# EXPLORACIÓN Y COMPARACIÓN DE MÉTODOS DE INTELIGENCIA ARTIFICIAL PARA LA CLASIFICACIÓN TAXONÓMICA EN ANÁLISIS METAGENÓMICOS


**WIDERMAN STID MONTOYA RAMÍREZ**


**Trabajo de grado para optar al título de**

**Ingeniero informático**


**ISIS BONET CRUZ**

**Ph.D**




**ESCUELA DE INGENIERÍA DE ANTIOQUIA**
**INGENIERÍA INFORMÁTICA**
**ENVIGADO**
**2014**

# AGRADECIMIENTOS

A mi familia a quienes son lo más importante en mi vida, también muy agradecido con mis profesores, compañeros y todas las personas que me ayudaron y me dieron fuerzas para estar aquí en este momento.

# CONTENIDO

pág.

# LISTA DE TABLAS

pág.

# LISTA DE FIGURAS

pág.

# LISTA DE ANEXOS

pág.

# RESUMEN

La mayor diversidad genética está presente en las comunidades de microorganismos, el conocer estas especies, sus funciones y diferencias constituye un papel importante para solucionar problemas diversas áreas, como la salud, la alimentación y el medio ambiente. El método tradicional para realizar este tipo de investigaciones consiste en aislar el microorganismo de una muestra del entorno y así estudiar su constitución genética, sin embargo menos del 1% de los microorganismos pueden ser aislados y cultivados en los laboratorios. Gracias a las técnicas de secuenciación modernas cada vez más accesibles surge la metagenómica proponiendo una alternativa para poder estudiar el otro 99%. La metagenómica se encarga de estudiar la secuenciación de una muestra del entorno para descubrir a qué organismos pertenecen los fragmentos secuenciados. Sin embargo el problema radica en que los procesos necesarios para identificar el tipo de organismos en la muestra demandan mucho tiempo y recursos computacionales.

En este trabajo se utilizan diferentes algoritmos de inteligencia artificial para agrupar los fragmentos de secuencias según su similitud en conjuntos puros, es decir, conjuntos cuyos fragmentos pertenezcan a un solo organismo o a un mismo grupo taxonómico de organismos. Además se propone un nuevo algoritmo que se basa en la aplicación del k-means de manera iterativa perfeccionando los grupos según la distancia entre ello. Se compararon los resultados con métodos de agrupamientos clásicos y se comprobó que con este último método se obtienen grupos más puros. Este resultado ayuda a que los procesos de ensamblado o de comparación serán más eficientes y rápidos, debido a que se tiene como entrada inicial una muestra más condensada y uniforme, disminuyendo el tiempo y los recursos consumidos durante los proyectos metagenómicos, al mismo tiempo que pueden realizarse de una forma más enfocada.



Palabras clave: Metagenómica, taxonomía, clusterización, K-mer, inteligencia artificial

# ABSTRACT

The biggest genetic diversity is present in the microorganisms communities, to know these communities their functions and differences plays an important role at the moment to solve problems in many areas as the health, environment and more. The classic method to realize these research kinds consist in incubate the microorganism's cells in the laboratory and then study their genetic information, but is proven that only less of the 1% of the microorganisms cells can be culturable in laboratory media. Thanks to the new sequencing techniques emerge the metagenomic proposing a new alternative to study the other 99%. The metagenomic is in charge of study the study sequence the environment sample and study their genetic constitution and find the microorganisms fragments that compose the sample. However the problem is that the process required to identified the organisms in the sample demands much **time and computational resources**

During this exploration are use different Artificial intelligence algorithms to group the sequence fragments into bins depending of their similarity, in other words, sets where their fragments are belong to the same organism or the same taxonomical group. It also propose a new algorithm based on the K-means application in an iterative manner, improving the bins quality depending of their distance. The results where compared from different classic binning methods and was probe that with the last method the cleanest bins were found. This result helps to the assemble or comparison process to be faster and more efficient, due to it has as input a more uniform and condensed sample, reducing the time and the consumed resources during the metagenomic process are more focused.



Key words: Metagenomic, clustering, K-mer, artificial intelligence, taxonomy

# INTRODUCCIÓN

La mayor diversidad genética del planeta está presente en las comunidades de microorganismos (Papamichail, Skiena, Lelie, & Mccorkle, 2004). Entender estas especies, sus diferencias y funciones constituye la clave para la solución de diversos problemas relacionados con la salud (Fancello, Raoult, & Desnues, 2012), el medio ambiente (Yousuf, Keshri, Mishra, & Jha, 2012), alimentación, entre otros. Estas especies, en su gran mayoría, no son cultivables (Saleh-Lakha et al., 2005), es decir, dada una muestra del entorno no es posible aislar y cultivar las células de un microorganismo para extraer su información genética.

Gracias a los avances tecnológicos, que permiten secuenciar el ADN cada vez a costos más bajos (Hall, 2007), surge la metagenómica. Este nuevo campo de la investigación se dedica al estudio de la composición genética de las comunidades de microorganismos presentes en diversos tipos de ecosistemas, agua, rizosfera (Yousuf et al., 2012) e incluso en el interior de seres humanos (Fancello et al., 2012) y animales. A partir de secuenciar y estudiar muestras tomadas directamente del entorno es importante conocer los microorganismos están presentes en dichas comunidades y sus diferentes funciones, aprovechando la capacidad de procesamiento de los equipos modernos (Wu & Ye, 2011), esto suprime la necesidad de separar y extraer los microorganismos de la muestra para cultivarlo de forma individual. Se tiene una larga y compleja cadena de nucleótidos que representa la información genética de diferentes individuos, sobre la cual el nuevo objetivo es encontrar a que individuos pertenecen las diferentes secuencias, o al menos identificar los diferentes tipos de microorganismos que componen la muestra y sus diferentes funciones. Durante los estudios metagenómicos, se llevan a cabo varias tareas, entre las cuales pueden estar incluidas, la definición del perfil taxonómico, identificación de las funciones metabólicas, ensamblaje de secuencias, análisis comparativos, entre otras.

Una de las tareas más importantes es el ensamblaje de secuencias. Las herramientas que realizan dichos procesos se dividen en 2 grandes tipos (Ngs, n.d.): basados en composición y en alineamientos. Los ensambladores basados en la composición de las secuencias, también conocidos como Novo, hacen un estudio no supervisado sobre los diferentes segmentos generados a partir de la muestra metagenómica, ensamblando las secuencias que tengan segmentos en común (Zerbino & Birney, 2008), se dividen en tres tipos dependiendo del método que implementen, ensambladores Greedy (Logares et al., 2012), los OLC y los ensambladores basados en grafos de Brujin (Namiki, Hachiya, Tanaka, & Sakakibara, 2012). El segundo tipo de ensambladores son los que se basan en alineamientos, estos realizan el ensamblado haciendo comparaciones contra una base de datos (Reddy, Mohammed, & Mande, 2012). En la tabla 1 se muestran algunos ensambladores, su tipo, el método que implementan, y la plataforma de secuenciación a la cual pueden ser aplicados.

Las herramientas informáticas encargadas de ejecutar estas funciones para hacer una identificación acertada de los genes, reciben como insumo largos conjuntos de información generados por las plataformas de secuenciación, esto conlleva a problemas de rendimiento y almacenamiento (Zakrzewski et al., 2013). Un ejemplo de esto son las técnicas de alineamiento como el BLAST, que clasifican secuencias genéticas contra una compleja base de datos y que en la mayoría de los casos resulta ser un proceso desgastante y demorado (Reddy et al., 2012).

La propuesta realizada en esta exploración, para mitigar el consumo de tiempo y recursos, es segmentar las secuencias que componen la muestra metagenómica usando métodos de aprendizaje no supervisado. De esta manera las herramientas que realizan las tareas en los procesos metagenómicos posteriores, no tendrán que trabajar con toda la información que tienen las secuencias del entorno, en su lugar van a trabajar con una muestra reducida y segmentada . Por ejemplo en los procesos de comparación ya no sería necesario hacer una comparación contra todos los elementos de la base de datos, en su lugar se podrán realizar comparaciones más orientadas, debido a que la mayoría de las secuencias de entrada pueden compartir cierto parentesco taxonómico.

Basados en la exploración de varios proyectos metagenómicos, donde se analizaron las herramientas, métodos y rasgos utilizados, se seleccionaron aquellos con los que se habían obtenido mejores resultados. Se escogieron varios rasgos para describir las secuencias que son comparados y finalmente se seleccionan los que mejores resultados muestran. Además se escogieron k-means y expectation maximization como algoritmos de clusterización clásicos muy utilizados en esta área. Para mejorar los resultados obtenidos hasta el momentos y buscando encontrar métodos que me generaran clústeres de elementos puros se propone un método basado en el k-means. Se comparan todos estos métodos, con los diferentes rasgos seleccionados.

# 1. PRELIMINARES

## 1.1 PLANTEAMIENTO DEL PROBLEMA

A partir de los nuevos métodos de secuenciación genética (NGS), capaces de secuenciar mayor cantidad de información en menor tiempo y a menor costo, surge la metagenómica, que pretende estudiar la composición genética de las comunidades de microorganismos a partir de secuenciar una muestra completa del entorno. Existen muchas aplicaciones informáticas que son capaces de identificar cada fragmento a qué organismo pertenece. Pero reunir e identificar esos fragmentos es una tarea que demanda más tiempo y recursos para la entrega de los resultados debido a que utilizan como insumo la información genética de todo el entorno y estos archivos tienden a ser muy pesados. Gran parte de estos sistemas se basan en hacer comparaciones de los segmentos de la secuencia metagenómica contra secuencias conocidas de una base de datos. Las desventajas en este tipo de sistemas, además de no reconocer secuencias de organismos nuevos que no se encuentren en sus bases de datos, es que el número de comparaciones a realizar demanda una gran cantidad de tiempo. Si bien en este momento hay sistemas que basados en la composición de los fragmentos de la muestra buscan ensamblar los fragmentos en secuencias conocidas, aún hay que mejorar esos tiempos de respuesta. En cambio, si desde el principio de la investigación se agrupan las secuencias genéticamente semejantes, los procesos posteriores de clasificación y ensamblaje de las secuencias podrán ser realizados de una manera más rápida y económica.

## 1.2 OBJETIVOS DEL PROYECTO

### 1.2.1 Objetivo General

Agrupar secuencias genómicas, a partir de bases de datos de varias muestras metagenómicas, para mejorar el tiempo de su posterior clasificación taxonómica, utilizando métodos de inteligencia artificial.

### 1.2.2 Objetivos Específicos

- Explorar y estudiar sistemas informáticos y métodos de inteligencia artificial utilizados en estudios metagenómicos, así como la forma de representación de las secuencias y las bases de datos utilizadas.

- Seleccionar los métodos de inteligencia artificial y la representación de datos que producen una mejor agrupación taxonómica, a partir de la comparación de los

resultados de diferentes métodos, utilizando diferentes bases de datos metagenómicos y diferentes formas de representación de los datos.

- Utilizar los métodos de inteligencia artificial seleccionados para realizar pruebas de agrupamiento de las secuencias con una base de datos propia.

## 1.3 MARCO DE REFERENCIA

### 1.3.1 Aspectos teóricos

La genómica está compuesta por un conjunto de ciencias y disciplinas que se basan en el conocimiento y la caracterización del genoma (Laviades, 1999), la cual según el DRAE es "El Conjunto de los genes de un individuo o de una especie, contenido en un juego haploide de cromosomas". Gracias a los avances realizados por empresas como Roche (Bekel et al., 2009) e Illumina (Logares et al., 2012) que permiten hacer secuenciar las cadenas de ADN en sus diferentes nucleótidos (A, C, G y T) a menor costo y en menor tiempo, es posible hacer una exploración de las comunidades de microorganismos y sus funciones. El método clásico para identificar un microorganismo dentro de un ambiente, consiste en separar y cultivar individualmente dicho microorganismo, a fin de poder extraer su información genética de manera directa. Sin embargo, está comprobado que está metodología no funciona para más del 99% de los microorganismos, ya que no son cultivables (Hall, 2007)(Saleh-Lakha et al., 2005). La metagenómica es una ciencia que surge precisamente para enfrentar este reto, estudiar los microorganismos que no pueden ser aislados a partir de muestras del entorno (Wu & Ye, 2011). La meta es la reconstrucción de genomas completos a partir de los fragmentos que se obtienen con las herramientas de secuenciación genómica(Singh, Gautam, Verma, Kumar, & Singh, 2008).

Históricamente en la comunidad científica un acercamiento concreto a la metagenómica aparece por primera vez en 1998 (Handelsman, Rondon, Brady, Clardy, & Goodman, 1998), como nueva alternativa al nuevo reto de esa época, encontrar la secuencia genética sin necesidad de aislar y clonar la muestra, teniendo como significado literal "Más allá del genoma"(Susana & Raimondi, n.d.). En general durante los estudios metagenómicos se llevan a cabo tareas que van desde la agrupación de las secuencias generando perfiles taxonómicos hasta la identificación de secuencias completas y funciones metabólicas (Zakrzewski et al., 2013). Dentro de estas tareas, la clasificación taxonómica o al menos la agrupación de diferentes subconjuntos de la secuencia juega un papel muy importante en este tipo de estudios (Reddy et al., 2012), ya que de su óptima implementación, depende la efectividad de los resultados en las tareas posteriores. Comúnmente la metagenómica aborda el problema de identificación y separación de las secuencias a partir de dos perspectivas. La primera consiste en un proceso de alineamiento basado en algoritmos de comparación directa (BLAST) (Altschul et al., 1997) contra una base de secuencias conocidas, a fin de encontrar especies conocidas. Sin embargo, este proceso tiene la desventaja de consumir muchos recursos en procesamiento y ser poco eficiente debido a la gran cantidad de información a comparar. También tiene problemas para secuencias cortas ya que sus características pueden

coincidir con varias partes de secuencias conocidas (Wu & Ye, 2011). La segunda perspectiva está basada en métodos de composición, los cuales tratan de identificar patrones dentro de la secuencia completa, que permitan caracterizar y agrupar segmentos similares con rasgos similares (Li & Waterman, 2003)(Reddy et al., 2012).

**Tabla 1.** Ensambladores basados en composición y comparación (Ngs, n.d.)

| Nombre | Tipo | Método | Plataformas | Autor |
|---|---|---|---|---|
| SSAKE | de novo | Greedy | Solexa | Warren, R. et al. |
| SHARCGS | de novo | Greedy | Solexa | Dohm et al. |
| VCAKE | de novo | Greedy | Solexa | Jeck W. et al. |
| Newbler | de novo | Greedy/OLC | 454, Sanger | 454/Roche |
| Celera Assembler | de novo | OLC | Sanger | Myers G. et al. |
| Arachne | de novo | OLC | 454, Solexa | Batzoglou S. et al. |
| CAP | de novo | OLC | 454, Solexa | Kolehmainen et al. |
| PCAP | de novo | OLC | 454, Solexa | Kolehmainen et al. |
| CABOG | de novo | OLC | Sanger, 454, Solexa | Miller G. et al. |
| Euler | de novo | DBG | Sanger, 454 | Pevzner P. et al. |
| Velvet | de novo | DBG | Sanger, 454, Solexa, SOLiD | Zerbino D. et al. |
| ABySS | de novo | DBG | Solexa, SOLiD | Simpson J. et al. |
| AllPaths | de novo | DBG | Solexa | Butler J. et al. |
| SOAPdenovo | de novo | DBG | Solexa | Li R. et al. |

Es necesario destacar, que el proceso de agrupamiento de las secuencias pertenecientes a diferentes microorganismos, puede llevarse a cabo asignando las lecturas a grupos taxonómicos (que comparten características similares), o grupos filogenéticos (dado su línea de evolución), sin embargo, en la mayoría de los casos los resultados de diferentes agrupaciones no difieren mucho, dado que las especies genéticamente cercanas suelen compartir las mismas características.

Para los investigadores del Centro Nacional de Secuenciación Genómica una agrupación inicial de los segmentos de la secuencia metagenómica, podría significar un filtro considerable al momento de hacer un procesamiento previo de la información con el fin de reducir el volumen de la información a manejar, optimizando así los procesos utilizados en los estudios metagenómicos.

Con el paso de los años, la mayoría de los sistemas de información utilizados para ejecutar las tareas de clasificación y agrupamiento de las secuencias, han comenzado a implementar diversos métodos de inteligencia artificial, enfocados en la clasificación y el agrupamiento de los datos (Logares et al., 2012; Papamichail et al., 2004; Reddy et al., 2012; Wu & Ye, 2011; Yousuf et al., 2012).

Algunos de los métodos de inteligencia artificial utilizados para la clasificación de los datos, contra una base de conocimiento establecida son:

- Naive Bayes: Opuesto al 1R debido a que este usa todos los atributos, les da igual valor y asume su independencia, en este modelo el valor de la clase se denomina como la evidencia, y se presume que cada evidencia es consecuencia de todos los atributos de la instancia (Ali, Shamsuddin, & Ismail, 2012; Farid, Zhang, Rahman, Hossain, & Strachan, 2013).

- Árbol j48: Selecciona un atributo para ser el nodo raíz, luego cada una de sus ramas son los posibles valores del atributo, los nodos hijos serán los demás atributos y el árbol se irá construyendo en el orden que vaya generando mayor ganancia de la información, es decir valores más puros de la clase (Andrea & Hern, 2004; Farid et al., 2013).

- KNN (haciendo variaciones al valor de K): Clasifica cada una de las instancias con el valor más repetidos de sus k vecinos cercanos (dada una función de distancia para definir tal cercanía) (Saini, Singh, & Khosla, 2013).

- Redes neuronales: estos modelos tratan de imitar la estructura y el proceso de aprendizaje de las neuronas biológicas, a un ritmo de aprendizaje y retroalimentación (Javier, Germ, Una, & Neuronal, n.d.; Salas, n.d.).

Estos métodos anteriores son supervisados, es decir, se basan en bases de datos de organismos reconocidos, por lo que pueden servir para reconocer los algoritmos que se encuentren en su base de datos.
Cuando la investigación se realiza de forma no supervisada, es decir, para agrupar simplemente los segmentos de la misma secuencia que son semejantes entre sí, sin necesidad de comparar contra una base de datos externa, se utilizan métodos de agrupamiento, como por ejemplo:

- Expectation Maximization (EM): El cual se basa en el estimador de máxima verosimilitud de la distribución de los datos, estos son los estimadores que maximizan la semejanza entre los datos, para esta implementación podemos parametrizar para un número de grupos óptimo, y un grupo que se ajuste al número de valores de la clase, sin embargo la teoría también afirma que es conveniente aumentar el número de clúster para generar grupos cada vez más limpios (Shokrzadeh, Khorsandi, & Haghighat, 2012)(Wu & Ye, 2011).
- K-means: Este modelo funciona a partir de la parametrización del número de grupos, luego el define los centros para cada grupo de forma arbitraria, el paso siguiente es coger cada uno de los elementos restantes, y calcular su distancia con respecto a cada centro finalmente lo agrupa al centroide con el que tuvo la menor distancia, al terminar la iteración define nuevos centros para cada clúster, y reinicia el proceso, hasta cumplir el número de iteraciones debido. Estas simulaciones se realizaron con la distancia euclidiana y Manhatan (Zheng, Yoon, & Lam, 2014).
- DB scan: Es un método de densidad al igual que el Optics, funciona con un mínimo de puntos cercanos a una distancia dada, sin embargo para esta investigación se suprimió el Optics por sus in aportantes resultados, ya que dejaba muchos elementos sin agrupar.
- Optics: Este algoritmo al igual que el DB Scan a partir de 2 parámetros ε(Distancia máxima) y mp(Puntos mínimos por grupo), genera grupos de al menos mp miembros, que estén separados por una distancia máxima ε, sin embargo a diferencia del DB Scan es que este algoritmo tiene en cuenta los elementos que hacen parte de la frontera del clúster.

### 1.3.2 Estado del arte

La mayoría de métodos supervisados requieren comparaciones contra una base de datos de muestras ya conocidas, y dado que el 99% de los microorganismos no pueden ser aislados(Saleh-Lakha et al., 2005) para ser estudiados de manera individual, en la mayoría de los casos las bases de datos podrían no tener la suficiente información para hacer una comparación precisa. Además hacer la comparación directa de toda la muestra metagenómica contra todas las secuencias de la base de datos es un proceso que consume muchos recursos. Precisamente teniendo en cuenta esto, en este trabajo vamos a utilizar métodos no supervisados.

Entre los proyectos estudiados se encontraron varias tendencias en cuanto a las técnicas, las funciones de distancia y los criterios de comparación. En Estados Unidos unos investigadores realizaron un análisis de ciertas comunidades de bacterias (Papamichail et al., 2004), el estudio lo hicieron partiendo del sistema clásico basado en el algoritmo de clasificación Naïve Bayes utilizando como rasgo para describir las secuencias el vector de frecuencias K-mer para la clasification realizado por (Sandberg et al.). Este tipo de rasgo consiste en la frecuencia con la que aparece una combinación de K o L nucleótidos en la cadena dada. El uso de Naive Bayes hace la comparación de los k-mers basada en probabilidades condicionales, es decir dado que se tiene un segmento de nucleótidos, calculan la probabilidad de que aparezca el K-mer N, dado que apareció el K-mer N-1, luego la probabilidad más alta ayuda a secuenciar e identificar las diferentes cadenas.

A diferencia del proyecto anterior un conjunto de estudiantes en la universidad de indiana abordaron la problemática de la metagenómica como un problema de aprendizaje no supervisado desarrollando Abundance bin (Wu & Ye, 2011). Un nuevo sistema para la agrupación de secuencias metagenómicas estableciendo como primer hecho que el orden de los nucleótidos en una secuencia dada sigue una distribución de Poisson, dada esta hipótesis, se utiliza el algoritmo EM para determinar los parámetros de la distribución que mejor se acomodan a la muestra dada, para luego realizar la agrupación. Los rasgos que se utilizaron en este proyecto fueron las frecuencias de nucleótidos conocidas como K-mer o L-tuplas (Lebovka, Khrapatiy, & Pivovarova, 2014).También el año pasado dos científicos en España propusieron un método para hacer agrupaciones y clasificaciones taxonómicas(Castellanos-Garzón & Díaz, 2013). Inicialmente se parte de agrupar las secuencias genéticas en clústeres jerárquicos (similar a una clasificación taxonómica), sin embargo el algoritmo para agrupar estas secuencias en una jerarquía (dendograma), tiende a caer en soluciones locales. Una forma de solucionar este problema, según los autores de este artículo, es a través de algoritmos genéticos los cuales a partir de diferentes dendogramas, tratan de llegar a una solución global del problema.

Otra forma de abordar el problema puede ser optando por una solución híbrida como la que crearon diferentes investigadores en la India presentando en el 2012 TWARIT(Reddy et al., 2012), una herramienta extremadamente rápida (comparada con las que fueron analizadas en esta investigación) orientada a la clasificación de secuencias apoyada en procesos de clusterización, en general el algoritmo evalúa inicialmente la longitud de la secuencia, si esta es menor a 150bp, utiliza el BWA TOOL. Si la longitud supera los 150 bp realiza un HPBA un algoritmo de 'hit-pair based assignment' (HPBA) para hacer comparaciones con los dos extremos de la secuencia, en lugar de toda la secuencia, básicamente el alineamiento se logra cuando los extremos y la distancia que los separa coincide para este procedimiento los rasgos analizados son los extremos de la cadena, y la longitud intermedia entre los extremos, para hacer el proceso de alineación directa. Cuando el BWA o el HPBA no generan los resultados esperados, se recurren al 'signature sorting based assignment' (SSBA) un algoritmo de clusterización que utiliza K-means (Manhatan 256D frecuencia de nucleótidos(Wu & Ye, 2011), también es lo mismo que utilizar el K-mer con K=4) para agrupar todas las secuencias parecidas luego con la distancia de cada una de las secuencias a cada uno de los centroides se genera una firma de referencia(concatenándose) y la base de datos se ordena con respecto a esa

firma, dejando los elementos vecinos con mayor parecido, para hacer posteriormente la comparación. Así mismo en Canadá el año pasado, un grupo de investigadores desarrollaron Phoenix2 (Soh, Dong, Caffrey, Voordouw, & Sensen, 2013) Básicamente este algoritmo en su flujo de trabajo incorpora varios tipos de algoritmos, primero hace una limpieza de los datos, eliminando lecturas repetidas, o que generan ruido, luego hace un breve alineamiento, posteriormente con las secuencias alineadas se realiza una clusterización iterativa basada en la matriz de distancia entre las secuencias, definiendo las diferentes OTUs(Unidades taxonómicas), para finalmente hacer la asignación taxonómica final, con las secuencias representativas de cada OTU.

Algunos investigadores en lugar de optar por hacer una clasificación o una agrupación de las secuencias similares se basaron en mejorar la calidad de las muestras para optimizar los experimentos con muestras más compactas y útiles como lo hicieron varios investigadores de china presentando este año Meta-QC-Chain (Zhou, Su, Jing, & Ning, 2014). Esta herramienta sirve para evaluar la calidad de las muestras metagenómicas, utilizando diferentes métodos de pre filtrado, como longitud de las lecturas, algunos marcadores biológicos, recortes de las lecturas, entre otros. Una de las primeras evaluaciones más relevantes que efectúa esta herramienta, es mirar el contenido de GC la aplicación está disponible para el consumo público en la siguiente dirección: http://computationalbioenergy.org/meta-qc-chain.html. El otro procedimiento importante dentro de su flujo de trabajo, es identificar las posibles especies eucariotas, ya que según los autores, generan contaminación en las muestras metagenómicas. En los primeros pasos del flujo de trabajo se tratan de identificar las secuencias GC, que suelen repetirse contaminando las muestras, Mientras que en el procedimiento que pretende eliminar la información de especies eucariotas utilizan el marcador 18S rRNA, como un marcador biológico de las células eucariotas (y que podría no otorgar mucha información), posteriormente, hacen una alineación con una base de datos para identificar los elementos que contaminan para extraerlos.

Por otro lado existen herramientas con el objetivo de analizar las diferentes secuencias de la muestra metagenómica y ensamblar los segmentos que tienen nucleótidos en común. MetaCAA (Reddy, Mohammed, & Mande, 2014), es una metodología para ensamblar la información metagenómica de una manera eficiente. Este trabajo fue realizado en la India y presentado este año, su procedimiento se divide en tres pasos, el primero es agrupar los segmentos de la secuencia metagenómica utilizando el método de "Ángulo Coseno", en el cual recibía como parámetro la frecuencia de tetra nucleótidos de las secuencias también conocido como el k-mer con k=4(Reddy et al., 2012). Luego todos los elementos de cada clúster son ensamblados con la ayuda de la herramienta para ensamblar secuencias de tipo "Greedy" llamada CAP3, y en el último paso las secuencias que no fueron agrupadas en el paso anterior son agrupados y ensamblados nuevamente en el CAP3.

Otro ejemplo hablando de las herramientas para ensamblar secuencias genéticas es la que realizaron Investigadores Japoneses en 2012 MetaVelvet (Namiki et al., 2012) que extiende las capacidades de otra herramienta para ensamblar secuencias llamada Velvet (Zerbino & Birney, 2008), básicamente MetaVelvet está optimizados para secuencias

cortas, básicamente parten del principio que el grafo de Bruijn al ser construido a partir de una secuencia metagenómica de diferentes especies, en esencia este también puede estar construido de pequeños grafos, también argumentan que dos especies evolutivamente separadas no comparten ningún K-mer, luego los grafos de Bruijn que estas generarían serían dos grafos independientes. A partir de una secuencia meta genética, realizan el grafo de Bruijn, luego crean un histograma de frecuencias de los diferentes nodos del grafo, los picos generados en el histograma representan el cubrimiento que tiene cierta especie en la muestra, y por tanto con los nodos pertenecientes a ese pico se calcula la probabilidad de cubrimiento aproximando los picos a distribuciones gaussianas, a partir de estas probabilidades son generados los sub grafos desde los nodos que tienen varias entradas y son considerados como quimeras (proceden de varias especies), finalmente sobre estos sub grafos se pueden generar las secuencias y sus clasificaciones taxonómicas. Para este trabajo, sobre cada secuencia se toman los diferentes K-mer, 2 k-mer consecutivos se solapan en K-1 nucleótidos estos se convierten en los nodos del grafo de Bruijn y los k-mer unen esos nodos, teniendo un grafo el objetivo es encontrar el camino euleriano sin embargo, para este proyecto los caminos se hallan sobre los sub grafos, y estos se extraen de los picos en el histograma generado por el peso de los nodos.

Otros proyectos se enfocaron en generar rasgos para los métodos tanto de clasificación como de clusterización que optimizaran dichos algoritmos, como por ejemplo la propuesta que se le hace a mejorar las operaciones realizadas con los ya mencionados vectores K-mer, ya que al condensar las secuencias, en este vector, según los autores(Wen, Chan, Yau, He, & Yau, 2014) se pierde información evolutiva, por tanto se reemplaza el K-mer convencional, con el K-mer natural un nuevo vector que garantiza correspondencia 1 a 1, lo cual implica que cualquier secuencia puede ser identificada inequívocamente con su vector K-mer natural. La mejora en esta metodología radica principalmente en la composición del rasgo para realizar las operaciones de clusterización, el vector natural se compone a partir de concatenar los 3 siguientes vectores, el vector K-mer común v1=(n1, n2, nl), donde n es el número de veces que se repite el segmento de longitud k, en la secuencia de nucleótidos de longitud l, el segundo vector V2= (u1, u2, ul) donde cada u(i) es el promedio aritmético desde la ocurrencia de cada K-mer a la primera base de la secuencia, el último vector v3=(D1, D2, ...Dl) es el vector de momentos centrales normalizado, finalmente este rasgo permite definir a cada una de las lecturas, y la distancia entre ellas. Otro ejemplo de los proyectos enfocados en la condensación de las secuencias genéticas para su análisis es la representación que se hace de las secuencias genéticas como matrices de densidad de los nucleótidos en donde Martin T. Swain (Swain, 2013) realizó un Proyecto de rápida comparación de secuencias de micro organismos usando la “Chaos Games Representation” para aplicaciones metagenómicas, el proceso consiste en graficar los nucleótidos de una secuencia de forma iterativa con una función acumulativa (en la que el resultado producido por el nucleótido n, depende del resultado de f(n-1)) en un espacio bidimensional, delimitado en un plano unitario, luego se la matriz unitaria se divide en r regiones permitiendo calcular la densidad de puntos por región y finalmente dependiendo de la densidad de las regiones, se hace una comparación de tipo linear (cadena a cadena) contra una base de datos conocida. Según

la investigación es mucho más óptimo comparar las densidades de las matrices que las cadenas de nucleótidos.

# 2. METODOLOGÍA

Para realizar la exploración de los estudios metagenómicos, serán consultadas distintas bases de datos académicas tales como ScienceDirect, Scopus, e investigación del trabajo realizado por parte de instituciones locales, la información extraída será estructurada en un cuadro comparativo, en el cual se hará énfasis en variables cualitativas como el sistema de información utilizado para el proyecto, los métodos de inteligencia artificial que implementó dicho sistema de información, la muestra utilizada, y las caracterización de esa muestra escogida para realizar el estudio. También como parte cuantitativa de la extracción se va a consignar en la tabla, los indicadores del error simple o la precisión para así poder comparar de forma numérica los resultados de los diferentes estudios. Además según la disponibilidad de la información presentada en los artículos, se tomará en forma cuantitativa o cualitativa la estimación de los tiempos y recursos optimizados en el estudio, con el fin de resaltar los métodos más óptimos para el desarrollo de los proyectos.

A partir de la información obtenida en el comparativo, se va a realizar una selección cualitativa, buscando los proyectos que tengan mayor similitud en las muestras estudiadas, y las herramientas seleccionadas, para así generar posteriormente la selección de los métodos más precisos al momento de agrupar las secuencias, y que sean afines a la base de conocimientos con la que serán probados.

Finalmente las simulaciones y evaluaciones de los métodos, serán realizadas con el módulo libre de inteligencia artificial WEKA, el cual contiene diversas y flexibles implementaciones de varios métodos de agrupación y clasificación. Con los resultados obtenidos en estas simulaciones finalmente se permiten descubrir los métodos y mecanismos capaces de agrupar las secuencias metagenómicas similares, sin tener que encontrar necesariamente el grupo específico de correspondencia, para hacer el análisis visual de los resultados, vamos a usar la versión de prueba de la herramienta Tableau (http://www.tableausoftware.com/).

# 3. APLICACIÓN DE LAS TÉCNICAS RECOLECTADAS

Durante la primera parte de esta exploración se estudiaron diversos proyectos realizados alrededor del mundo con el fin de generar o mejorar métodos y sistemas que apoyen los estudios metagenómicos dichos trabajos implementaron diversos métodos desde modelos Bayesianos (Papamichail et al., 2004) hasta algoritmos evolutivos (Castellanos-Garzón & Díaz, 2013). Algunos proyectos presentaban valor agregado en la característica de comparación de las secuencias utilizando funciones o modelos para representar la secuencia de una forma más resumida y no tener que hacer las comparaciones nucleótido por nucleótido(Swain, 2013; Wen et al., 2014). Estos estudios sirvieron para complementar substancialmente el marco de referencia de esta, dichos proyectos constituyen la base para elegir los métodos, rasgos y funciones de distancia que serán comparados para encontrar que combinaciones de estas características pueden agrupar de manera más efectiva las secuencias genéticas que mejor se relacionen entre sí.

Al trabajar con los grupos sintetizados en lugar de toda la muestra los estudios metagenómicos podrán realizarse de una forma más rápida y dirigida, debido a que van a trabajar con un subconjunto de la muestra, y las comparaciones de ese subconjunto ya no serán tan “a ciegas” porque la muestra tiene características en común.

## 3.1 DATOS

En total 166618 secuencias genéticas ensambladas fueron descargadas del FTP del instituto Sanger, compuestas por virus, bacterias, y eucariotas, ilustradas en la siguiente tabla.

**Tabla 2 Secuencias y fuentes de datos**

| **Organism** | **Data source** |
|---|---|
| Aspergillus fumigatus | ftp://ftp.sanger.ac.uk/pub/project/pathogens//A_fumigatus/AF.contigs.031704 |
| Ascaris suum | ftp://ftp.sanger.ac.uk/pub/project/pathogens//Ascaris/suum/genome/assembly/contigs.fasta |
| Dengue | ftp://ftp.sanger.ac.uk/pub/project/pathogens//Dengue/Dengue.fasta |
| Glossina | ftp://ftp.sanger.ac.uk/pub/project/pathogens//Glossina/morsitans/Assemblies/tsetseGenome-v1.tar.gz |
| Bacteroides dorei | ftp://ftp.sanger.ac.uk/pub/project/pathogens//Bacteroides/dorei/D8/454LargeContigs.fna |
| Bifidobacterium longum | ftp://ftp.sanger.ac.uk/pub/project/pathogens//Bifidobacterium/longum/454LargeContigs.fna |
| Candida parasilopsis | ftp://ftp.sanger.ac.uk/pub/project/pathogens//Candida/parapsilosis/contigs/CPARA.contigs.fasta |

| | |
|---|---|
| Influenza | ftp://ftp.sanger.ac.uk/pub/project/pathogens//Influenza/Santiago_7981_06.fasta |

Las secuencias pertenecientes al reino de las bacterias, compuestas por Bacteroides dorei con un total de 1948 lecturas con un tamaño máximo de 83484 bases y Bifidobacterium longum con un total de 2377370 bases.

Las Eucariotas elegidas fueron 2 hongos el “Mold aspergillus fumigatus” y el “Yeast candidda parasilopsis” un nematodo y un insecto.

Las secuencias de virus están compuestas por 64 secuencias de Denge, donde su longitud varía entre 10392 y 10785 bp y su promedio de contenido G+c es de 45.95%, 8 secuencias de influenza con una longitud menor en comparación a las secuencias de Denge desde 853 hasta 2309.

## 3.2 K-MEANS

El K-means es uno de los métodos más utilizados para los ejercicios de agrupación de datos, este algoritmo genera iterativamente un conjunto de K centroides y asocia cada dato de la muestra al centroides más cercano, dependiendo de la función de distancia, cuando todos los datos están asignados a un centroides en específico. Se calculan los nuevos centroides de cada conjunto y se repite la el proceso iterativamente hasta que la función converja, o se llegue a un límite de iteraciones parametrizado. Dicho método fue utilizado para identificar bacterias en una muestra heterogénea(Papamichail et al., 2004). En esta simulación se utilizó la función de distancia euclidiana (Ecuación 1) la cual será tenida en cuenta para parametrizar el algoritmo durante las simulaciones de esta investigación. El K-means parametrizado con la función de distancia Manhattan (Ecuación 2) también fue usado para la construcción de la herramienta TWARIT (Reddy et al., 2012) que mostró contundentes tiempos de respuesta al ser comparada con herramientas similares. También será utilizada la función Coseno (Ecuación 3) implementada en la herramienta MetaCAA (Reddy et al., 2014) para realizar la clusterización en la primera fase del proceso, también usado para medir la distancia entre los vectores de frecuencia K-mer Naturales (Wen et al., 2014).

$$dE(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

$$dM(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (2)$$

$$dC(x,y) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i} \times \sqrt{\sum_{i=1}^{n} x_i}} \quad (3)$$

Donde los vectores X y Y representan las dos instancias de longitud *n* a comparar.

## 3.3 EXPECTATION MAXIMIZATION

El segundo método de clusterización que será utilizado en las simulaciones es el Expectation Maximization (EM), un método de clusterización que busca el estimador de máxima verosimilitud que describa como distribuir la muestra dada de una muestra dada, alternando de forma iterativa dos pasos el paso Expectation (E) que consiste en el cálculo de la esperanza a partir de la suposición de algunos parámetros de entrada, y en el paso de la maximización(M) en donde a partir de la esperanza calculada se generan los nuevos parámetros que serán usados en el paso de E de la siguiente iteración hasta que el algoritmo converga.(Keshavarz & Huang, 2014). El algoritmo está propuesto especialmente para problemas en donde hay información faltante, por tal motivo se usa mucho en los estudios metagenómicos, dado a uno de los mayores retos en estos estudios es la inconsistencia de la información. Una de las propiedades de este algoritmo afirma que bajo condiciones usuales, la precisión del algoritmo aumenta mientras lo hace el volumen de información a analizar. Sin embargo este método tiende a caer en mínimos locales, dado a que una función de probabilidad puede tener máximos estimadores de máxima verosimilitud puede no tenerlo. Este método fue implementado en el proyecto Abundance Bin (Wu & Ye, 2011), en el que establecen que el comportamiento de las frecuencias K-mer sigue una distribución Poisson. También fue utilizado en una investigación que pretendía estimar las especies en una muestra metagenómica (Seok, Hong, & Kim, 2014) asumiendo que su comportamiento seguía una distribución gaussiana.

## 3.4 K-MEANS ITERATIVO

Es una extensión del algoritmo K-means, si bien es sabido que el K-Means ya es un algoritmo iterativo, en esta implementación la estrategia es ejecutarlo de forma completa repetidamente, reduciendo el conjunto de información a agrupar tras cada ejecución del algoritmo. En esta investigación se demostró que bajo ciertas circunstancias, los grupos más alejados tienden a tener los resultados menos útiles para el proyecto. Este hecho es muy importante para la investigación, dado que si en los proyectos metagenómicos no se conoce bien la pureza o los resultados de cada grupo, si se puede saber la distancia entre ellos, y bajo esa premisa se procede a elaborar una matriz de distancia entre los grupos para identificar los kf (K farthest, más lejanos) grupos con la mayor distancia promedio y realizar nuevamente el algoritmo K-means para agrupar las secuencias pertenecientes a los kf grupos más alejados con el fin de generar nuevos grupos con mejores resultados e incorporarlos a los grupos cercanos generados en la iteración anterior.

La ilustración 1 muestra el proceso del algoritmo, donde se parte de la base de datos metagenómica original y se ejecuta k-means. A los k grupos obtenidos se les aplican métodos de medición de distancia intra-grupo e inter-grupo. Los grupos que están muy cercanos entre sí, pueden mezclarse, formando un único grupo. Los grupos que tienen una distancia intra-grupo muy grande, se considera que no es un grupo compacto y pueden existir muchas secuencias que no se parezcan, o sea está propenso a ser dividido al menos una vez más. Estos grupos con estas características (poco compactos) son mezclados y con estas secuencias se vuelve a construir una base de datos que es ahora la entrada al k-means para volver a generar k grupos. Este proceso se repite N veces hasta que los grupos generados son compactos. La salida del algoritmo es la unión de los grupos compactos generados en cada iteración de k-means.

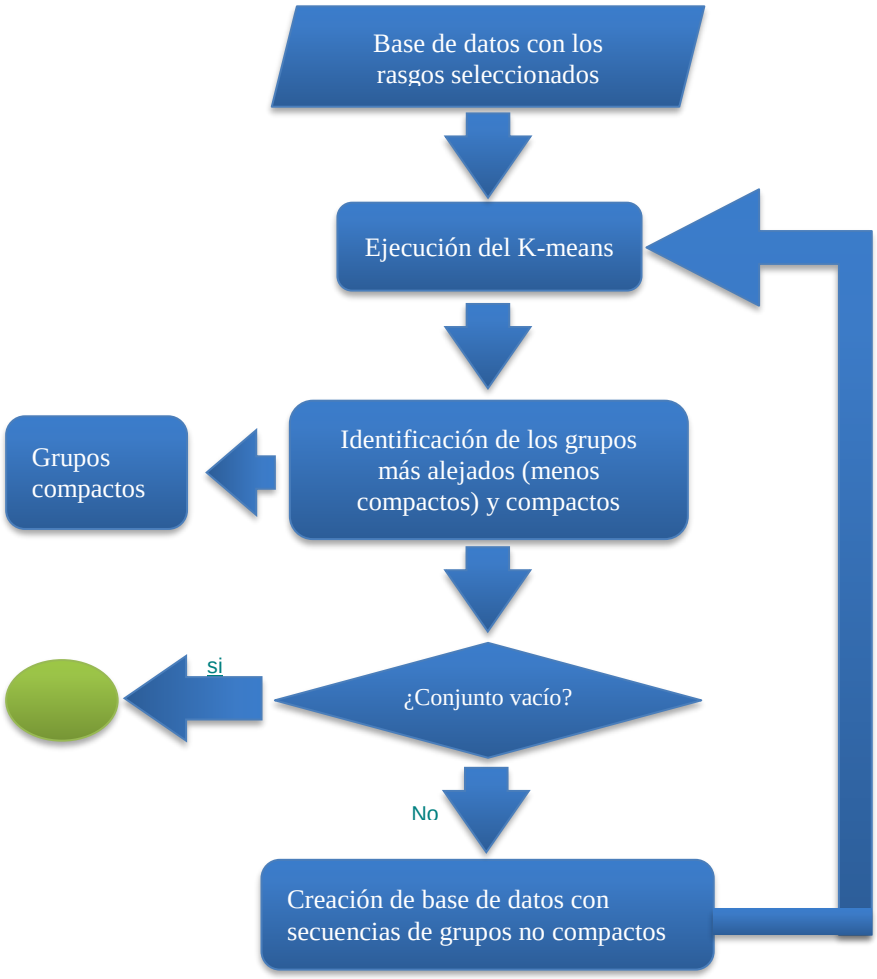


**Ilustración 1 Proceso Kmeans Iterativo**

## 3.5 CARACTERÍSTICAS

Para hacer las comparaciones de una manera más ágil y óptima entre las secuencias, se van a generar representaciones más compactas de las cadenas de nucleótidos sin que éstas pierdan sus características esenciales.

El primer rasgo a considerar para las pruebas será el vector de frecuencias K-mer debido a su amplia aplicación en varios de los proyectos investigados (Papamichail et al., 2004), aunque en algunas investigaciones se le dio el nombre de “l-tuples”(Li & Waterman, 2003; Wu & Ye, 2011), también se le dan diferentes nombres dependiendo el valor de K, por ejemplo con k = 3 se conoce este rasgo como codones, y con k = 4 se utiliza como

frecuencia de tetranucleótidos (Reddy et al., 2012; Wu & Ye, 2011) un vector de 256 dimensiones.

El otro rasgo a utilizar será el contenidos G+C, el cual indica la razón entre el contenido G y C con respecto la longitud total de la cadena, mencionados y utilizados en algunas investigaciones como una característica comparación entre diferentes secuencias (Seok et al., 2014; Wu & Ye, 2011), también se va a añadir esta característica al vector de frecuencia K-mer con el fin de generar un nuevo rasgos en el cual el contenido G+C complemente la información proporcionada por el vector de frecuencias.

$$GC = \frac{G + C}{A + G + C + T}$$

Finalmente el último rasgo tenido en cuenta para el análisis es la frecuencia de nucleótidos (nucl), un vector de 4 dimensiones compuesto por en la participación o razón que tiene cada una de las diferentes bases sobre la longitud total de la secuencia N, y se representa mediante la siguiente expresión.

$$nucl = < \frac{A}{N}, \frac{T}{N}, \frac{G}{N}, \frac{C}{N} >$$

Finalmente con cada uno de los cuatro tipos de rasgos mencionados se va a generar un archivo de 166618 registros, en el que cada fila va a representar la traducción de la secuencia original para dicha característica. Al hacer las comparaciones con estos archivos y no con las secuencias originales el objetivo es lograr tiempos de respuesta más óptimos, sin perder la precisión del sistema. Además de eso se van a generar nuevos rasgos a partir de la permutación de los cuatro rasgos iniciales, por ejemplo un rasgo <GC, nucl>, un vector de 5 dimensiones en el que la primera es su contenido GC, y las otras cuatro son las dimensiones de las frecuencias de nucleótidos.

# 4. DISCUSIÓN DE RESULTADOS

A partir de la base de datos original, se generaron 15 archivos producto de los 4 rasgos propuestos y sus posibles combinaciones. El contenido GC, frecuencia de tetranucleótidos( 4mer) , codones (cod), y la frecuencia simple de nucleótidos (nucl), GC-nucl, GC-cod, GC-nucl-cod-4mer, y así sucesivamente.

Cada simulación generó por cada instancia un número entero, que representa el número del grupo al que se asigna dicha secuencia y en el caso del K-means se generaron las distancias entre cada uno de los grupos con respecto a los demás grupos, para el EM no se pudo generar este último resultado dado a que no es un método basado en distancia. Para cada archivo se evaluó el K-means 6 veces, variando las 3 funciones de distancias y los valores de K =15 y K = 20. Con el algoritmo EM, se realizaron menos simulaciones, debido a que este algoritmo consumía muchos recursos y demoraba mucho tiempo en comparación con las simulaciones realizadas para el K-means, razón por la cual no puede ser considerado como un método que optimice los procesos metagenómicos. Finalmente para el K-means por todos los archivos y todas sus variaciones se realizaron 90 simulaciones, mientras que el EM fue evaluado 5 veces.

## 4.1 ANÁLISIS GENERAL DE LOS RESULTADOS

Luego de extraer y tabular los resultados de las 95 simulaciones, es necesario definir cuáles simulaciones entregaron los mejores resultados, teniendo en cuenta que los mejores resultados son los que generan los grupos más limpios, es decir los que generen el mayor número de grupos que solo tengan una clase de secuencias en común, ya sea a nivel de dominio (grupos que solo contengan bacterias, solo eucariotas, o solo virus) o a un nivel más específico del árbol taxonómico. Por tal motivo se presenta el indicador de pureza del método, que se define como el porcentaje de grupos limpios generados por la simulación.

$$pureza = \frac{Grupos\ con\ 1\ taxón}{Número\ de\ grupos}$$

Donde la pureza es 0 si todos los grupos generados por la simulación tienen secuencias de varios grupos taxonómicos, evidenciando la baja capacidad de la configuración del método para identificar cadenas genéticas con el mismo taxón. Estos resultados son válidos para estas simulaciones debido a que tenemos los atributos de cada secuencia, pero en la aplicación no tenemos este tipo de información. Por tal motivo, se desea

seleccionar los métodos y configuraciones que generen la mayor pureza de los grupos, para que en la aplicación se garantice cierto nivel de confianza sobre la limpieza de los grupos generados por la simulación, para que posteriormente puedan ser usados de una manera más óptima y enfocada en los demás procesos de los estudios metagenómicos.

Los mejores resultados en términos de limpieza para las simulaciones realizadas con el K-means se lograron haciendo las simulaciones con:

- los rasgos combinados de nucl-cod (frecuencia de nucleótitos y codones) y nucl-cod-4mer (frecuencia de nucleótidos, codones y k-mer con k=4).
- la función de distancia Coseno
- K=20.

Se logró una limpieza del 90%, es decir de los 20 grupos que formaron las simulaciones, 18 quedaron totalmente limpios. Si bien estos rasgos funcionaron bien con la función de distancia Coseno, no presentan resultados tan favorables con las demás funciones de distancia, por ejemplo el nucl-cod-4mer con la función de distancia Manhattan obtuvo 0 y 5% de pureza respectivamente, y el rasgo nucl-cod para la función de distancia euclidiana logró una pureza de 0% para K = 20. Sin embargo cuando se combinan estos rasgos con la función de distancia Coseno el funcionamiento del modelo, presenta los mejores resultados, que mejoran a medida que aumenta el valor de K.

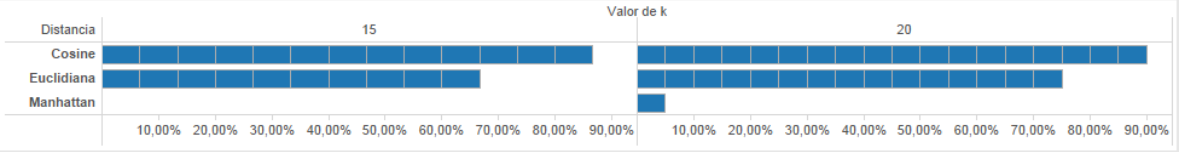


*Figura 1 Resultados de limpieza para las simulaciones con nucl-cod-4mer*

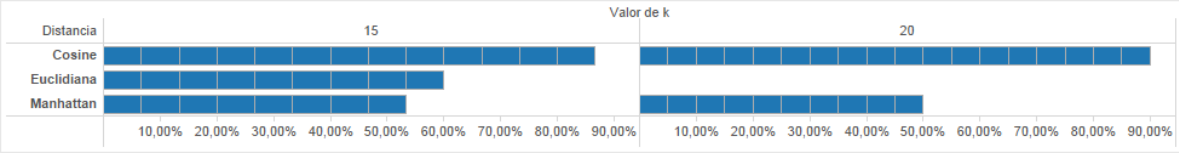


*Figura 2 Resultados de limpieza para las simulaciones con nucl-cod*

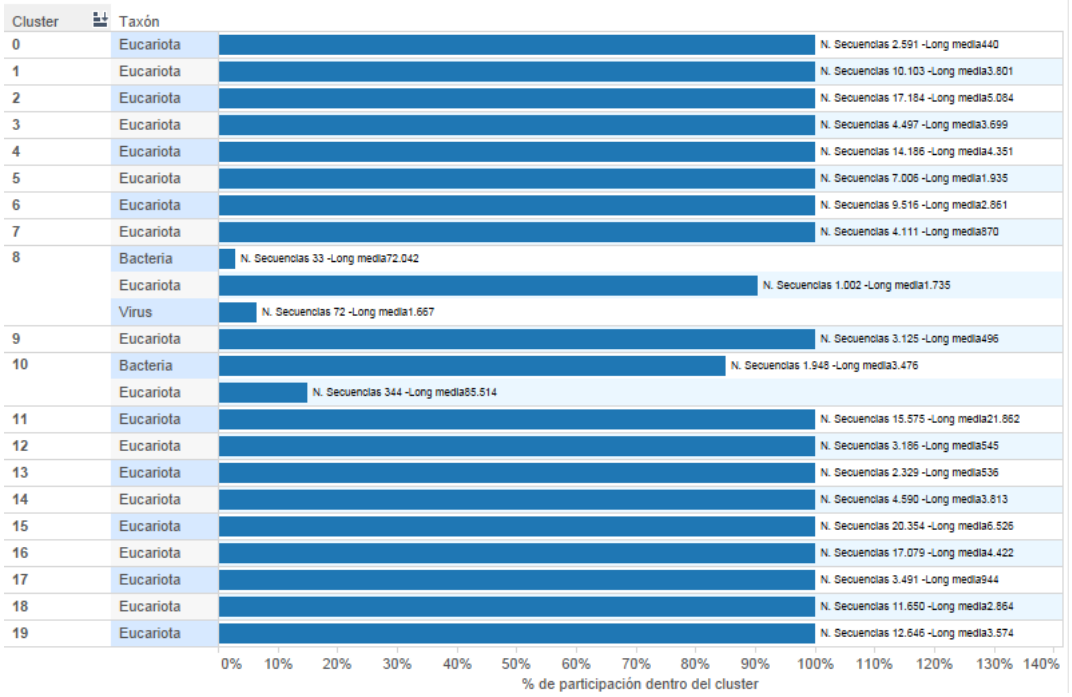


*Figura 3 Descripción de los grupos generados con el rasgo nucl-cod-4mer, coseno y K=20*

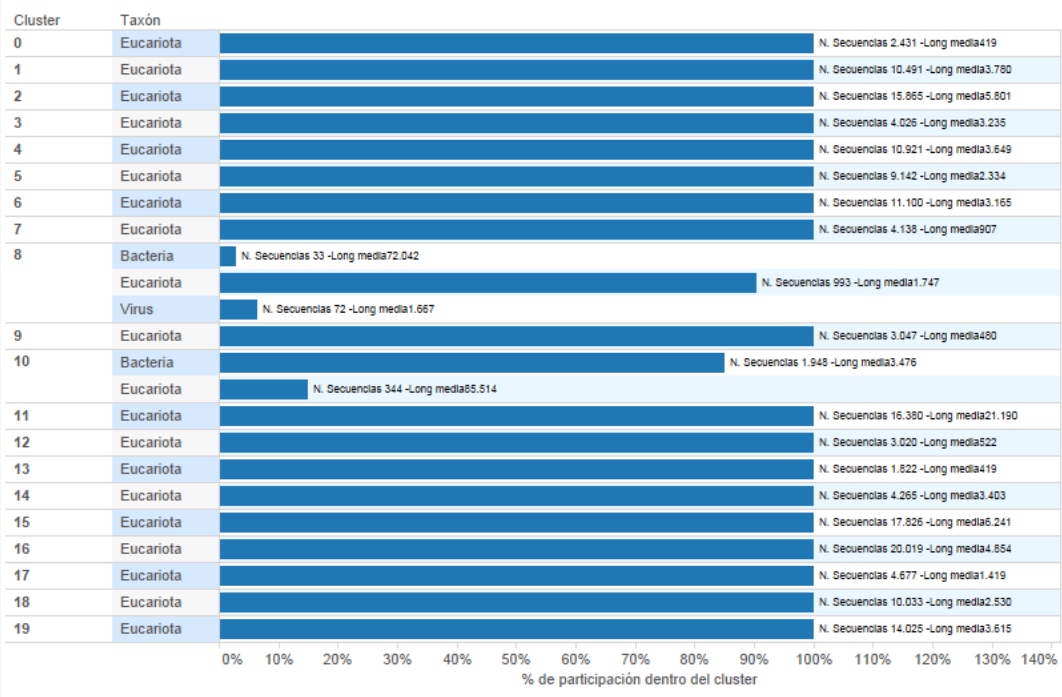


*Figura 4 Descripción de los grupos generados con el rasgo nucl-cod, coseno y K=20*

Si bien no hay mucha diferencia entre la descripción de los grupos generados en la Figura 3 y la Figura 4, para las demás simulaciones con el K-means, las diferencias son más significativas, entre un rasgo y otro, como por ejemplo los resultados presentados en la figura 8, usando el rasgo GC.

Es muy relevante señalar la composición de la base de datos en cuanto distribución de los diferentes taxones, los virus con 72 secuencias no representa ni el 1% del total de las secuencias de las muestra y sin embargo en las simulaciones anteriores, el algoritmo parametrizado con estos dos rasgos y la distancia coseno fue capaz de identificar su similitud entre ellos, a pesar de la inferioridad numérica como lo podemos observar en la simulación con una limpieza del 85% realizada con el rasgo nucl-cod-4mer con k=15 también en la Figura 5.

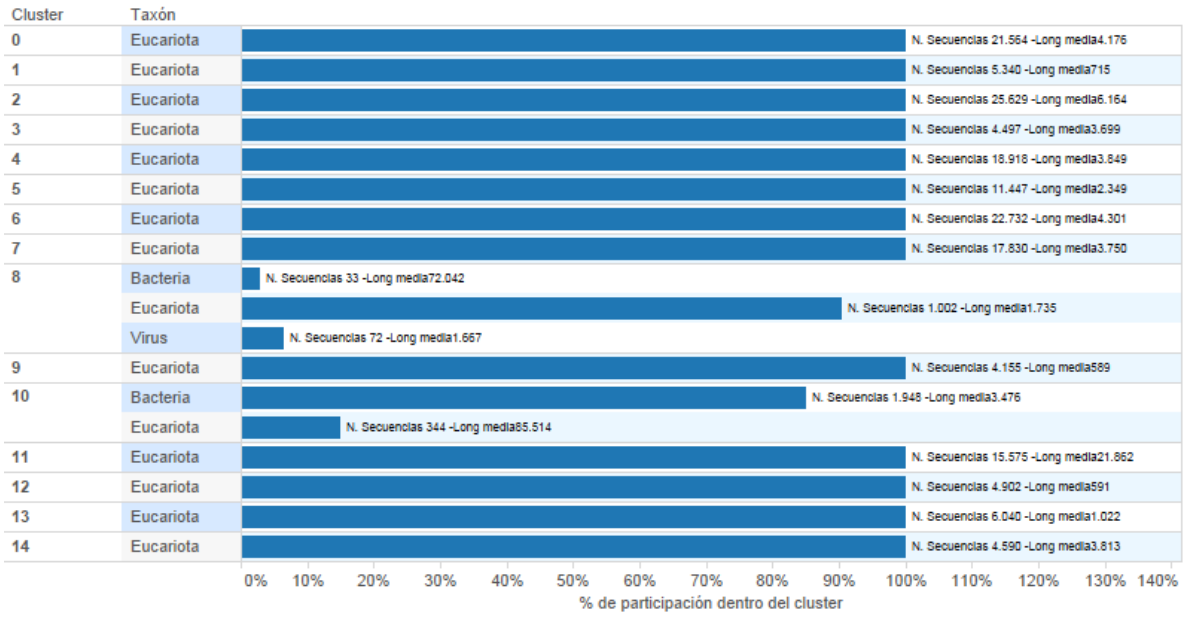


*Figura 5 Descripción de los grupos generados con el rasgo nucl-cod-4mer, coseno y K=15*

Los resultados más equilibrados para el proyecto fueron generados por el rasgo nucl, debido a que el objetivo del estudio es encontrar algoritmos que ayuden a agilizar los procesos durante los estudios metagenómicos, por ende la ejecución del algoritmo tiene que ser más eficiente que los mismos procesos posteriores a este. Y mientras más rápido sea conservando la suficiente precisión en sus resultados, más van a aportar a la solución del problema inicial. La eficiencia de este rasgo, se debe a que representa toda la cadena de nucleótidos en un vector de 4 dimensiones, dicho rasgo unido con la distancia coseno y k=20 generó resultados con una limpieza del 85% figura 6, en tiempos menores a los generados por los dos rasgos anteriores debido a que el K-means, para generar los resultados del rasgo nucl-cod-4mer tiene que operar con un vectores de 324 dimensiones, estamos hablando de más o menos 10 minutos de ejecución contra casi 3 horas que demoró el K-means parametrizado con el segundo rasgo.

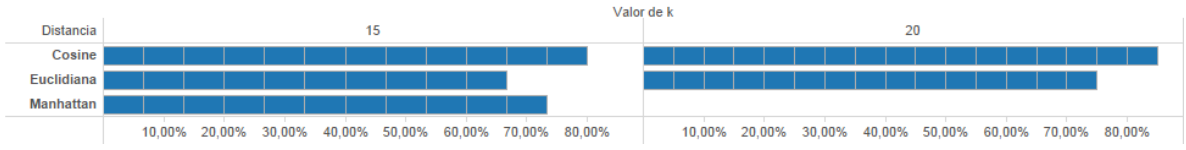


*Figura 6 Resultados de limpieza para las simulaciones con nucl*

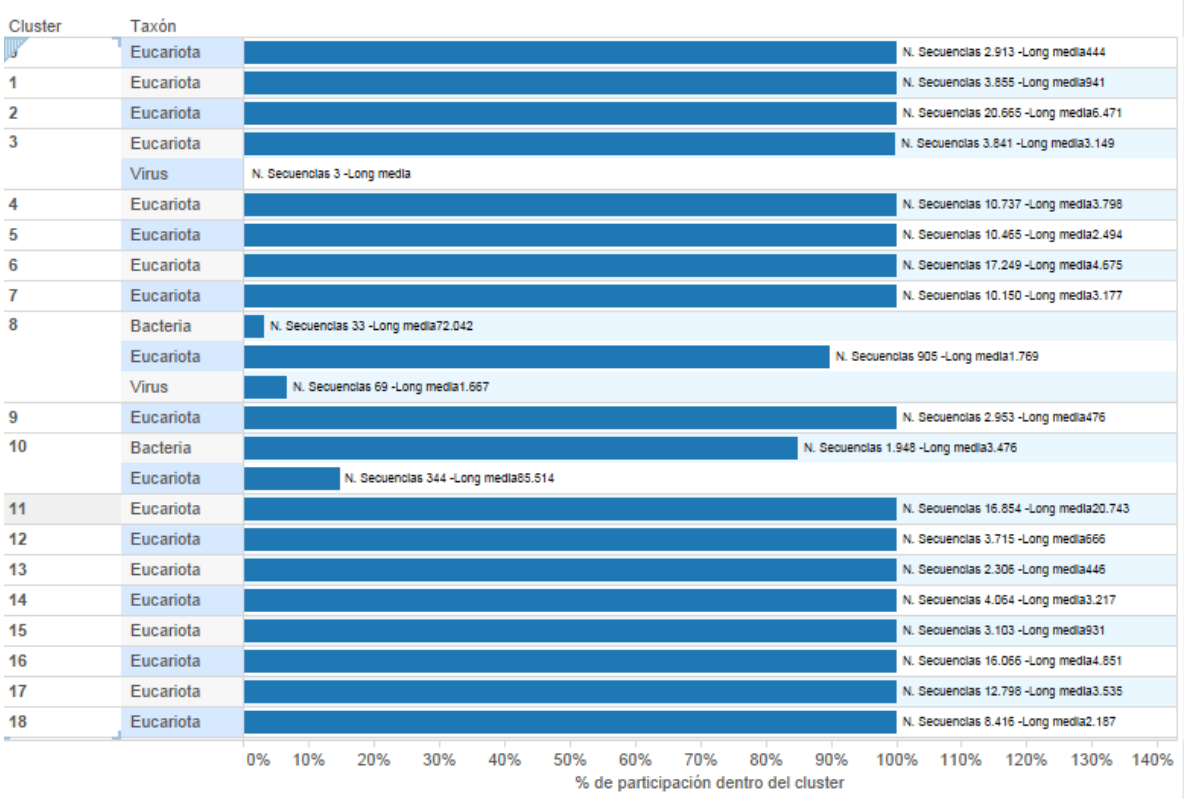


**Figura 7 Descripción de los grupos generados con el rasgo nucl, coseno y K=20**

El mejor indicador de limpieza para la función de distancia manhattan 75%, fue obtenido usando el rasgo de codones, para un valor de K=15, sin embargo dicha función no generó tan buenos resultados en las demás simulaciones, razón por la cual se considera que en futuras simulaciones, su uso podría no agregar mucho valor.

En términos generales el rasgo GC no presentó muy buenos resultados con el algoritmo K-means generando a lo sumo un clúster limpió por cada simulación en la que se utilizó, esto demuestra que por sí solo este rasgo no le da suficientes argumentos al modelo cómo para distinguir las secuencias a nivel general. Sin embargo al ser utilizado en conjunto con el rasgo cod-4mer generando el rasgo gc-cod-4mer, ayudó mejorar la limpieza de los grupos.

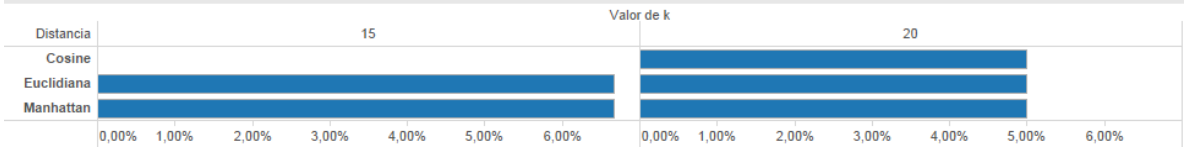


Figura 8 Resultados de pureza de los grupos para las simulaciones con gc

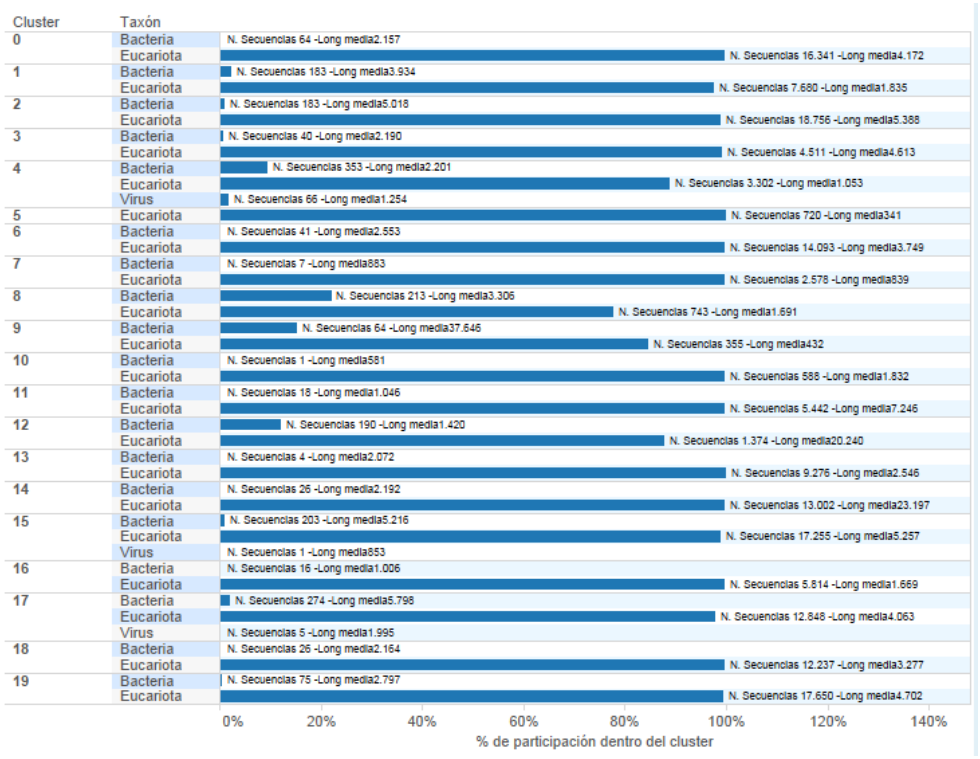


*Figura 9 Descripción de los grupos generados con el rasgo GC, distancia coseeno y k=20*

A pesar de los recursos y tiempo invertidos en las simulaciones con el algoritmo EM, no se consiguieron resultados más favorables que los arrojados por el método K-means, un hecho peculiar durante estas simulaciones fue que los grupos más limpios fueron generados utilizando el rasgo GC, que si bien fue uno de los rasgos que entragaron los peores resultados, el las simulaciones del EM entregó los resultados más estables Figura 9. Las simulaciones con el EM en un equipo de 6GB de memoria Ram y con un procesador Intel core i7 usando los archivos con los rasgos más sencillos (3 MB de peso), demoraban más de 6 horas, por tal motivo se decidió arrendar un servidor en Amazon AWS que contaba con un procesador intel de 3.5GHZ y 60MB de memoria RAM. En dicho servidor las simulaciones sencillas tardaban alrededor de una hora sin embargo

para los rasgos más complejos como por ejemplo el archivo gc-nucl-cod-4mer con un peso de 318MB este servidor no pudo realizar la simulación con dicho rasgo corriendo durante 7 horas. Comparado con las simulaciones realizadas con el K-means que en promedio demoraron 40 minutos cada una, utilizando el equipo de 6GB de memoria RAM.

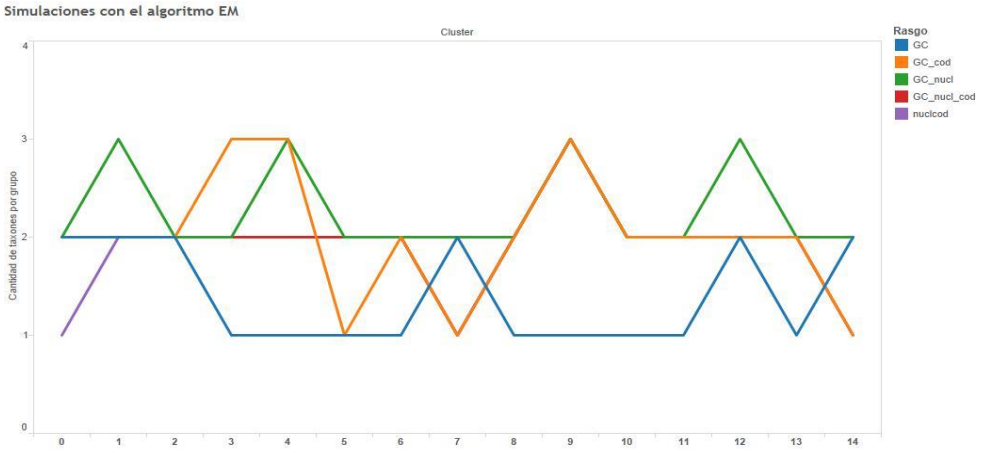


*Figura 10 Comparación entre las simulaciones realizadas con el EM*

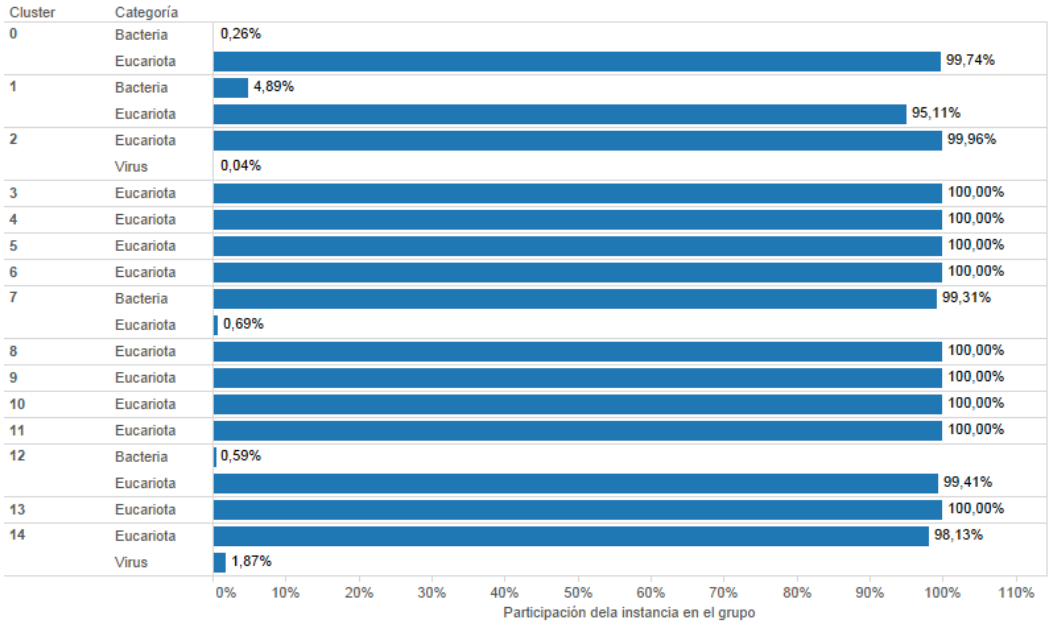


*Figura 11 Simulación del EM con el rasgo GC para 15 grupos*

En la siguiente figura se representa la duración promedio en minutos, para cada uno de los rasgos utilizados en las diferentes simulaciones del Kmeans, como era de esperarse la duración de las simulaciones es proporcional al número de fragmentos que se deben agrupar y la complejidad espacial del rasgo, y es aquí donde se destaca la sencillez del rasgo nucl al momento de generar resultados tan precisos.

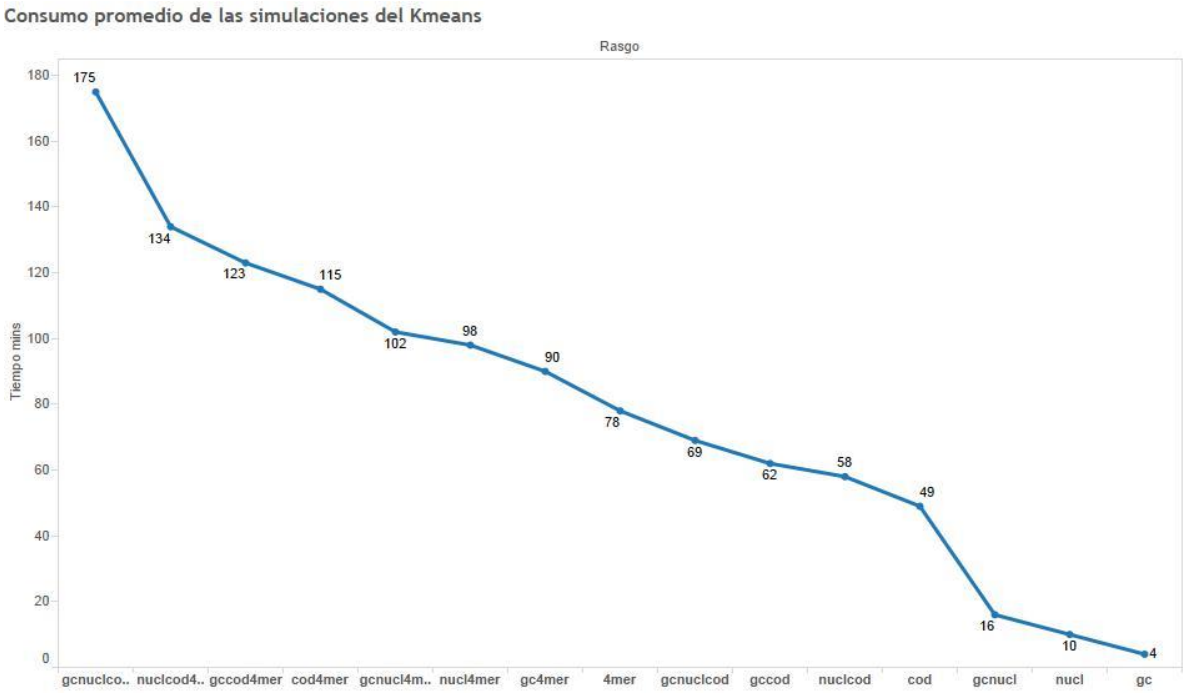


**Figura 12 Rendimiento promedio de las simulaciones del Kmeans**

## 4.2 INDENTIFICACIÓN DE TENDENCIAS Y ANÁLISIS ESPECÍFICOS

En otras investigaciones se ha hablado del reto que implican las lecturas cortas, para la identificación de las secuencias durante los estudios metagenómicos(Reddy et al., 2012; Wu & Ye, 2011). Sin embargo en los resultados de las simulaciones del algoritmo K-means con la distancia coseno y usando los rasgos compuestos por la frecuencia de nucleótidos (nucl), han mostrado igual capacidad para identificar los registros pequeños, además al nivel de los taxones más generales se ha pudo identificar durante estas simulaciones que los grupos generados con el menor número de instancias eran los grupos menos limpios.

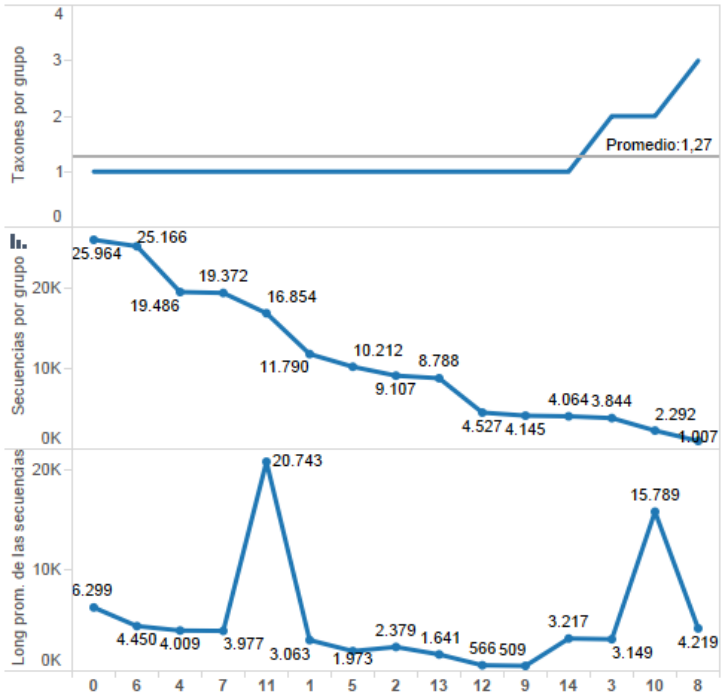


*Figura 13 Relación pureza, Tamaño de los grupos y longitud media coseeno, k=20, nucl*

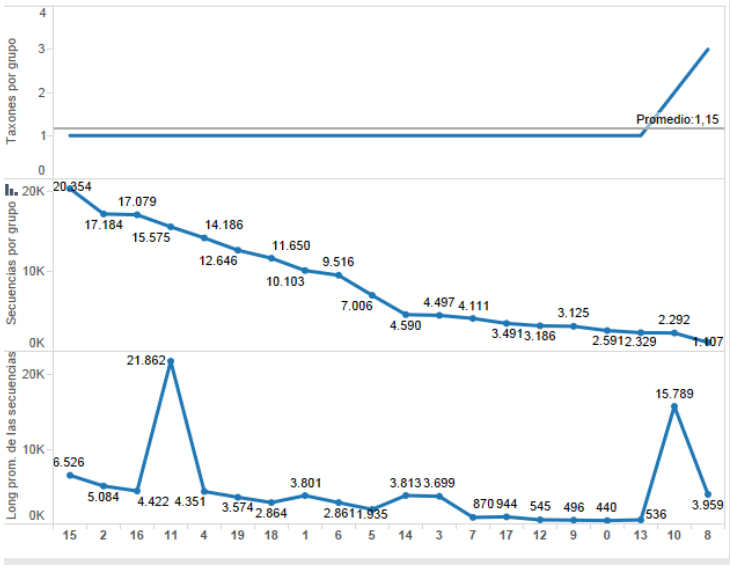


*Figura 14 Relación pureza, Tamaño de los grupos y longitud media coseno, k=20, nucl-cod-4mer*

Además de tener la división de las secuencias entre bacterias, virus y eucariotas, el conjunto de datos cuenta con un nivel de clasificación taxonómica más específico, una división por phylum, con un total de 9 taxones o grupos taxonómicos en este nivel el indicador de limpieza de los grupos es menor, dado a que es más difícil agrupar las secuencias a un mayor nivel de detalle, sin embargo el impacto del cambio de la granularidad no generó tanto impacto entre las simulaciones con los rasgos compuestos por nucl y la distancia Coseno presentan un comportamiento similar, disminuyendo su pureza al 75%, es decir de 20 grupos, 15 son totalmente limpios a nivel de phylum.

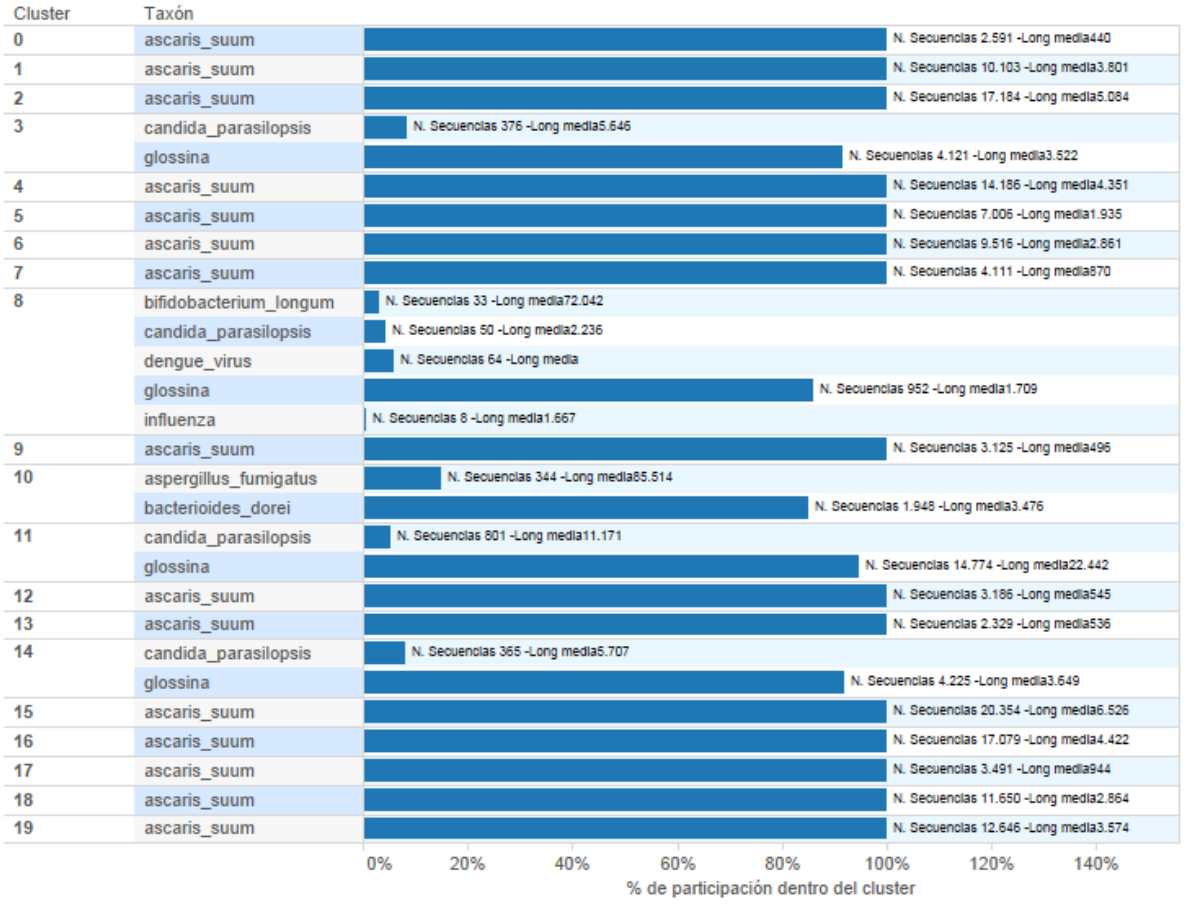


*Figura 15 Descripción de los grupos generados con el rasgo nucl-cod-4mer, coseno y K=20, a nivel de phylum.*

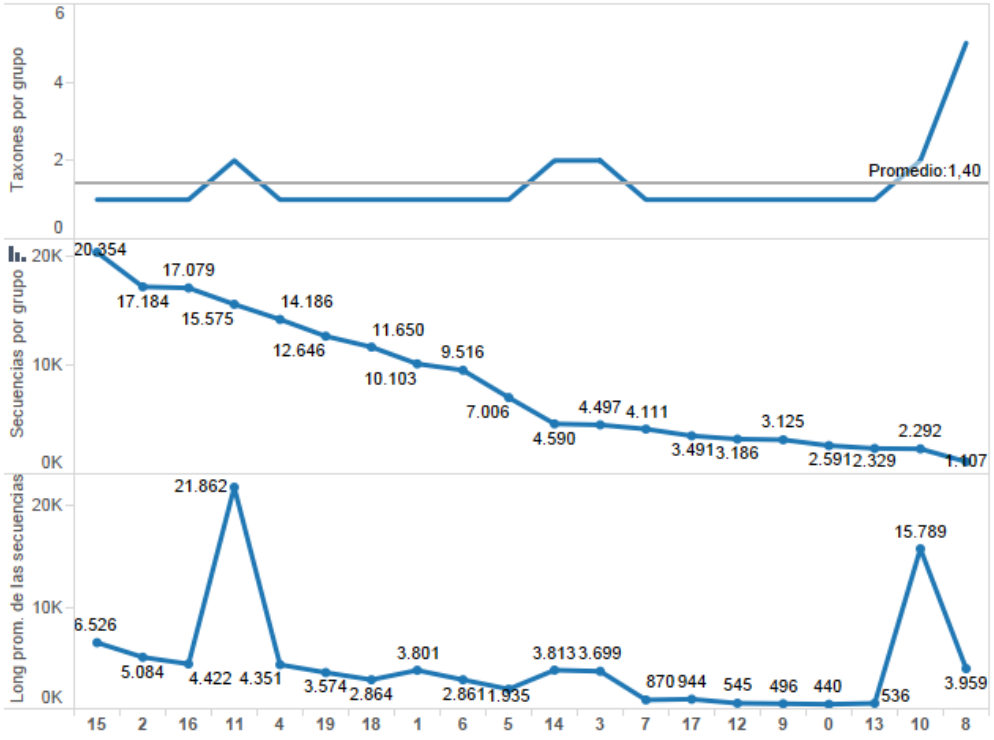


*Figura 16 Figura 13 Relación pureza, Tamaño de los grupos y longitud media coseno, k=20, nucl-cod-4mer a nivel de phylum, cambiando los taxones por grupo.*

Aparte de hacer los análisis de la limpieza por cada uno de los grupos, también fueron generados como parte de los resultados las distancias entre los centroides de los grupos, para encontrar alguna relación entre la pureza de los grupos y la distancia entre ellos, este indicador se representa cómo un mapa de color, en el cual el tamaño de la intersección de dos grupos representa la distancia entre ellos. Para las simulaciones con el mayor nivel de pureza, el análisis de distancia no muestra mucha información, sin embargo en la figura 16 si se observa la información a nivel de detalle de phylum, se puede observar que los grupos más fragmentados (3, 8,10,11 y 14) son los que tienen mayor distancia con respecto a los demás. Esta relación entre la limpieza de los grupos y su distancia a los demás abre las posibilidades a un paso adicional en la exploración.

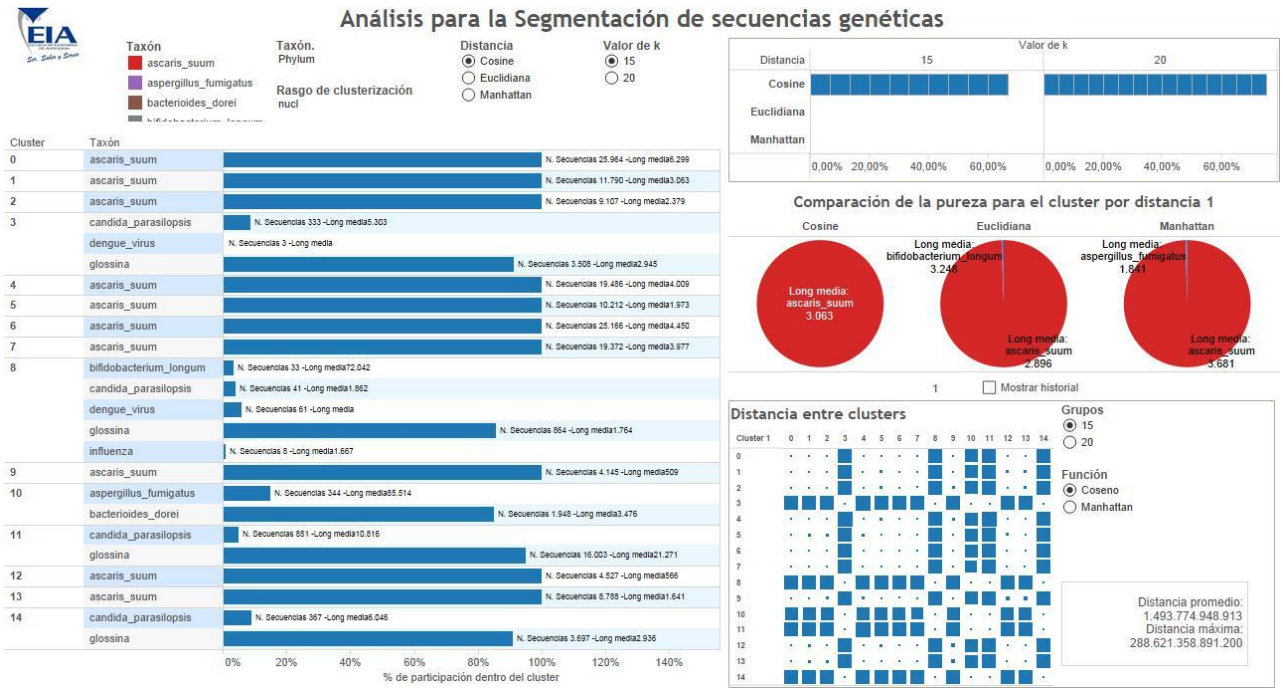


**Figura 17 Relación entre la pureza y la distancia entre grupos**

## 4.3 RESULTADOS K-MEANS ITERATIVO

La implementación del K-means Iterativo, una extensión de las capacidades del algoritmo K-means, para realizar una segunda segmentación con las secuencias pertenecientes a los K1 grupos que obtuvieron las mayor distancia promedio con respecto a los demás conjuntos durante la primera segmentación. Dado a que en la mayoría de las simulaciones con la distancia Coseno y los rasgos relacionados con la frecuencia de nucleótidos los grupos con la mayor distancia resultaron ser los grupos más fragmentados, al realizar una nueva segmentación sobre estos grupos se espera mejorar el indicador de pureza reemplazando los grupos fragmentados por un número mayor de conjuntos limpios.

La simulación del K-means iterativo se realizó con el rasgo nucl y la distancia Coseno para un valor de K=15, a partir de los resultados generados en la primera parte del algoritmo(parte izquierda de la figura 17) se calcularon los tres grupos con la mayor distancia promedio, el grupo 14, el 8 y el 11, en estos grupos hay 22 mil secuencias pertenecientes a 5 taxones diferentes, que luego de la segunda iteración del K-means quedaron divididas en 15 nuevos grupos), 13 de ellos totalmente limpios. Estos conjuntos se incorporaron a los 12 grupos restantes de la primera iteración, en otras palabras, los primeros 12 grupos de los resultados de la parte derecha de la fitura 17, son los mismos

resultados presentados por la simulación del K-means original a excepción de los grupos 8,11 y 14 de dicha iteración, que fueron utilizados para generar los 15 grupos adicionales presentados en los resultados, El resultado total de la simulación presenta 27 de los cuales 24 están totalmente limpios lo cual genera una pureza del 88%, mejorando ampliamente los resultados presentados por el K-means, incluso aumentando el nivel de detalle estudiando la división de las secuencias a nivel de phylum.

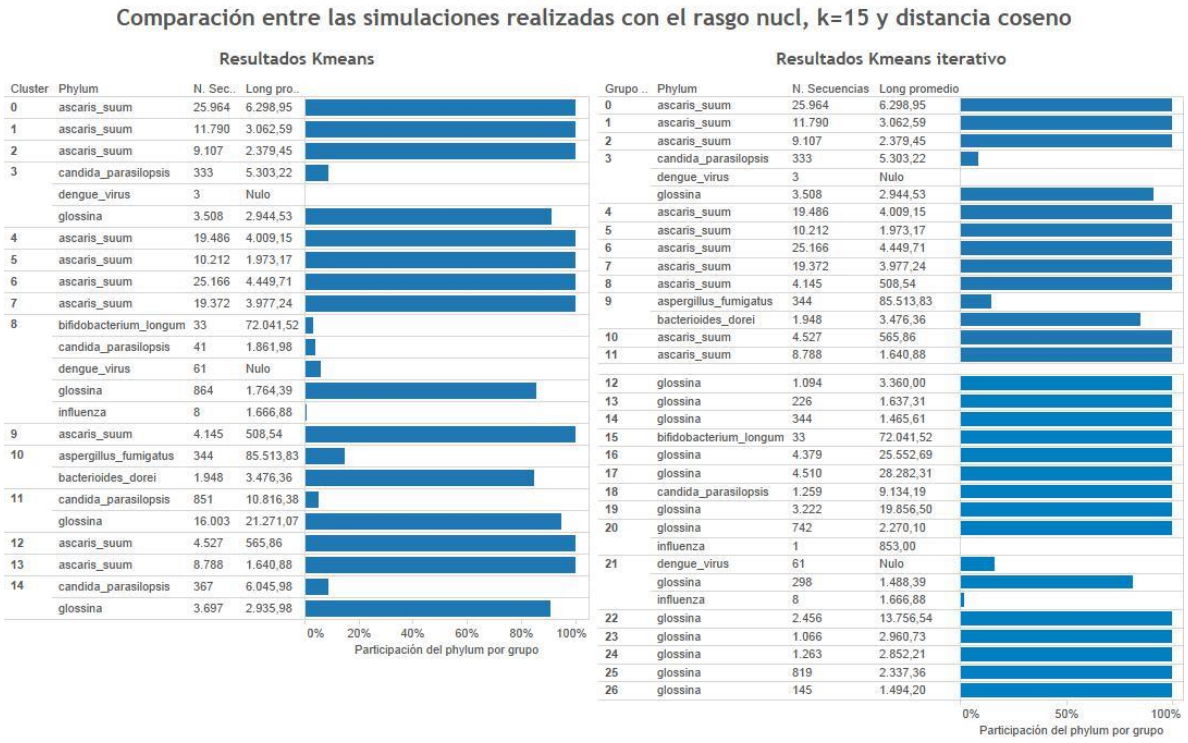


**Figura 18 Resultados del K-means Iterativo contra el K-means original**

# 5. CONCLUSIONES Y CONSIDERACIONES FINALES

Se ha comprobado la capacidad del K-means para identificar y segmentar las secuencias genéticas de una forma precisa y eficiente, si bien es cierto que al aumentar el nivel de detalle se pierde la pureza de los grupos, tal impacto puede ser mitigado utilizando el K-means iterativo, que trabaja específicamente con los grupos más separados de los demás.

Durante las simulaciones de los métodos de inteligencia artificial en el servidor de Amazon con 60GB de memoria RAM, se observó que el algoritmo solo consumía el 35% de la memoria y menos del 20% del procesador, en este orden de ideas una forma de optimizar en general este, y los demás procesos en los estudios metagenómicos, es programando en paralelo aquellas actividades en las que sea posible paralelizar, de forma que puedan ser usados todos los recursos de los equipos mientras que se optimizan los tiempos de respuesta.

A partir de las 95 simulaciones realizadas a las 166618 secuencias, se generaron en total 15828710 asignaciones, el generar y controlar esta cantidad de datos de forma óptima día a día durante los proyectos metagenómicos, se convierte en un problema de Big data, por tal motivo se recomienda para futuros trabajos, guardar y gestionar la información en bases de datos analíticas que tengan estructuras óptimas para el análisis de grandes cantidades de datos, en lugar de utilizar bases de datos transaccionales Además posterior a su procesamiento y almacenamiento, estos datos necesitan ser analizados de forma ágil y sencilla para convertirse en información.

Durante esta investigación se realizaron tableros de mando e informes dinámicos para realizar minería de datos sobre los resultados y encontrar tendencias y patrones en ellos, como por ejemplo la relación entre las distancias de los grupos y su nivel de limpieza. Esta investigación prueba que para generar valor en los proyectos metagenómicos es necesario realizar estrategias de minería y análisis de datos que permitan identificar tendencias y patrones sobre los resultados.

Como recomendación para trabajos futuros es relevante optimizar la variante del K-means iterativo, programando los métodos y las interfaces necesarias, que permitan parametrizar la selección de los grupos más alejados, no sólo por los que tengan la mayor distancia en promedio a los demás, sino también los que estén alejados ciertas desviaciones estándar, adicionalmente permitir realizar las simulaciones sobre los grupos más alejados con otros algoritmos, variando las parametrizaciones del mismo algoritmo a medida que avanzan las iteraciones, o que el algoritmo pueda establecer un punto de convergencia para las iteraciones extendidas.

# BIBLIOGRAFÍA


Ali, W., Shamsuddin, S. M., & Ismail, A. S. (2012). Intelligent Naïve Bayes-based approaches for Web proxy caching. *Knowledge-Based Systems*, *31*, 162–175. doi:10.1016/j.knosys.2012.02.015

Altschul, S. F. ., Madden, T. L. ., Schäffer, A. A. ., Zhang, J. ., Zhang, Z. ., Miller, W. ., & Lipman, D. J. . (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. doi:10.1093/nar/25.17.3389

Andrea, P., & Hern, C. (2004). Aplicaci ´ on de ´ arboles de decisi ´ on en modelos de riesgo crediticio Introducci ´ on Definiciones, 139–151.

Bekel, T., Henckel, K., Küster, H., Meyer, F., Mittard Runte, V., Neuweger, H., … Goesmann, A. (2009). The Sequence Analysis and Management System – SAMS-2.0: Data management and sequence analysis adapted to changing requirements from traditional sanger sequencing to ultrafast sequencing technologies. *Journal of Biotechnology*, *140*(1-2), 3–12. doi:10.1016/j.jbiotec.2009.01.006

Castellanos-Garzón, J. a., & Díaz, F. (2013). An evolutionary computational model applied to cluster analysis of DNA microarray data. *Expert Systems with Applications*, *40*(7), 2575–2591. doi:10.1016/j.eswa.2012.10.061

Fancello, L., Raoult, D., & Desnues, C. (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology*, *434*(2), 162–74. doi:10.1016/j.virol.2012.09.025

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. a., & Strachan, R. (2013). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*. doi:10.1016/j.eswa.2013.08.089

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology, 1518–1525. doi:10.1242/jeb.001370

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9818143

Javier, C., Germ, T., Una, R., & Neuronal, R. (n.d.). Redes Neuronales Artificiales Introducci ´ Modelo neuronal de McCulloch-Pitts, 22–30.

Keshavarz, M., & Huang, B. (2014). Expectation Maximization method for multivariate change point detection in presence of unknown and changing covariance. *Computers & Chemical Engineering*, *69*, 128–146. doi:10.1016/j.compchemeng.2014.06.016

Laviades, J. D. C. (1999). Los retos de la hipertensión arterial en el siglo XXI, *XIX*, 487–491.

Lebovka, N., Khrapatiy, S., & Pivovarova, N. (2014). Barrier properties of -mer packings. *Physica A: Statistical Mechanics and Its Applications*, *408*, 19–27. doi:10.1016/j.physa.2014.04.019

Li, X., & Waterman, M. S. (2003). Estimating the repeat structure and length of DNA sequences using L-tuples. *Genome Research*, *13*(8), 1916–22. doi:10.1101/gr.1251803

Logares, R., Haverkamp, T. H. a, Kumar, S., Lanzén, A., Nederbragt, A. J., Quince, C., & Kauserud, H. (2012). Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *Journal of Microbiological Methods*, *91*(1), 106–13. doi:10.1016/j.mimet.2012.07.017

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155. doi:10.1093/nar/gks678

Ngs, A. (n.d.). Estado del Arte NGS.

Papamichail, D., Skiena, S. S., Lelie, D. V. A. N. D. E. R., & Mccorkle, S. R. (2004). Bacterial population assay via k -mer analysis (extended abstract), 1–10.

Reddy, R. M., Mohammed, M. H., & Mande, S. S. (2012). TWARIT: an extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. *Gene*, *505*(2), 259–65. doi:10.1016/j.gene.2012.06.014

Reddy, R. M., Mohammed, M. H., & Mande, S. S. (2014). MetaCAA: A clustering-aided methodology for efficient assembly of metagenomic datasets. *Genomics*. doi:10.1016/j.ygeno.2014.02.007

Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of Advanced Research*, *4*(4), 331–344. doi:10.1016/j.jare.2012.05.007

Salas, R. (n.d.). Redes Neuronales Artificiales, 1–7.

Saleh-Lakha, S., Miller, M., Campbell, R. G., Schneider, K., Elahimanesh, P., Hart, M. M., & Trevors, J. T. (2005). Microbial gene expression in soil: methods, applications and challenges. *Journal of Microbiological Methods*, *63*(1), 1–19. doi:10.1016/j.mimet.2005.03.007

Seok, H.-S., Hong, W., & Kim, J. (2014). Estimating the composition of species in metagenomes by clustering of next-generation read sequences. *Methods (San Diego, Calif.)*. doi:10.1016/j.ymeth.2014.07.009

Shokrzadeh, H., Khorsandi, S., & Haghighat, A. T. (2012). Optimized query-driven appointment routing based on Expectation-Maximization in wireless sensor networks. *Journal of Network and Computer Applications*, *35*(6), 1749–1761. doi:10.1016/j.jnca.2012.06.007

Singh, B., Gautam, S. K., Verma, V., Kumar, M., & Singh, B. (2008). Metagenomics in animal gastrointestinal ecosystem: Potential biotechnological prospects. *Anaerobe*, *14*(3), 138–44. doi:10.1016/j.anaerobe.2008.03.002

Soh, J., Dong, X., Caffrey, S. M., Voordouw, G., & Sensen, C. W. (2013). Phoenix 2: a locally installable large-scale 16S rRNA gene sequence analysis pipeline with Web interface. *Journal of Biotechnology*, *167*(4), 393–403. doi:10.1016/j.jbiotec.2013.07.004

Susana, P., & Raimondi, G. De. (n.d.). Metagenómica: más allá del genoma de los microorganismos, 1–6. Retrieved from http://revistas.unc.edu.ar/index.php/Bitacora/article/view/5573

Swain, M. T. (2013). Fast Comparison of Microbial Genomes Using the Chaos Games Representation for Metagenomic Applications. *Procedia Computer Science*, *18*, 1372–1381. doi:10.1016/j.procs.2013.05.304

Wen, J., Chan, R. H. F., Yau, S.-C., He, R. L., & Yau, S. S. T. (2014). K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene*, *546*(1), 25–34. doi:10.1016/j.gene.2014.05.043

Wu, Y.-W., & Ye, Y. (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *18*(3), 523–34. doi:10.1089/cmb.2010.0245

Yousuf, B., Keshri, J., Mishra, A., & Jha, B. (2012). Application of targeted metagenomics to explore abundance and diversity of $CO_2$-fixing bacterial community using cbbL gene from the rhizosphere of Arachis hypogaea. *Gene*, *506*(1), 18–24. doi:10.1016/j.gene.2012.06.083

Zakrzewski, M., Bekel, T., Ander, C., Pühler, A., Rupp, O., Stoye, J., … Goesmann, A. (2013). MetaSAMS--a novel software platform for taxonomic classification, functional annotation and comparative analysis of metagenome datasets. *Journal of Biotechnology*, *167*(2), 156–65. doi:10.1016/j.jbiotec.2012.09.013

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, *41*(4), 1476–1482. doi:10.1016/j.eswa.2013.08.044

Zhou, Q., Su, X., Jing, G., & Ning, K. (2014). Meta-QC-Chain: comprehensive and fast quality control method for metagenomic data. *Genomics, Proteomics & Bioinformatics*, *12*(1), 52–6. doi:10.1016/j.gpb.2014.01.002

# ANEXO 1

| Referencia | Algoritmo | Rasgos | Base de datos | Resultado | | Equipos utilizados |
|---|---|---|---|---|---|---|
| | | | | Precisión en la clasificación y otros indicadores | Tiempo y eficiencia | |
| Proyecto abundance bin (Wu & Ye, 2011) | Primero establece que la aparición de un nucleótido en una secuencia dada, sigue una distribución de Poisson, dada esta hipótesis, se utiliza el algoritmo de E.M para determinar los parámetros de la distribución que mejor se acomodan a la muestra dada, para luego realizar la clasificación | El rasgo a utilizar en este experimento eran las frecuencias de nucleótidos conocidas como K-mer o L-tuplas, este rasgo consiste en la frecuencia con la que aparece una combinación de *l* nucleótidos en la cadena dada | Para generar un conjunto de datos artificial, utilizaron un Programa llamado MetaSim, que toma como datos de entrada, un conjunto de genomas conocido, y genera una muestra de datos "Metagenómica" a partir de estos genomas que para este caso fueron tomados del NCBI(National Center For Biotecnology Information ) | Por ejemplo para una relación de abundancia entre las especies menor a 1 ,5:1 se obtuvo un error del 20,6%, y el error para definir el número de especies, aumentaba al 11% para 6 genomas. | El rendimiento de la aplicación, depende de la longitud de los rasgos(*l*-tuplas), la cual está dada por l, sin embargo para el experimento, se comprobó que el mejor rendimiento se lograba con l igual a 20, también se recomienda excluir los rasgos que solo aparezcan 1 vez | (using single CPU on Intel(R) Xeon(R)@2. 00GHz) |
| BACTERIAL POPULATION ASSAY VIA K-MER ANALYSIS (Papamichail et al., 2004) | El sistema clásico basado en el vector de frecuencias Kmer para la clasification (Sandberg et al.) Utiliza el algoritmo de clasificación Naïve Bayes | Ambos algoritmos usan el mismo tipo de rasgos, calculan el vector de frecuencias K-mer para las secuencias en la base de datos, sin embargo en esta investigación también realizan un trabajo previo sobre las secuencias, y es dividirlas en Genomic Sequence Tags (GSTs) son lecturas cortas (21 bp) capaces de proporcionar información suficiente de manera independiente, luego sobre estas lecturas cortas se calcula el vector K-mer | A partir de 25 genomas, formó secuencias aleatorias para su trabajo, comparándolo con el método Naïve. Con diferentes valores para K | Tanto el método Bayesiano clásico como el método basado en probabilidades condicionales fueron comparados evaluando muestras de diferentes longitudes(35bp , 60bp , 100bp , 200bp , 400bp y 100bp), en general con respecto a este criterio de evaluación podemos concluir que a mayor longitud delas secuencias la clasificación es más precisa, sin embargo esto requiere mayor tiempo de procesamiento, también al evaluar los resultados se utilizaron dos rasgos de evaluación para los dos métodos en todas las medidas evaluadas ambos rasgos están basados en la codificación Kmer y su única diferencia fue el valor de K que para este ejercicio fue evaluado en 3 y 8, los resultados mostraron que utilizando la frecuencia 8Mer logró mejores resultados que al evaluar K=3, sin embargo este tiende a estabilizarse a medida que aumenta la longitud de las muestras, finalmente el método clásico logró una máxima precisión del 85% con K = 8, el segundo método basado en las probabilidades condicionales mejora la precisión del sistema a más del 90% utilizando el mismo valor de K y secuencias de las misma longitud 100bp. | | No presenta |
| | Mejora el algoritmo anterior, utilizando probabilidades condicionales, es decir dado a que se tiene un segmento se nucleótidos, calculan la probabilidad de que aparezca el K-mer N, dado que apareció el K-mer N-1, luego la probabilidad más alta ayuda a secuenciar e identificar las diferentes cadenas | | | | | No presenta |

| | | | | | | |
|---|---|---|---|---|---|---|
| TWARIT: An extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences (Reddy et al., 2012) | En general el algoritmo evalúa inicialmente la longitud de la secuencia, si esta es menor a 150bp, utiliza el BWA TOOL | Inicialmente el criterio de evaluación utilizado por el sistema es la longitud de las lecturas | NCBI contiene la base de datos de los genomas de 952 procariotas(ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz/) luego, fueron indexadas usando el algoritmo 'bwtsw' incluido en la herramienta BWA, para evaluar la precisión de TWARIT comparado con otras herramientas, fueron utilizadas 4 bases de datos simuladas nombradas como Sanger, 454–400, 454–250 and 454–100 los conjuntos de datos contienen 35,000 lecturas sus longitudes son Sanger (Longitud por lectura alrededor de 800 bp), 454-Titanium (~400 bp), 454-Standard (~250 bp), y 454-GS20 (~100 bp) llamadas así por las tecnologías utilizadas para ser y simuladas a partir del conjunto de datos del NCBI. | La comparación entre las 5 herramientas, muestra nuevamente que la relación entre la precisión de los sistemas, y la longitud de la lecturas debido a que el mayor índice de clasificación por la herramienta SOrt-ITEMS para la base de datos, sin embargo TWARIT mostró una mayor tolerancia a la reducción del tamaño de las lecturas, superando al SOrt-ITEMS que muestra tener de las peores cualidades entre las 5 herramientas para las base de datos con las lecturas más pequeñas ( 250 bp y 100bp), por otro lado la herramienta MEGAN tuvo el mayor índice de clasificaciones erróneas para las bases de datos con las lecturas de mayor longitud, sin embargo este porcentaje también bajo al momento de clasificar lecturas pequeñas, sin embargo también perdió precisión para clasificar correctamente las lecturas. | Los sistemas también fueron comparados en cuanto a rendimiento, y aquí es en donde la velocidad de TWARIT sale a relucir; los 5 sistemas fueron comparados evaluando el tiempo que toman procesar 10.000 lecturas de las 4 bases de datos en términos generales las bases de datos con las lecturas más largas tomaron más tiempo para ser evaluadas, los sistemas MEGAN y SOrt-ITEMS fueron las herramientas más lentas al tomar alrededor de 5 horas en promedio de las cuatro bases de datos, para procesar las 10.000 lecturas, por otro lado las herramientas más rápidas fueron SPHINX y TWARIT que en promedio tomaron 15 y 3.5 minutos en procesar la misma cantidad de información respectivamente. | Intel Xeon quad core processor (4 GB RAM) |
| | Si la longitud supera los 150 bp realiza un HPBA un algoritmo de 'hit-pair based assignment' (HPBA) para hacer comparaciones con los dos extremos de la secuencia, en lugar de toda la secuencia, básicamente el alineamiento se logra cuando los extremos y la distancia que los separa coincide | Los rasgos analizados son los extremos de la cadena, y la longitud intermedia entre los extremos, para hacer el proceso de alineación directa | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Cuando el BWA o el HPBA no generan los resultados esperados, se recurren al 'signature sorting based assignment' (SSBA) un algoritmo que utiliza K-means(Manhatan 256D frecuencia de nucleótidos) para agrupar todas las secuencias parecidas luego con la distancia de cada una de las secuencias a cada uno de los centroides se genera una firma de referencia(concatenándose) y la base de datos se ordena con respecto a esa firma, dejando los elementos vecinos con mayor parecido, para hacer posteriormente la comparación | En esta parte se utiliza la frecuencia de nucleótidos de cada frecuencia, un vector de 256 dimensiones formado con todas las posibles combinaciones de los cuatro nucleótidos, y el número de veces que una de estas combinaciones se repite en la secuencia. Posteriormente se a partir de un proceso de K-means se generan los centroides de referencia, y estos a su vez sirven para sacar la firma de la secuencia, concatenando la distancia de cada una de las secuencias a los centroides de referencia | | | Finalmente vale la pena resaltar que en cuestiones de eficiencia TWARIT no depende de la longitud de las lecturas, ya que para la base de datos de SANGER(con lecturas de 800 bp) demoró lo mismo que al procesar las lecturas de 100bp. | |
| Meta-QC-Chain: Comprehensive and Fast Quality Control Method for Metagenomic Data (Zhou et al., 2014) | Es una herramienta que sirve para evaluar la calidad de las muestras metagenómicas, utilizando diferentes métodos de pre filtrado, como longitud de las lecturas, algunos marcadores biológicos, recortes de las lecturas, entre otros… una de las primeras evaluaciones más relevantes que efectúa esta herramienta, es mirar el contenido de GC la aplicación está disponible para el consumo público en la siguiente dirección : http://computationalbioenergy.org/meta-qc-chain.html. | En este paso del flujo de trabajo se tratan de identificar las secuencias GC, que representan las islas de Guanina unidas por medio de Fosfatos con citosina, que suelen repetirse mucho en las secuencias | En ambos procedimientos utilizaron dos muestras reales de saliva humana, los datos se pueden extraer de la siguiente dirección (http://computationalbioenergy.org/meta-qc-chain.htm) | Cómo el objetivo del proyecto no era realizar una asignación metagenómica, no se presentan resultados de precisión, sino del tiempo que tomaron las diferentes operaciones | En cuanto al tiempo requerido para realizar cada una de las tareas en las diferentes bases de datos, los resultados mostraron que para las pruebas técnicas demoraron entre 1 y 2.30 minutos, la limpieza de la contaminación demoró entre 2 y 10 minutos, sin embargo el tiempo requerido para realizar el ajuste de calidad de las lecturas está un poco más relacionado con el tamaño de la base de datos ya que demoró para la base de datos R1 8min y 33s para la base R2 14 min y para la base S1 4 min, finalmente los tiempos totales fueron aproximadamente de 11, 19 y 17 minutos para las tres bases de datos respectivamente. | Rack server with an Intel dual Xeon E5-2650 CPU (2.0 GHz, 16 cores in total, supporting 32 threads), 64GB DDR3 ECC RAM and 2TB HDD. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | El otro procedimiento importante dentro de su flujo de trabajo, es identificar las posibles especies eucariotas, ya que según los autores, generan contaminación en las muestras metagenómicas | Aquí utilizan el marcador 18S rRNA, como un marcador biológico de las células eucariotas(y que podría no otorgar mucha información), posteriormente, hacen una alineación con una base de datos para identificar los elementos que contaminan para extraerlos | Para la evaluación el Proyecto fueron utilizadas tres bases de datos llamadas R1, R2 y S1 con 19185960 , 33134512 y 22127714 lecturas y un peso de 5 , 8.3 y 2.2 GB respectivamente, aunque después de la limpieza de los datos, las bases quedaron con un total de longitudes de 9414926 , 20951704 y 22127714 y un peso final de 1.2 , 2.8 y 2.2 GB respectivamente. | | | |
| MetaCAA: A clustering-aided methodology for efficient assembly of metagenomic datasets (Reddy et al., 2014) | Su procedimiento se divide en tres pasos, el primero es clusterizar los segmentos de la secuencia metagenómica utilizando el método "Cosine Angle" utilizando como criterio de distancia la función Coseno, luego todos los elementos de cada clúster, los ensambla con la ayuda de una herramienta de ensamble de secuencias de tipo "Greedy" llamada CAP3 | En el primer paso utilizan como rasgo la frecuencia de tetra nucleótidos, usada también en la referencia del TWARIT, y en la segunda parte, la aplicación CAP3 ensambla las secuencias que tengan segmentos de mayor longitud en común | Utilizaron 12 base de datos, nueve elaboradas artificialmente y 3 bases de datos reales, las bases de datos las subdividen en grupos de complejidad, que representan espacios con mayor o menor diversidad de especies | Los factores de comparación entre estos dos ensambladores MetaCAA y CAP3, fueron divididos de acuerdo a la base de datos y a la complejidad del medio, dependiendo del número de especies dominantes(En términos de porcentaje de participación en la muestra), así el medio con baja complejidad tendría pocas especies dominantes, luego los criterios para comparar fueron el número de bases ensambladas en cóntigos en donde los resultados fueron más favorables para los medios más complejos, y MetaCAA superó por más del 34% al otro ensamblador para la base de datos Sanger y por un valor muy parecido en la base 454-T, Finalmente en términos generales se puede concluir que MetaCAA tuvo un mejor rendimiento que CAP3 tanto en el número de secuencias ensambladas, como en la mayor pureza obtenida, también que se concluyó que la pureza disminuye a medida que aumenta la complejidad de la muestra. | Para medir la eficiencia de MetaCAA se utilizaron 7 conjuntos de datos, variando el número de secuencias, desde el primero que tiene 100 secuencias, hasta el último con 25.000, en donde sus tiempos de ejecución variaron desde 0.46 segundos hasta 37.98 respectivamente. | 2.33 GHz desktop with 2 GB of RAM |

| | | | | | | |
|---|---|---|---|---|---|---|
| Fast comparison of microbial genomes using the Chaos Games Representation for metagenomic applications (Swain, 2013) | Consiste en graficar los nucleótidos de una secuencia de forma iterativa con una función acumulativa(en la que el resultado producido por el nucleótido n, depende del resultado de f(n-1)) en un espacio bidimensional, delimitado en un plano unitario, luego se divide la matriz unitaria en r regiones permitiendo calcular la densidad de puntos por región y finalmente dependiendo de la densidad de las regiones, se hace una comparación de tipo alignment contra una base de datos conocida, según la investigación es mucho más óptimo comparar densidades de las matrices que las cadenas de nucleótidos. | Luego de tener graficados los diferentes nucleótidos, el espacio se divide en r cuadrados, y la densidad por cada uno de los paneles sirve como rasgo de comparación | El software Genometa(C. Davenport, J. Neugebauer, N. Beckmann, B. Friedrich, B. Kameri, S. Kokott, M. Paetow, B. Siekmann, M. Wieding-Drewes, M. Wienhofer,S. Wolf, B. Tommler, V. Ahlers, F. Sprengel, Genometa - a fast and accurate classifier for short metagenomic shotgun reads., Plos One 7 (5) (2012) e41224.), tiene una base de datos de 2550 genomas, de los cuales tienen 1551 genomas de especies diferentes, luego este último conjunto de datos lo utilizaron como la base de datos de destino, y los genomas restantes fueron utilizados para consulta | En el experimento se utilizaron 977 genomas de los datos de consulta, y los compararon contra los 1550 genomas de cada especie, haciendo un proceso de identificación, los resultados de este método se compararon tanto en tiempo y eficiencia contra los resultados presentados por la herramienta BLAST, luego sobre estos datos GCR presentó una mayor precisión cuando utilizaba las longitudes de 1/16 y 1/32 en donde las lecturas presentaban una mayor longitud, para tales casos estamos hablando de una precisión alrededor del 67%. | En comparación con la herramienta BLAST la eficiencia mostrada por CGR es un factor de decisión contundente, si bien la relación entre los tiempos en minutos para la secuencias pequeñas(100bp) es de 1 a 7, a medida que la longitud de las secuencias aumenta, esta diferencia se hace mucho más pronunciada, hasta el punto de decir que por cada minuto de procesamiento de la herramienta GCR el BLAST demora 100, finalmente la precisión de BLAST varía desde un 68% hasta un 83% mientras que aumentan la longitud de las secuencias, desde 100 bp hasta 100Kbp | 2.93 Ghz Intel Xeon chip |
| Phoenix 2 (Soh et al., 2013) | Básicamente este algoritmo en su flujo de trabajo incorpora varios tipos de algoritmos, primero hace una limpieza de los datos, eliminando lecturas repetidas, o que generan ruido, luego hace un breve alineamiento, posteriormente con las secuencias alineadas se realiza una clusterización iterativa definiendo las diferentes OTUs(Unidades taxonómicas), para finalmente hacer la asignación taxonómica final, con las secuencias representativas de cada OTU | Para realizar la agrupación en OTUs, se define la matriz de distancia entre las secuencias, siguiendo el concepto de identidad relativa entre las secuencias. | Para este proyecto utilizaron una base de datos enviada al servidor en abril 30 del 2012, que tenía 47 muestras de entornos de hidrocarbono, tomados de "Athabasca oil sands", un campo de petróleo convencional | Para el análisis de las 47 muestras el número de lecturas que superaron la etapa del control de la calidad fueron 245865 el (61%) de la muestra, luego en el paso de des-replicación quedaron 18215 el 7.4% de la muestra total estas representan el número total de lecturas agrupadas en esta etapa y finalmente después de la alineación quedaron 5651 luego con estas lecturas se hizo la comparación entre Phoenix2 y Mothur(una herramienta similar), finalmente los resultados mostraron un mayor índice de agrupación por parte de Phoenix 2, sin mencionar que la comparación entre los tiempos de ejecución de las etapas de clusterización fueron de 2 minutos que tomo Phoenix2 en realizar esta tarea, contra 142 minutos que demoró Mothur. | Tomó aproximadamente 34,7 horas para realizar el estudio de las 47 muestras que ocupaban un espacio en disco de 15Gb, también a partir de un análisis realizado entre los tiempos y los números de lecturas en los trabajos realizados en esta herramienta, llegaron a la conclusión de que el tiempo en horas podía estar dado por la siguiente Función teniendo como R el número de lecturas $t = 0{,}0001 * R + 0{,}7763$ | SunFire 6800 with 20 of 1250 MHz UltraSPARC IV processors and four 1.5 GHz UltraSPARC IV+ processors with 96GB RAM. |

| | | | | | | |
|---|---|---|---|---|---|---|
| GAHierarchical clustering (Castellanos-Garzón & Díaz, 2013) | Inicialmente se parte de la idea de agrupar las secuencias genéticas en clúster jerárquicos(Similar a una clasificación taxonómica), sin embargo el algoritmo para agrupar estas secuencias en una jerarquía(Dendograma), tiende a caer en soluciones locales, una forma de solucionar este problema, según los autores de este arartículoes a través de algoritmos genéticos los cuales a partir de diferentes Dendograma, tratan de llegar a una solución global del problema. | La agrupaciones jerárquicas se realizar a partir de las matrices de similitud entre las diferentes secuencias. Luego cada una de estas clasificaciones, sirve como cromosoma para el algoritmo genético, y como función de evaluación utilizan el nivel de concentración al interior del Dendograma(Una combinación entre las distancias entre los elementos al interior del clúster, y las distancias entre diferentes grupos) | Fueron utilizadas dos bases de datos diferentes, la primera una base compuesta por 384 genes la cual está publicada en esta dirección http://faculty.washington.edu/kayee/cluster la segunda base de datos llamada Sorlie está compuesta de 456 genes tomados de 85 muestras. | Uno de los resultados más destacables, es en el que se compara este método de clusterización, junto con otros 5 métodos, en términos de la calidad de los clústeres para la primera base de datos presentada, la calidad de los clústeres fue medida en términos de la homogeneidad al interior de los grupos y la separación entre estos, los métodos con mejores resultados fueron subrayados, señalando el método estudiado en este artículo como el mejor en 2 de los 3 aspectos a evaluar ya que la los clúster más homogéneos fueron obtenidos por la herramienta "Hybrid clust" | | 3 GHz computer of 2 GB of memory, using Debian GNU/Linux as an operating system |
| K-mer natural vector and its application to the phylogenetic analysis of genetic sequences (Wen et al., 2014) | Es una mejora que se le hace a las operaciones realizadas con los vectores K-mer, ya que al condensar las secuencias, en este vector, según los autores se pierde información evolutiva, por tanto se reemplaza el K-mer convencional, con el K-mer natural un nuevo vector que garantiza correspondencia 1 a 1, lo cual implica que cualquier secuencia puede ser identificada inequívocamente con su vector K-mer Natural | La mejora en esta metodología radica principalmente en la composición del rasgo para realizar las operaciones de clusterización, el vector natural se compone a partir de concatenar los 3 siguientes vectores, el vector K-mer común v1=(n1, n2, nl), donde n es el número de veces que se repite el segmento de longitud k, en la secuencia de nucleótidos de longitud l, el segundo vector V2= (u1, u2, ul) donde cada u(i) es el promedio aritmético desde la ocurrencia de cada K-mer a la primera base de la secuencia, el último vector v3=(D1, D2, ...Dl) es el vector de momentos centrales normalizado, finalmente este rasgo permite definir a cada una de las lecturas, y la distancia entre ellas. | Para esta investigación se usaron dos bases de datos diferentes, la primera consta de 53 genomas humanos, y la segunda de 40 secuencias de diferentes especies | Los resultados de esta investigación, para las dos bases de datos, fueron dos árboles filogenéticos, los cuales según sus autores, estaban muy relacionados con las estructuras reales, también en las conclusiones, se resalta a esta modificación a los vectores de frecuencias, como una poderosa herramienta, por su capacidad de identificar cada secuencia de forma única, y por llevar en su estructura relaciones evolutivas. | No presentan resultados de tiempo, sin embargo en las conclusiones mencionan que estos procedimientos son más óptimos, ya que las secuencias solo tienen que ser leídas completamente al momento de su transformación en los vectores de frecuencia, y que el proceso de identificación se hace sobre las estructuras reducidas. | No presenta |

| | | | | | | |
|---|---|---|---|---|---|---|
| MetaVelvet : An extension of Velvet assembler to de novo metagenome assembly from short sequence reads (Namiki et al., 2012) | Este proyecto extiende las capacidades del ensamblador Velvet, para secuencias cortas, básicamente parten del principio que el grafo de Bruijn al ser construido a partir de una secuencia metagenómica de diferentes especies, en esencia este también puede estar construido de pequeños grafos, también argumentan que dos especies evolutivamente separadas no comparten ningún *K-mer, luego los grafos de Bruijn que estas generarían serían dos grafos independientes. A partir de una secuencia metagenómica, realizan el grafo de Bruijn, luego crean un histograma de frecuencias de los diferentes nodos del grafo, los picos generados en el histograma representan el cubrimiento que tiene cierta especie en la muestra, y por tanto con los nodos pertenecientes a ese pico se calcula la probabilidad de cubrimiento aproximando los picos a distribuciones gaussianas, a partir de estas probabilidades son generados los sub grafos desde los nodos que tienen varias entradas y son considerados como quimeras(Proceden de varias especies), finalmente sobre estos subgrafos se pueden generar las secuencias y sus clasificaciones taxonómicas* | Para este trabajo, sobre cada secuencia se toman los diferentes Kmer, 2 k-mer consecutivos se solapan en K-1 nucleótidos estos nucleótidos se convierten en los nodos del grafo de Bruijn y los k-mer unen esos nodos, teniendo un grafo de Bruijn el objetivo es encontrar el camino euleriano sin embargo, para este proyecto los caminos se hallan sobre los sub grafos, y estos se extraen de los picos en el histograma generado por el peso de los nodos | Las simulaciones fueron realizadas en bases de datos simuladas con el software MetaSim, para generar secuencias cortas, ya que una de las bondades de este proyecto sería su aplicabilidad para este tipo de secuencias, y también probaron la aplicación con una base de datos real de la flora intestinal humana | El ensamblador MetaVelved fue comparado con otras 4 herramientas similares, Separate assembly, Velvet, ABySS, y SOAPdenovo las simulaciones en las bases de datos artificiales se tomaron basados en 4 especies diferentes, y estas fueron mezcladas para diferentes simulaciones, un hecho para resaltar es que a medida que se aumenta el número de especies también aumenta el número de secuencias ensambladas, también aumenta la longitud máxima para bases de datos con pocas especies ABySS presentó los mejores resultados, sin embargo estos resultados no mejoraron tanto como lo hicieron los de Velvet y MetaVelvet, que lograron los incrementos más pronunciados, presentando Velvet los mejores resultados en la mayoría de los casos; Sin embargo en las dos últimas simulaciones realizadas con 3 y 6 especies de pseudomonas ABySS logró ensamblar casi el doble de secuencias que velvet y MetaVelvet en resultados de 7617 contra 3412 y 3758 respectivamente. | En cuanto al rendimiento de Meta Velvet, para la base de datos que solo tenía 2 especies, demoró 35 minutos aproximadamente para lograr generar 76 secuencias ensambladas(En esta simulación no obtuvo los mejores resultados), y alrededor de 40 minutos para generar 444 secuencias ensambladas, y finalmente para la simulación de pseudomonas con 6 especies Metavelvet demoró casi 10 horas, cinco veces más que el ensamblador Velvet, logrando casi los mismos resultados a excepción de la pureza máxima de la secuencias. | No presenta |

## Anexo 1 Comparación entre los proyectos investigados