) doi: 10.1097/AOG.0000000000001932
11. Liu, D., Chen, X., Peng, D.: Some cosine similarity measures and distance measures between q-rung orthopair fuzzy sets. International Journal of Intelligent Systems, vol. 34, no. 7, pp. 1572–1587 (2019) doi: 10.1002/int.22108
12. Gamage, G., Gimhana, N., Wickramarachchi, A., Mallawaarachchi, V., Perera, I.: Alignment-free whole genome comparison using k-mer forests. In: 19th International Conference on Advances in ICT for Emerging Regions, vol. 250, pp. 1–7 (2019) doi: 10.1109/ICTer48817.2019.9023714
13. Fan, H., Ives, A. R., Surget-Groba, Y., Cannon, C. H.: An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC genomics, vol. 16, no. 1, pp. 1–18 (2015)

14. Choi, I., Ponsero, A. J., Bomhoff, M., Youens-Clark, K., Hartman, J. H., Hurwitz, B. L.: Libra: Scalable k-mer–based tool for massive all-vs-all metagenome comparisons. GigaScience, vol. 8, no. 2 (2019) doi: 10.1093/gigascience/giy165
15. Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M., Tiedje, J. M.: The ribosomal database project (RDP-II): Sequences and tools for high-throughput rRNA analysis. Nucleic Acids Research, vol. 33, no. 1, pp. D294–D296 (2005) doi: 10.1093/nar/gki038
16. Marrazzo. J. M.: Interpreting the epidemiology and natural history of bacterial vaginosis: Are we still confused? Anaerobe, vol. 17, no. 4, pp. 186–190 (2011) doi: 10.1016/j.anaerobe.2011.03.016
17. Gamboa-Unsihuay, J. E.: Topic modeling en twitter: Determinación de la agenda política peruana en el periodo de enero a setiembre del 2018. Anales Científicos, vol. 80, no. 2, pp. 308–327 (2019)
18. Boyd-Graber, J., Blei, D., Zhu, X.: A topic model for word sense disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1024–1033 (2007)
19. Nieto, J. J.: Comparación de secuencias (2005)
20. 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, pp. 201 (2019) doi: 10.3389/fmed.2019.00201
21. Angelone, L.: Análisis del muestreo Gibbs para detección de motivos en secuencias biológicas (2005)
22. Moreno-Arévalo, M. A.: Descripción e implementación del muestreador de Gibbs en versión bivariada (2016)
23. del-Rosario-Rodicio, M., del Carmen-Mendoza, M.: Identificación bacteriana mediante secuenciación del ARNr 16s: Fundamento, metodología y aplicaciones en microbiología clínica. Enfermedades infecciosas y microbiología clínica, vol. 22, no. 4, pp. 238–245 (2004) doi: 10.1016/S0213-005X(04)73073-6
24. la Rosa, M., Fiannaca, A., Rizzo, R., Urso, A.: Probabilistic topic modeling for the analysis and classification of genomic sequences. BMC Bioinformatics, vol. 16, no. 6, pp. 1–9 (2015) doi: 10.1186/1471-2105-16-S6-S2
25. Nei, M., Kumar, S.: Molecular evolution and phylogenetics. Oxford University Press (2000)
26. Pace, N. R., Stahl, D. A., Lane, D. J., Olsen, G. J.: The analysis of natural microbial populations by ribosomal RNA sequences. Advances in microbial ecology, vol. 9, pp. 1–55 (1986) doi: 10.1007/978-1-4757-0611-6_1
27. Elango, P. K., Jayaraman, K.: Clustering images using the latent dirichlet allocation model. University of Wisconsin, pp. 1–18 (2005)
28. Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R.: Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, vol. 73, no. 16, pp. 5261–5267 (2007) doi: 10.1128/AEM.00062-07
29. Arun, R., Suresh, V., Veni Madhavan, C. E., Murthy, N.: On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, vol. 6118, pp. 391–402 (2010) doi: 10.1007/978-3-642-13657-3_43
30. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Hoopen, P. T., Vaughan, R., Zalunin, V., et al.: The european nucleotide archive. Nucleic Acids Research, vol. 39, pp. D28–D31 (2010) doi: 10.1093/nar/gkq967

31. Zhang, R., Cheng, Z., Guan, J., Zhou, S.: Exploiting topic modeling to boost metagenomic reads binning. BMC Bioinformatics, vol. 16, no. 5, pp. 1–10 (2015) doi: 10.1186/1471-2105-16-S5-S2
32. Kim, S., Narayanan, S., Sundaram, S.: Acoustic topic model for audio information retrieval. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 37–40 (2009) doi: 10.1109/ASPAA.2009.5346483
33. Shehab, S. A., Keshk, A., Mahgoub, H.: Fast dynamic algorithm for sequence alignment based on bioinformatics. International Journal of Computer Applications, vol. 37, no. 7, pp. 54–61 (2012)
34. Vinga, S., Almeida, J.: Alignment-free sequence comparison—a review. Bioinformatics, vol. 19, no. 4, pp. 513–523 (2003) doi: 10.1093/bioinformatics/btg005
35. Aso, T., Eguchi, K.: Predicting protein-protein relationships from literature using latent topics. Genome Informatics 2009: Genome Informatics Series, vol. 23, pp. 3–12 (2009) doi: 10.1142/9781848165632_0001
36. Griffiths, T. L., Steyvers, M.: Finding scientific topics. In: Proceedings of the National Academy of Sciences, vol. 101, pp. 5228–5235 (2004) doi: 10.1073/pnas.0307752101