INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA

MAESTRÍA EN SISTEMAS COMPUTACIONALES

"*EXTRACCIÓN DE SECUENCIAS FRECUENTES MAXIMALES, MEDIANTE EL ALGORITMO DIMASP, PARA ANALIZAR CADENAS DE ADN*"

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRA EN SISTEMAS COMPUTACIONALES
P R E S E N T A

MA. DEL REFUGIO VELAZCO HERMOSILLO

DIRECTOR: DR. ALEJANDRO DEL REY TORRES RODRÍGUEZ
CODIRECTOR: DR. LUIS ALBERTO MORALES ROSALES

FEBRERO 2018

AGRADECIMIENTO

“Siempre hay que encontrar el tiempo

para agradecer a las personas

que hacen una diferencia en nuestras vidas”.

— John F. Kennedy

Gracias a Este Ser Maravilloso que es Dios, por darme el ser y la libertad para ser quien soy.

Gracias al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo para lograr la meta académica más importante en mi vida.

Gracias, Dr. Alejandro del Rey Torres Rodríguez, por aceptar ser el Director de esta Tesis, por su disponibilidad y apoyo. Gracias por compartir su conocimiento de manera generosa.

Gracias infinitas, Dr. Luis Alberto Morales Rosales, por tener la disponibilidad de apoyarme, en todo momento de manera generosa y desinteresada. Gracias por su paciencia, tolerancia y comprensión. Gracias por su humildad y humanidad. Sobre todo gracias por confiar en mí y motivarme para sacar lo mejor de mí.

Gracias a mi hermana Sara Alicia, a mis sobrinas: Mónica y Saraly, por su apoyo y por formar parte de mi vida.

Gracias a los compañeros que me brindaron su apoyo y su amistad.

Gracias a la Dra. Guadalupe Corelly Salazar Salazar, por el tiempo que me dedicó.

DEDICATORIA

“Un poco de ciencia aleja de Dios,

pero mucha ciencia devuelve a Él”

—Louis Pasteur.

Quiero dedicar esta Tesis, al Dr. Luis Alberto Morales Rosales, a quien considero una persona con vastas cualidades académicas, las cuales se reflejan en su pasión por la investigación. Es una persona que con su testimonio ha impactado mi vida al descubrir, que es una persona de Ciencia pero la misma no lo ha alejado de Dios. Considero que conciliar estos dos elementos demuestra que es una persona humilde y consciente del fin último del ser humano en el tiempo.

TABLA DE CONTENIDO

RESUMEN

Esta Tesis se enfoca en el diseño de un Mecanismo Computacional para extraer patrones de las secuencias de ADN. La característica de los patrones es que deben ser Secuencias Frecuentes Maximales (SFM), esto quiere decir que son secuencias frecuentes que no son subsecuencias de ninguna otra secuencia frecuente. El mecanismo se basa en el algoritmo DIMASP, que ha sido utilizado por sus desarrolladores para descubrir patrones en texto (KDT). Este algoritmo tiene dos características principales: se basa en el crecimiento de patrones, y en la extracción de todas las SFM con independencia del umbral de frecuencia. El aporte general de las SFM es que se reducen considerablemente las Secuencias Frecuentes, obteniendo una compresión del total de patrones secuenciales, por ende, se reduce el espacio de almacenamiento y de búsqueda de datos. En lo referente a las cadenas de ADN, la aportación de esta técnica complementada con otras tecnologías, es una alternativa para analizar múltiples secuencias de ADN. De esta manera se colabora con el problema de desfase entre las secuencias que se generan y las que se analizan.

INTRODUCCIÓN

La Minería de Datos (MD), se define como “el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos”. (Witten, 2005). De acuerdo con esto, la Minería de Secuencias, es un caso particular de la Minería de Datos, que enfocada a este proyecto consiste en extraer Patrones Frecuentes Maximales, de colecciones de datos que están representados de forma secuencial. (Mabrouketh, 2010). Siendo el objeto de estudio, la extracción de Secuencias Frecuentes Maximales de las cadenas de Ácido Desoxirribonucleico (ADN), el trabajo se desarrolla en el área de Bioinformática.

La Bioinformática es una disciplina emergente que utiliza las tecnologías de la información para conocer, organizar, analizar y distribuir información biológica con el propósito de responder preguntas complejas en biología. Es un área de investigación multidisciplinaria, que puede definirse como la interfaz entre dos ciencias: la biología y la computación, impulsada por la incógnita del genoma humano y la promesa de una nueva era en la que la investigación genómica puede ayudar a mejorar la condición y la calidad de vida humana. (Cañedo, 2004). Uno de los retos de la bioinformática consistió en la secuenciación del genoma humano, lo cual originó grandes volúmenes de información que debe analizarse para poder ser utilizada de manera conveniente.

La secuenciación del Ácido Desoxirribonucleico (ADN) se refiere a los métodos para determinar el orden de las bases de nucleótidos, denominados estos como Adenina (A), Guanina (G), Citosina (C) y Timina (T) [4]. (Ver Figura 1). El problema de analizar las secuencias de genomas es particularmente complejo por el gran tamaño de las cadenas de ADN, como es el caso de las eucariotas (células con núcleo definido por una membrana, en el

cual se almacena toda la información genética), (Ver Figura 2). Esto representa para el área de computación un problema de tipo NP-Completo, lo cual significa que el tiempo de ejecución crece exponencialmente con relación a la longitud de las secuencias, ya que para obtener patrones frecuentes los algoritmos tienen que revisar explícitamente 2^{m-1} conjuntos de elementos (*m* es el número de elementos de la secuencia), (Dao-I, 1998). Así pues, para encontrar un conjunto Maximal de tamaño 100, se tendrían que revisar 2^{100-1}, resultando aproximadamente 10^{30} subconjuntos. (García et. al, 2006). Por lo tanto se requiere investigar maneras adecuadas para la implementación de algoritmos que mejoren el análisis, la estructura, identificación, velocidad, precisión y almacenamiento de los datos.

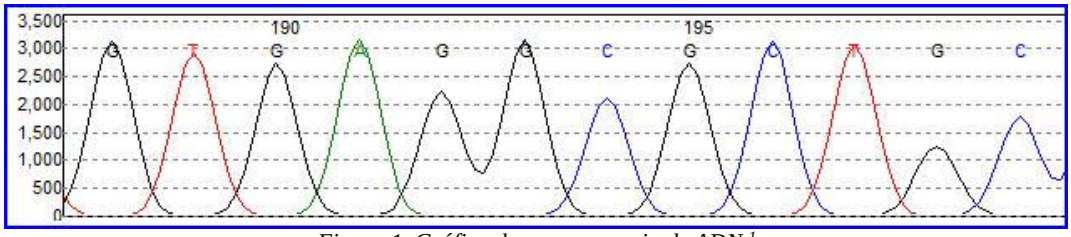

Figura 1. Gráfico de una secuencia de ADN.[1]

Figura 2. Estructura de célula eucariota.[2]

[1] https://es.wikipedia.org/wiki/Secuenciaci%C3%B3n_del_ADN
[2] http://funcionde.com/celula-eucariota/

En particular, en esta tesis se aborda el problema de la extracción de las Secuencias Frecuentes Maximales (SFM) de cadenas de ADN, (Definiendo SFM como una Secuencia que no es subsecuencia de ninguna otra Secuencia Frecuente Maximal) (García et. al, 2006). Es imprescindible encontrar la manera de comprimir los datos para analizar, almacenar y recuperar las secuencias cuando se requiera. La implementación de la Minería de Datos, ocupa un lugar sobresaliente, al transformar los datos en información y ésta a su vez en conocimiento. Esta transformación implica el utilizar heurísticas para obtener, seleccionar y estructurar los datos, ya que la cantidad de información que comprende un genoma humano suscita un problema por la cuantía de combinaciones al extraer las SFM. En este trabajo se propone un mecanismo computacional, para la extracción de las SFM, de las cadenas de ADN, basándose en un algoritmo para Descubrir Patrones de Secuencias Maximales (DIMASP del inglés Discovery all the Maximal Sequential Patterns), el algoritmo se basa en la técnica patrón de crecimiento, además es independiente del umbral de soporte. Con esta extracción de patrones, se comprime la información de las secuencias frecuentes, esto ayuda a reducir el espacio de almacenamiento y de búsqueda de las secuencias. Este algoritmo se ha implementado en la Minería de texto, con información de noticieros. (García et. al, 2006).

La estructura del resto de este documento se describe de la siguiente manera: El Capítulo 1. Generalidades. Comprende el planteamiento del problema, propuesta de solución, justificación, objetivo general, objetivos específicos, hipótesis, metodología y antecedentes. El Capítulo 2. Estado del Arte. Contiene el Estado del Arte y preliminares. El Capítulo 3. Desarrollo de la metodología. Describe los objetivos específicos como parte del desarrollo de la metodología. El Capítulo 4. Experimentos y Análisis de resultados. Al final se presentan

con el siguiente orden las conclusiones, trabajo a futuro, referencias, índice de figuras e índice de tablas.

CAPÍTULO 1.

GENERALIDADES

1.1 Planteamiento del problema

En los últimos años, la Secuenciación del ADN (entiéndase por Secuenciación: los métodos para determinar el orden de las bases de nucleótidos, denominados estos como Adenina (A), Guanina (G), Citosina (C) y Timina (T) (Green, 2015), (Ver Figura 1), ha incrementado el volumen de datos de manera exponencial, gracias a la creación e implementación de plataformas de Secuenciación de Próxima Generación. Algunas de esas plataformas son Illumina, SOLiD, 454 (Roche), HeliScope y Complete Genomics. Las secuencias de nucleótidos son datos primarios, por lo que es necesario analizar las cadenas de ADN implementando algoritmos computacionales para descubrir conocimiento. El problema es, el desfase que hay entre las secuencias que se generan y las que se analizan. Se han desarrollado diferentes herramientas efectivas para el análisis. No obstante, todavía existe una brecha de desarrollo entre las secuencias y los resultados del análisis. (Mckenna, 2010). Una de las fases de la Secuenciación del ADN, es el alineamiento para comparar secuencias y determinar las relaciones funcionales o evolutivas de acuerdo a la similitud y homología, entre genes consultados.

Existen diferentes técnicas para el análisis de información, una de ellas es la Minería de Datos (MD). La Minería de Secuencias es un caso particular de la MD. La tarea de Minería de Secuencias Frecuentes Maximales de las cadenas de ADN, se considera un problema, debido al crecimiento exponencial de información generada por la secuenciación del ADN. (Ver Figura 3). Este problema hace difícil el análisis de las secuencias de ADN para conocer la función de los genes, la similitud y homología entre secuencias o padecimientos de algún individuo. Con el análisis de SFM también se puede saber qué regiones son evolutivas y que regiones son conservadas. (Escobar, 2011)

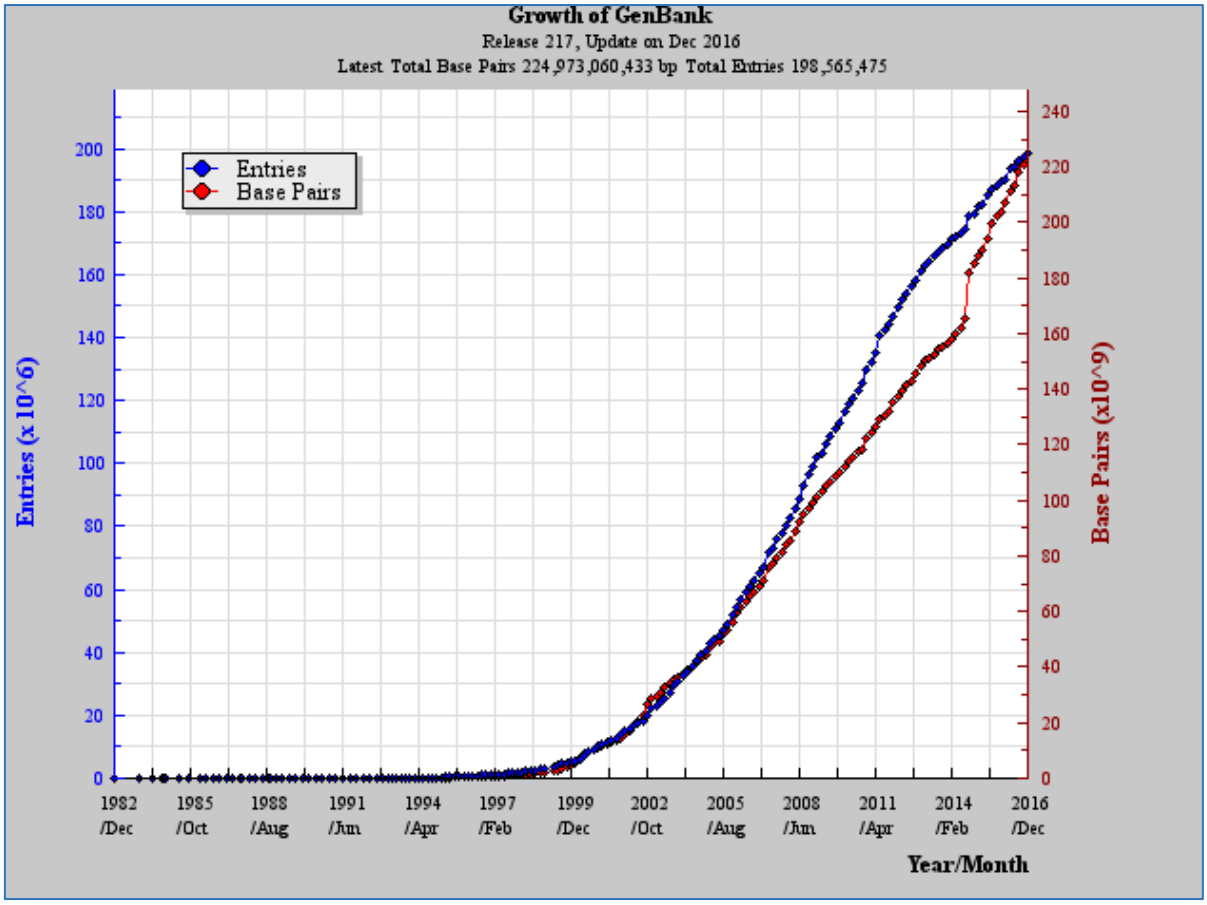

Figura 3. Crecimiento del volumen de información del ADN.[3]

El análisis de las cadenas de ADN enfrenta situaciones complejas por la cantidad de información que se maneja, como: el problema de combinatoria (ver Figura 4) y las limitaciones computacionales por la cantidad de procesos, el espacio en memoria y el tiempo de ejecución. El National Institute General Medical of Sciences, publicó en su sitio web, que: se tiene conocimiento de 3, 200, 000,000 Mbps (millones de pares de bases) que forman el genoma humano (Toledo, 2012). La National Library of Medicine, reportó que GenBank contiene 185 millones de secuencias de más de 365,000 especies diferentes (Humphreys, 2016), (Ver Figura 5). Para hacer una comparación de una secuencia solicitada en tal cantidad

[3] http://www.gen-info.osaka-u.ac.jp/~uhmin/reference/GrowthOfGenBank/graph.html

de información resulta imposible resolverlo de manera exhaustiva. El costo computacional del problema de hallar SFM en cadenas de ADN (Ácido Desoxirribonucleico), lo clasifica como un problema con un grado de complejidad NP-Completo (Liu, 2000). Se requiere de estrategias heurísticas para su solución.

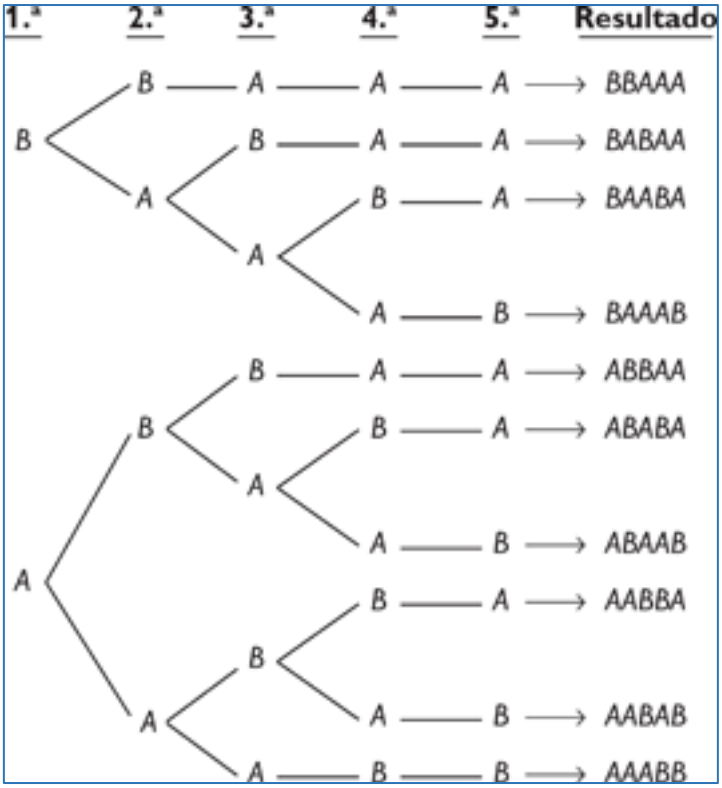

Figura 4. Ejemplo de combinatoria.[4]

La Figura 4, muestra parte de la permutación de 2 objetos diferentes para una secuencia de longitud 5, ya que una combinatoria donde el orden importa, da como resultado una permutación de 32 secuencias diferentes, de acuerdo con la fórmula n^r. Para una secuencia de 2 objetos diferentes y con una longitud de 100 elementos, resultan: 1.26765060022823e+30,

[4] https://www.pinterest.es/pin/384283780678225428/

secuencias diferentes. Como se puede observar la cantidad de secuencias crece exponencialmente.

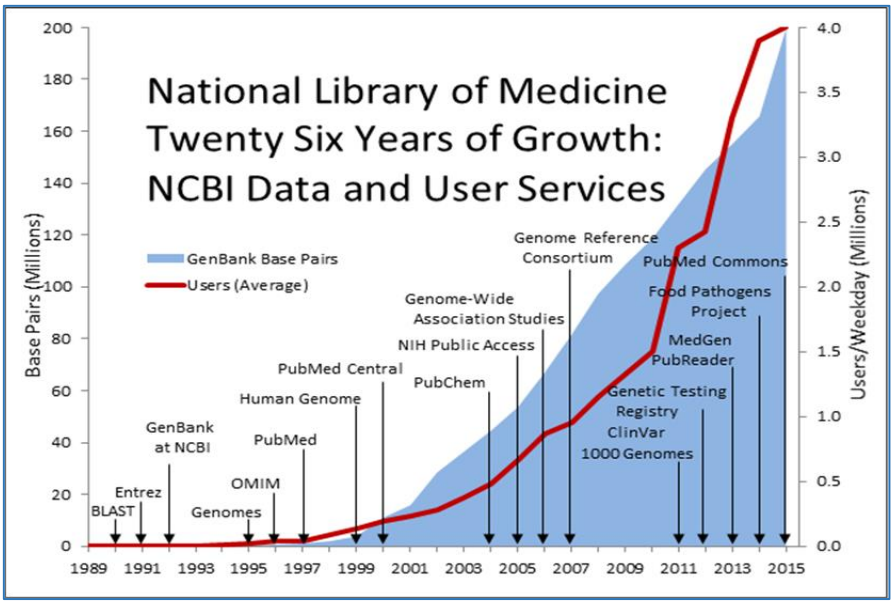

Figura 5. Datos de NCBI y servicios de usuario.[5]

En consecuencia de lo anterior se exponen algunos de los problemas al llevar a cabo el análisis de las secuencias de ADN.

- Las Instancias reales de cadenas de ADN, son muy largas. Para poder leer estas cadenas se debe ir fragmentando la información, de esta manera se obtienen todas las subsecuencias, de "*e"* elementos de una cadena dada. Para *n* secuencias individuales, el método requiere construir el equivalente *n*-dimensional de la matriz formada en el alineamiento estándar de pares de secuencias de la programación dinámica. De esta forma, el espacio de búsqueda se incrementa exponencialmente conforme se incrementa el número de secuencias y la longitud de las secuencias. Analizar de esta forma *n* secuencias ha mostrado ser un problema NP-completo (Edgar, 2004).

[5] https://www.ncbi.nlm.nih.gov/

- El no saber dónde empieza una secuencia solicitada y dónde termina, requiere de todas las combinaciones posibles que se originan de una cadena de gran tamaño, generando problemas de requerimientos de procesos y espacio en memoria al cargar una cadena completa de ADN en una computadora.
- Identificar donde empiezan y acaban los genes y averiguar qué actividad realiza, es lo que técnicamente se conoce como ‘anotar el genoma’. Entender qué significan las letras de las cadenas de ADN guardadas en las bases de datos, es un trabajo menos mecánico que el de simplemente leer el ADN y, por tanto, más complejo y largo. Aunque el número de genes en nuestro genoma es limitado (Unos 25.000), las numerosas interacciones entre ellos y la multitud de funciones que tienen hacen predecir que se llevará unas cuantas décadas conocer, cómo se regula el cuerpo humano a nivel genético. (Macip, 2010).
- Comparar secuencias a través de la búsqueda, tiene un alto costo computacional. En la práctica se emplean diversas heurísticas que, muchas veces distan de ser eficientes, si bien son más rápidas, deben garantizar que el alineamiento óptimo sea encontrado, lo cual no es una tarea fácil, por la integración de los diferentes elementos como clúster de computadoras, estructuras de datos, almacenamiento, carga en memoria, repartición de procesos y tiempo. Aquí radica la importancia de aplicar la potencia de cómputo de plataformas paralelas para acelerar el procesamiento de las secuencias sin perder precisión en los resultados (Rucci, 2013).

Por lo tanto en esta tesis se busca aplicar el algoritmo DIMASP, que tiene la característica de crecimiento de patrones y maneja umbrales diferentes, para extraer todas las Secuencias Frecuentes Maximales, esto se adaptará al proceso del análisis de las cadenas de ADN, reestructurando los documentos de entrada del algoritmo.

1.2 Propuesta de solución

En el campo de la Minería de Datos y extracción de conocimiento, la minería de secuencias es un caso particular de la minería de datos estructurados. Consiste en encontrar patrones estadísticamente relevantes en colecciones de datos que están representados de forma secuencial (Mabroukeh, 2010). En este caso se aplica a secuencias de ADN, extrayendo un conjunto de patrones que tienen la característica de ser Maximales y que representan todas las secuencias frecuentes que hay en una colección de secuencias. En este trabajo se pretende extraer las Secuencias Frecuentes Maximales, que se definen como: secuencias frecuentes que no son sub-secuencias de ninguna otra SFM. La aportación final es poder comparar múltiples secuencias de ADN, de longitud considerable, como una alternativa para descubrir similitud, evolución, funciones, detectar y prevenir enfermedades en los individuos, diseñar medicamentos para la medicina personalizada, con el objetivo de que sean menos los efectos colaterales. (Ginsburg, 2001)

Como solución al problema de Minar las secuencias de las bases de datos de ADN, se propone desarrollar un mecanismo integrando elementos computacionales para la extracción de Secuencias Frecuentes Maximales y así disminuir la cantidad de información para facilitar el análisis de secuencias. Algunos de los procesos de este mecanismo es: obtener los documentos con las secuencias de los nucleótidos, a través de instituciones que se dedican a la investigación genómica, por ejemplo NCBI (National Center for Biotechnology Information), este centro gestiona la base de datos GenBank. Otro proceso es, reestructurar el documento donde se almacena la colección de secuencias. Este documento se utiliza como entrada de datos. Por último utilizar un algoritmo para la extracción de las Secuencias Frecuentes Maximales.

Para trabajar en la extracción de SFM, se propone utilizar el algoritmo DIMASP. Este algoritmo extrae todas las SFM, de una colección de secuencias, que se basa en la técnica patrón de crecimiento, además es independiente del umbral de soporte (García et al, 2006). Adaptar los documentos de entrada que contienen los elementos de las cadenas de ADN es parte esencial para poder ejecutar el algoritmo. El algoritmo de manera general se desarrolla en 4 etapas: Primera etapa, se asigna un número entero como identificador, a cada uno de los elementos diferentes en la base de datos y a cada uno de los pares de elementos contiguos. En la segunda etapa, a partir de los pares de elementos de la BDD, se construye la estructura de datos, almacenando en ella todos los pares contiguos de los elementos, respetando el orden de la secuencia, esto permitirá buscar las SFM. En la tercera etapa, el algoritmo busca para cada par de elementos de la BDD y de acuerdo al umbral β especificado por el usuario, la Secuencia Frecuente (SF) más larga que se pueda formar a partir de ese par de elementos y los almacena como posibles SFM. En la cuarta etapa, se seleccionan todas las SFM, a partir del conjunto de Secuencias Frecuentes.

El algoritmo DIMASP se ha implementado para documentos que contienen texto, específicamente de noticieros. En el caso de las secuencias de ADN es un alfabeto de solo 4 caracteres (A, G, C y T), sin embargo son cadenas de tamaño considerable. Lo que hace el algoritmo es encontrar las Secuencias Frecuentes Maximales de las cadenas de ADN.

1.3 Justificación

La secuenciación del ADN se refiere a los métodos para determinar el orden de las bases de nucleótidos, denominados estos como: Adenina (A), Guanina (G), Citosina (C) y Timina (T) que forman el genoma de cada organismo (NIH, 2015). La tarea de analizar genomas es

particularmente compleja. Cuando se trabaja con genomas de gran tamaño, como es el caso de la mayoría de las células eucariotas (animales, plantas y hongos), es importante contar con mecanismos eficientes y precisos que permitan analizar esas secuencias (Minetti, 2011). Dentro de las técnicas computacionales para extraer conocimiento, está la minería de patrones, que en el caso de las cadenas de ADN, se trata de hacer el análisis de secuencias para extraer información que representa la estructura molecular de ciertos elementos o compuestos que forman un ser vivo y que puede ser utilizada por ejemplo: para saber qué repercusiones tienen los organismos y así resolver problemas por los que se ven afectados.

El conocimiento de las secuencias en una molécula de ADN, se ha hecho indispensable para el estudio de los procesos biológicos de investigación, pudiéndose aplicar en un diagnóstico médico o investigación forense (Alonso, 2005). El orden de las bases de datos a lo largo del ADN contiene el conjunto completo de instrucciones que son la causa de múltiples factores como la herencia genética.

La importancia de analizar las Secuencias de ADN:

- Es importante secuenciar para poder analizar la estructura y función de las cadenas de ADN. Se debe tener el conocimiento de su estructura primaria es decir su secuencia de ADN para analizarlas y conocer las funciones, similitud, homología, evolución de los genes de cada organismo.
- Conocer cuál es la función de una secuencia específica y saber si es responsable de alguna enfermedad, de esta manera se puede prevenir y actuar en consecuencia de la información obtenida.

- Al hacer un estudio comparativo en secuencias de ADN de la misma especie, podemos detectar mutaciones o regiones conservadas, esto ayuda a diseñar medicamentos efectivos tomando en cuenta posibles mutaciones en el futuro.
- Resuelve conflictos de progenitores y parentesco. El ser humano tiene en casi todas sus células 23 pares de cromosomas que los progenitores aportan en el momento de la procreación. Gracias a esto se puede hacer la prueba de paternidad y consanguinidad.
- La huella de ADN, es lo que nos hace únicos e irrepetibles, conocerla puede ayudar a resolver problemas particulares a cada individuo.
- Al conocer la secuencia completa del genoma, el objetivo del proyecto del genoma humano se completa, acercándose a la medicina personalizada. Por ahora la medicina personalizada es un paradigma que existe más en términos conceptuales que en la realidad, no obstante cada vez es más inminente debido a la mayor conciencia de las deficiencias en la administración de fármacos que tienen beneficios pero también tienen riesgos para los pacientes. Con esto se trataría de forma integral al paciente, evitando efectos colaterales que provocan otras enfermedades.
- El mercado de las pruebas de diagnóstico molecular se prevé que crezca a tasas extraordinarias en los próximos años, impulsado por la farmacogenética y el sueño de alcanzar la medicina personalizada (Lesko, 2007).

Como dato, el cromosoma de Escherichia coli tiene 4×10^6 pares de bases, 4Mpb. A un laboratorio de la mitad de la década de los setentas le habría tomado dos meses secuenciar 150 nucleótidos. Actualmente, un laboratorio especializado es capaz de secuenciar millones de nucleótidos al día. Desde esta perspectiva, resulta notoria la capacidad de las tecnologías disponibles actualmente, para el análisis de ácidos nucleicos; no obstante se puede mejorar

(Necochea, 2004). Si el caso de la Escherichia coli con 4 Mpb, se convierte en un problema computacional muy fuerte por la cantidad de combinaciones que se deben generar para poder identificar dónde se encuentra la información que yo estoy buscando, se hace exponencialmente más fuerte cuando un genoma tiene el tamaño de 3,200 Mpbs, como es el caso del genoma humano.

El costo de los equipos y el proceso de la secuenciación, siguen siendo elevados. Secuenciar el ADN de una persona tiene un costo alrededor de 1,000 dlls. Al reducir el tiempo de procesamiento se reducen los costos, los sistemas se hacen más eficientes al obtener los resultados en menor tiempo. Los biólogos investigadores son los primeros en beneficiarse de los adelantos en la secuenciación porque se avanza más rápido en el conocimiento. Otro beneficio es conocer las áreas reservadas de las secuencias de ADN para el desarrollo de medicamento. La medicina personalizada podrá desarrollar medicamentos de manera individual, reduciendo los efectos colaterales en los individuos.

El Alineamiento de secuencias, como se le conoce a la comparación de secuencias de ADN en Bioinformática, se divide en dos categorías: alineamiento global y alineamiento local. Estas técnicas tienen ciertas limitaciones o desventajas. Los alineamientos globales son útiles en secuencias muy similares y aproximadamente del mismo tamaño. Los alineamientos locales toman en cuenta solo regiones de mayor similitud entre las secuencias. Aunque se consideran métodos de optimización se reconoce que son lentos y solo hacen comparaciones por pares de secuencias.

Al aplicar el algoritmo DIMASP, se extraen las Secuencias Frecuentes Maximales, de las secuencias de ADN, con la ventaja de que este algoritmo, tiene la característica de crecimiento de patrones lo cual evita la generación de candidatos y se restringe la búsqueda en la base de

datos. La aportación de este proyecto es la aplicación del algoritmo para extraer las SFM de cadenas de ADN, con esto se reduce el espacio de almacenamiento y de búsqueda de la información ya que las SFM son una representación de todas las secuencias frecuentes.

1.4 Objetivo general

Desarrollar una propuesta de mecanismo computacional, basado en el algoritmo DIMASP, para extraer las Secuencias Frecuentes Maximales de una colección de secuencias de ADN.

1.5 Objetivos Específicos

- Diseñar la propuesta, con elementos computacionales como: bases de datos, normalización, pre-procesamiento, algoritmo, datos de entrada, datos de salida.
- Recopilar los datos de los repositorios de bioinformática que corresponden a las secuencias de nucleótidos.
- Analizar el algoritmo DIMASP para conocer cuáles son sus entradas, salidas y diferentes etapas.
- Reestructurar los documentos en formato FASTA, para normalizar la estructura del contenido de los datos de entrada que utiliza el algoritmo DIMASP.
- Almacenar la colección de secuencias de nucleótidos en un solo archivo, que funciona como datos de entrada del algoritmo.
- Implementar el mecanismo para la extracción de Secuencias Frecuentes Maximales.
- Realizar pruebas del mecanismo para su valoración.

- Analizar los resultados, para mostrar que se logró el objetivo de extraer todas las SFM de la colección de secuencias de ADN.

1.6 Hipótesis

Es posible resumir los patrones frecuentes de una colección de secuencias de ADN, mediante un mecanismo computacional, basado en el algoritmo DIMASP, con el objetivo de extraer todas las Secuencias Frecuentes Maximales, para reducir el espacio de almacenamiento y de búsqueda de secuencias de ADN.

1.7 Metodología

En este apartado se describen los pasos y procedimientos, para lograr el objetivo general que implica desarrollar una propuesta de un mecanismo computacional basado en el algoritmo DIMASP, para extraer las Secuencias Frecuentes Maximales, de una colección de secuencias de ADN.

Los pasos de la metodología son los siguientes:

1. Obtener los documentos de las cadenas de ADN, accediendo a las plataformas web que proporcionan las instituciones que se dedican a la investigación relacionada con el genoma.

2. Analizar la estructura y características que tienen los documentos de las bases de datos de las cadenas de ADN, como el formato, el tamaño, el tipo de organismo al que pertenece dicha secuencia. Se trabaja con secuencias de organismos que tienen la característica de estar formados por células eucariotas.

3. Analizar el algoritmo DIMASP para identificar los procesos en cada etapa del algoritmo. Se ejecuta para explorar las salidas, los datos que arroja el resultado.

4. Implementar una función para reestructurar los documentos que contienen las secuencias de ADN. Los datos de entrada del algoritmo DIMASP, tienen los siguientes requerimientos: se integrará en un solo archivo la colección de secuencias de ADN, el formato del documento debe estar en texto plano, cada secuencia de las cadenas de ADN se almacena en una sola línea, cada cadena tiene que identificarse con un número seguido del signo igual, los nucleótidos representados por los caracteres A, C, G, T, deben estar separados por una coma, a excepción del último carácter de cada línea o secuencia; todo es sin espacios.

5. Diseñar el mecanismo de manera gráfica, integrando los elementos computacionales. Recabar la información de las bases de datos, normalizar los datos, integrar las secuencias en un solo archivo, integrar el algoritmo DIMASP para la Extracción de Secuencias Frecuentes Maximales.

6. Realizar experimentos con el mecanismo para su valoración. Se prueba el mecanismo con una colección de secuencias de ADN, luego con colecciones de diferentes longitudes y con diferentes umbrales, siempre con un GAP = 0.

7. Analizar los resultados con respecto al umbral, la cantidad de SFM, y su longitud.

1.8 Antecedentes

El avance en la secuenciación de los ácidos nucleicos ha generado un amplio conocimiento en el campo de la genómica, estos beneficios se extienden a la farmacogenómica aplicándose a la medicina personalizada. Actualmente se obtiene gran cantidad de información con aplicaciones innumerables. Entre otras cosas, la secuenciación ha permitido entender la

asociación de enfermedades con la variabilidad genética, la función de genes, el patrón de expresión de genes nuevos, la similitud o variación genética entre especies diferentes, la organización de la información genética, el origen de algunos genes, etc. (Hutchinson, 2007), (Shendure, 2008).

Las herramientas de software que facilitan la investigación en bioinformática pueden clasificarse en cuatro clases, sin embargo las que interesan para nuestro objetivo son las de comparación y alineación de secuencias y descubrimiento de patrones. (Meneses, 2011). Algunas de ellas son:

- Herramientas principales de gestión de bases de datos biológicas: GenBank (USA), EMBL (Europa) y DDBJ (Japón), para alineamiento local por pares de secuencias.
- BLAST, comparación y alineación secuencias, realiza búsquedas, en la totalidad de una base de datos no redundante en poco tiempo. Su principal característica es la velocidad.
- FASTA, se puede utilizar para hacer una comparación rápida de proteínas o de nucleótidos. Alcanza un alto nivel de sensibilidad para la búsqueda de similitud mediante la realización de búsquedas optimizadas para alineamientos locales utilizando una matriz de sustitución.
- ClustalW para alineación de secuencias múltiples, la cual se puede utilizar para alinear las secuencias de ADN o de proteínas con el fin de dilucidar sus relaciones, así como su origen evolutivo (Meneses et. al, 2011).

Minería Secuencial de Patrones, tiene como objetivo encontrar todas las subsecuencias que están contenidas al menos β veces en una colección de secuencias, donde β es el umbral de soporte especificado por el usuario.

Secuencias Frecuentes Maximales, es una secuencia frecuente que no es subsecuencia de ninguna otra secuencia frecuente. Estas son representaciones compactas de todo el conjunto de secuencias frecuentes.

Preliminares

Definición del Problema en el análisis de secuencias:

Una *secuencia S,* denotada por $<n_1, n_{2,\ldots}n_k>$, es una lista ordenada de *k* componentes llamados elementos. El número de elementos en una secuencia S es la longitud de la secuencia denotada por |S|. Una k-secuencia denota una secuencia de longitud k. Tenemos que $P=<p_1p_2\ldots p_n>$ y $S=<s_1s_2\ldots s_m>$ secuencias, P es una subsecuencia de S, denotada $P \subseteq S$, si existe un entero $i \geq 1$, de tal manera que $p_1=s_i, p_2=s_{i+1}, p_3=s_{i+2}, \ldots p_n=s_{i+(n-1)}$.

La frecuencia de una secuencia S, denotada por S_f o $<s_1,s_2,\ldots s_n>_f$, es el número de documentos donde S es una subsecuencia. Una secuencia S es β-*frecuente* si $S_f \geq \beta$, una secuencia β-*frecuente* es solo llamada un patrón secuencial. Un patrón secuencial S es *maximal* si S no es una subsecuencia de ningún otro patrón secuencial (García, 2006).

CAPÍTULO 2.

ESTADO DEL ARTE

En este apartado se citan trabajos de investigación referentes a las técnicas utilizadas para resolver problemáticas similares, que tienen relación con el tema de las secuencias frecuentes o que tienen relación con el tema de esta tesis, que trata de la extracción de Secuencias Frecuentes Maximales de una colección de secuencias de ADN, se fundamenta en la importancia que tiene el descubrir patrones de información. También se mencionan las técnicas que se utilizan para hacer comparaciones, alineamientos y búsquedas en bases de datos de las cadenas ADN. En lo particular, se hace referencia a trabajos relacionados con la Minería de Secuencias. Como ya se ha mencionado, descubrir patrones de los datos es relevante para obtener conocimiento de acuerdo al objeto de estudio que se trata y en este caso lo relevante de las SFM, es que son la representación de todas las secuencias frecuentes de las cadenas de ADN relacionadas.

2.1 Minería de Secuencias Frecuentes

(Agrawal, 1995) Los autores proponen un algoritmo para resolver el problema de extraer patrones secuenciales de una base de datos de secuencias. Este algoritmo es GSP (Por sus siglas en Inglés: Generalized Sequential Patterns), se trata de descubrir patrones secuenciales generalizados. La naturaleza de los datos que tratan son secuencias de una lista de transacciones, donde cada transacción es un conjunto de literales llamados elementos. El descubrimiento de los patrones se obtiene con el mínimo soporte que se calcula de acuerdo al porcentaje en que aparece el patrón en la base de secuencias. El algoritmo GSP tiene una estructura de varios pasos sobre la base de datos. Se implementa en dos fases. Primero determina el soporte de cada elemento, es decir el número de veces que las secuencias contienen el elemento, con esto el algoritmo conoce cuales elementos son frecuentes con el

mínimo soporte. Estos elementos frecuentes pasan a ser secuencias frecuentes que se toman como prefijos o semillas para seguir generando secuencias candidatas. Luego viene la fase de generación de secuencias candidatas, poda, conteo y eliminación de las k-secuencias cuyo soporte está por debajo del mínimo soporte. Este algoritmo aunque restringe la generación de candidatos, tiene la desventaja de escanear la base de secuencias más de una vez.

(Pei et. al, 2001) PrefixSpan. Este algoritmo es parte del desarrollo de métodos que buscan descubrir patrones. Su objetivo es reducir sustancialmente el número de combinaciones que debe examinarse. Es un reto ya que se tiene que examinar un número explosivo de combinatoria de patrones de las subsecuencias. Este algoritmo, propone la proyección de los prefijos en la extracción completa de las secuencias de patrones, reduce la generación de subsecuencias candidatas. Sin embargo se sigue teniendo problemas en cuanto a bases de datos muy grandes o cuando los patrones secuenciales extraídos son numerosos. La proyección de prefijos reduce el tamaño de proyección de bases de datos y conduce a un procesamiento eficiente. Supera a los algoritmos GSP y FreeSpan.

(Bergroth et. al, 2002) El objetivo de los autores es dar a conocer métodos para resolver el problema de Secuencia Común de Longitud mayor (Por sus siglas en Inglés LCS). Los métodos se utilizan para hacer comparaciones entre dos secuencias. Su utilidad se aplica a entornos para corregir errores de entradas de palabras buscando la más parecida. También se utiliza en Biología Molecular, para comparar pares de cadenas de ADN o secuencias de proteínas para saber si son homologas. Una medida obvia es encontrar el mayor número de coincidencias exactas entre dos cadenas, preservando el orden entre ellas. Tratan el problema de LCS como un caso a resolver el problema de distancias entre dos secuencias. La Distancia entre una secuencia X y una secuencia Y es definida como el número mínimo de operaciones

elementalmente necesarias para convertir la cadena fuente X en la cadena objetivo Y. En la práctica real está restringido para inserciones, eliminaciones y sustituciones. Utilizan la estructura de tabla donde los elementos son considerados como vértices en un gráfico y los valores definen los bordes. La tarea es encontrar el camino más largo entre los vértices en la esquina superior izquierda e inferior derecha de la tabla.

(Zaki, 2001) Extracción de secuencias frecuentes con algoritmo SPADE. Utilizan propiedades combinatorias para descomponer el problema principal en pequeños subproblemas, que pueden ser resueltos independientemente en memoria principal, usando técnicas eficientes de búsqueda de enrejado, las secuencias se obtienen a través de 3 escaneos a las bases de datos. Como resultado sí se extraen las secuencias. Pueden utilizarse en una aplicación real.

(Fogglino, 2009) FINDPAT es un algoritmo que encuentra repeticiones (patrones) maximales con matching exacto dentro de secuencias de ADN (alfabeto fA, C, G, Tg), y sin límite en la longitud de las repeticiones que este arroja. Esto lo hace de una manera conocida como ab initio, ya que necesita de alguna semilla (secuencia de referencia) previa para su funcionamiento. Solamente necesita como entrada una secuencia o dos (en el caso de buscar patrones entre dos secuencias), y la longitud mínima deseada para las repeticiones resultantes. Este algoritmo puede recibir secuencias genéticas de hasta 500 Megabytes, es así que brinda la posibilidad de analizar cromosomas enteros (un cromosoma, o parejas de cromosomas). Por una cuestión de simplicidad y más que nada por cuestiones de implementación del algoritmo, se buscan repeticiones con correspondencia exacta, es decir que no se permite modificación, eliminación de caracteres para la búsqueda de los patrones.

2.2 Algoritmos Minería de Secuencias Maximales

(García et al. 2006), desarrollaron el algoritmo DIMASP, para descubrir Patrones de Secuencias Maximales (PSM's), este algoritmo tiene la característica que se basa en el crecimiento de patrones con independencia del umbral que se requiera. Estos PSM's son una representación de todas las secuencias frecuentes con lo que se comprime la información, reduciendo el espacio de almacenamiento de la información y el espacio de búsqueda. El algoritmo comprende 4 Pasos Primero asigna un número entero como identificador a cada palabra diferente de las secuencias de la base de datos y genera una estructura de datos asignando un identificador a cada par contiguo de palabras, respetando el orden de las palabras. En el tercer paso, el algoritmo busca a partir de los pares de palabras la secuencia frecuente más larga de acuerdo al umbral requerido. En el cuarto paso selecciona todas las SFM, a partir del conjunto de Secuencias Frecuentes. Los resultados muestran que DIMASP supera a los algoritmos GSP, DELISP GenPrefixSpan and SPADE, por su buena escalabilidad en cuanto al umbral, descubre secuencias más largas con un umbral mayor a dos.

2.3 Algoritmos Alineación de Secuencias de cadenas de ADN

(Altschul et. al, 1990) BLAST, este algoritmo permite comparar una secuencia de ADN contra una, o un conjunto de bases de datos, e identificar secuencias dentro de estas con las cuales se asemeja. Introduce varios refinamientos a la búsqueda en bases de datos, que mejoran el tiempo de búsqueda. Hace énfasis en la velocidad por sobre la sensibilidad. Esto es fundamental para que el algoritmo sea práctico al buscar en las bases de datos gigantes que están disponibles hoy día. No está basado en un algoritmo que garantiza el alineamiento óptimo, sino que usa una heurística que funciona la mayoría de las veces en la práctica, así

que, podría fallar con algunas secuencias poco relacionadas entre sí. Es alrededor de 50 veces más rápido que otros algoritmos que garantizan el alineamiento local de secuencias óptimo y usan programación dinámica.

CAPÍTULO 3.

DESARROLLO DE LA METODOLOGÍA

En este capítulo se describen las acciones que se realizaron de acuerdo a la metodología, para lograr la propuesta de un mecanismo con el objetivo de extraer las SFM de las secuencias de ADN, mediante el algoritmo DIMASP. A continuación, se puntualizan los paso de la metodología:

3.1 Obtener las Secuencias de ADN

Para obtener las Secuencias de ADN, se investigó cuáles son las principales Instituciones internacionales que se dedican a la investigación del genoma. Las tres Instituciones más reconocidas son: El **C**entro **N**acional para la **I**nformación **B**iotecnológica (NCBI) de Estados Unidos de Norte América. El **L**aboratorio **E**uropeo de **B**iología **M**olecular (EMBL) con sede en Reino Unido. El **B**anco de **D**atos **D**NA de **J**apón (DDJB). Lo importante de estas instituciones es que sus bases de datos se actualizan todos los días para compartir la misma información a los usuarios. Después de explorar las plataformas se decidió obtener las secuencias ADN de la plataforma NCBI, por tener una interfaz más amigable.[6]

3.2 Analizar la estructura y características de los diferentes formatos de las cadenas de ADN.

En este paso se accedió a la plataforma NCBI para, analizar los diferentes formatos de los documentos que almacenan las secuencias de ADN, conocer cómo está estructurada la información que representa a las cadenas de ADN, saber qué longitudes tienen las cadenas y conocer el tipo de organismo al que pertenecen las secuencias.

[6] https://www.ncbi.nlm.nih.gov/

En esta parte se presentan algunos tipos de formatos que se manejan en las plataformas, para almacenar las secuencias de ADN:

```
LOCUS       EC750390               558 bp    mRNA    linear   EST     03-JUL-2006
DEFINITION  POE00005652 PL(light) Polytomella parva cDNA similar to frataxin protein
            -related, mRNA sequence.
ACCESSION   EC750390
VERSION     EC750390.1  GI:110064507
KEYWORDS    EST.
SOURCE      Polytomella parva
ORGANISM    Polytomella parva
            Eukaryota; Viridiplantae; Chlorophyta; Chlorophyceae;Chlamydomonadales;
            Chlamydomonadaceae; Polytomella.
REFERENCE   1  (bases 1 to 558)
  AUTHORS   Lee,R.W. and Borza,T.
  TITLE     The colorless plastid of the green alga Polytomella parva: a repertoire of its functions
  JOURNAL   Unpublished (2006)
COMMENT     Contact: TBestDB
            Departement de Biochimie, Universite de Montreal
            Montreal, Canada
            Email: tbestdb-curator@bch.umontreal.ca
            Plate: 4065.
FEATURES             Location/Qualifiers
     source          1..558
                     /organism="Polytomella parva"
                     /mol_type="mRNA"
                     /db_xref="taxon:51329"
                     /clone_lib="PL(light)"
ORIGIN
        1 gcggccgctt ttttttttt ttttttttt ttttcgtccg ttatttcttt tttaagaatg
       61 cagtcatctg tacatcgtca agtattcgga gtgttatctc gttttgtggg aaacaaagcg
      121 ggtatttta caaagcataa tcatggtgtc tcaaggttgt cttcatgcac ttcgtcatgc
      181 gtaaagatgt atactagcaa caaggccccc gaggatcttc aaacgttcca ccggcaagca
      241 gacgaaactc tagagcaagt cactgaagcc cttgaaaact atgtagatga gcatgaagtg
      301 gaaggcagcg acattgagca tacgcaagga gtgcttacta ttaagcttgg aactcttgga
      361 agttatgtaa ttaataaaca gactcctaat aagcagatat ggttatcctc tcccgtcagt
      421 ggacccttcc gatatgatct taaagaaggt gcctgggttt atgaacgggc tggcgaggct
      481 cggcgcgagc ttatttctca attagaaaca gaaatttcgg atttagttgg tgtcgaatta
      541 aagataagta actgaacg
//
```

Figura 6. Ejemplo de una secuencia de ADN en formato GenBank.[7]

[7] https://bioinf.comav.upv.es/courses/intro_bioinf/bases_datos.html

```
ID   EC750390; SV 1; linear; mRNA; EST; PLN; 558 BP.
XX
AC   EC750390;
XX
DT   04-JUL-2006 (Rel. 88, Created)
DT   04-JUL-2006 (Rel. 88, Last updated, Version 1)
XX
DE   POE00005652 PL(light) Polytomella parva cDNA similar to frataxin
DE   protein-related, mRNA sequence.
XX
KW   EST.
XX
OS   Polytomella parva
OC   Eukaryota; Viridiplantae; Chlorophyta; Chlorophyceae; Chlamydomonadales;
OC   Chlamydomonadaceae; Polytomella.
XX
RN   [1]
RP   1-558
RA   Lee R.W., Borza T.;
RT   "The colorless plastid of the green alga Polytomella parva: a repertoire of
RT   its functions";
RL   Unpublished.
XX
DR   UNILIB; 42732; 19932.
XX
CC   Contact: TBestDB
CC   Departement de Biochimie, Universite de Montreal
CC   Montreal, Canada
CC   Email: tbestdb-curator@bch.umontreal.ca
CC   Plate: 4065.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..558
FT                   /organism="Polytomella parva"
FT                   /mol_type="mRNA"
FT                   /clone_lib="PL(light)"
FT                   /db_xref="taxon:51329"
FT                   /db_xref="UNILIB:42732"
XX
SQ   Sequence 558 BP; 153 A; 105 C; 127 G; 173 T; 0 other;
     gcggccgctt ttttttttt ttttttttt ttttcgtccg ttatttcttt tttaagaatg        60
     cagtcatctg tacatcgtca agtattcgga gtgttatctc gttttgtggg aaacaaagcg       120
     ggtatttta caaagcataa tcatggtgtc tcaaggttgt cttcatgcac ttcgtcatgc       180
     gtaaagatgt atactagcaa caaggccccc gaggatcttc aaacgttcca ccggcaagca      240
     gacgaaactc tagagcaagt cactgaagcc cttgaaaact atgtagatga gcatgaagtg      300
     gaaggcagcg acattgagca tacgcaagga gtgcttacta ttaagcttgg aactcttgga      360
     agttatgtaa ttaataaaca gactcctaat aagcagatat ggttatcctc tcccgtcagt      420
     ggacccttcc gatatgatct taaagaaggt gcctgggttt atgaacgggc tggcgaggct      480
     cggcgcgagc ttatttctca attagaaaca gaaatttcgg atttagttgg tgtcgaatta      540
     aagataagta actgaacg                                                   558
//
```

Figura 7Ejemplo de una secuencia ADN en formato EMBL.[8]

[8] https://bioinf.comav.upv.es/courses/intro_bioinf/bases_datos.html

```
>gi|110064507|gb|EC750390.1|EC750390 POE00005652 PL(light) Polytomella parva cDNA
 similar to frataxin protein-related, mRNA sequence
GCGGCCGCTTTTTTTTTTTTTTTTTTTTTTTTTTCGTCCGTTATTTCTTTTTTAAGAATGCAGTCATCTG
TACATCGTCAAGTATTCGGAGTGTTATCTCGTTTTGTGGGAAACAAAGCGGGTATTTTTACAAAGCATAA
TCATGGTGTCTCAAGGTTGTCTTCATGCACTTCGTCATGCGTAAAGATGTATACTAGCAACAAGGCCCCC
GAGGATCTTCAAACGTTCCACCGGCAAGCAGACGAAACTCTAGAGCAAGTCACTGAAGCCCTTGAAAACT
ATGTAGATGAGCATGAAGTGGAAGGCAGCGACATTGAGCATACGCAAGGAGTGCTTACTATTAAGCTTGG
AACTCTTGGAAGTTATGTAATTAATAAACAGACTCCTAATAAGCAGATATGGTTATCCTCTCCCGTCAGT
GGACCCTTCCGATATGATCTTAAAGAAGGTGCCTGGGTTTATGAACGGGCTGGCGAGGCTCGGCGCGAGC
TTATTTCTCAATTAGAAACAGAAATTTCGGATTTAGTTGGTGTCGAATTAAAGATAAGTAACTGAACG
```

Figura 8. Ejemplo de una secuencia en formato FASTA. [9]

En la figura siguiente se muestra un ejemplo de la estructura de los datos de entrada que se requiere para el algoritmo DIMASP.

```
1=A,C,G,C,A,C,G,T,A,T,C,C,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,
C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,
A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,C,A,C,A,C,C,A,C,
,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,T,A,
C,C,G,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,T,A,
G,T,A,A,C,C,C,T,A,A,C,C,C,T,T,T,A,C,C,C,T,A,A,C,C,C,G,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,T,T,A,
,C,C,T,A,A,C,C,C,T,T,A,A,C,C,C,T,G,A,G,T,T,A,G,G,G,T,T,A,G,G,G,T,T,A,G
2=G,C,G,C,A,C,G,T,A,T,C,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,
C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,A,C,C,C,A
,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,A,C,A,C,C,
C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,
C,C,G,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,G,T,
A,A,C,C,C,T,A,A,C,C,C,T,T,T,A,C,C,C,T,A,A,C,C,C,G,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,T,A,C,
,T,A,A,C,C,C,T,T,A,A,C,C,C,T,G,A,C,C,G,G,T,T,A,G,G,G,T,T,A,G
3=C,G,C,A,C,G,T,A,T,C,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,
A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,
,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,A,C,C,C,A,C,A,
A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,C,
,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,C,
A,A,C,C,C,T,A,A,C,C,C,T,T,T,A,C,C,C,T,A,A,C,C,C,G,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,
C,T,A,A,C,C,C,T,T,T,A,A,C,C,C,T,G,A,C,C,G,G,T,T,A,G,G,G,T,T,A,G
4=T,C,G,C,A,C,G,T,A,T,C,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,
C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A
,C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,A,C,C,C,A,C,A,C,
C,A,C,C,A,C,A,C,C,C,A,C,A,C,A,C,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,
C,C,G,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,G,T,
A,A,C,C,C,T,A,A,C,C,C,T,T,T,A,C,C,C,T,A,A,C,C,C,G,A,A,C,C,C,C,T,A,A,C,C,C,C,T,A,A,C,C,C,C,T,T,A,C,
,T,A,A,C,C,C,T,T,A,A,C,C,C,T,G,A,C,C
```

Figura 9. Ejemplo de la estructura de los datos de entrada del algoritmo DIMASP.[10]

[9] https://bioinf.comav.upv.es/courses/intro_bioinf/bases_datos.html
[10] Elaboración propia.

Como se puede observar, en la figura 8, el formato FASTA es más simple ya que almacena menos atributos de información y semeja la estructura del documento de datos de entrada del algoritmo DIMASP. El formato FASTA es un documento donde la secuencia inicia con un signo de mayor que '>', seguido por el nombre de la secuencia de ADN y otras referencias, después de esta línea, inicia la cadena de nucleótidos representados por las letras 'A, C, G, T'. De acuerdo a esto se tomó la decisión de analizar las cadenas de ADN en formato FASTA.

3.3 Analizar el Algoritmo DIMASP

Después de consultar el estado del arte con el objetivo de encontrar una técnica para extraer patrones de grandes cantidades de información, se observa que el algoritmo DIMASP, es una técnica adecuada que extrae los Patrones Frecuentes Maximales. En este paso se analizó el artículo titulado: A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection (García et al. 2006) donde se expone el algoritmo DIMASP. El objetivo fue conocer las características del documento de datos de entrada que almacena la colección de secuencias y conocer el procedimiento del algoritmo. Para entender mejor el proceso se ejecuta el algoritmo utilizando un programa informático desarrollado en C++ por los creadores del algoritmo. Para los datos de entrada, se utilizó una colección de secuencias de palabras, relacionadas a noticieros, conocida como Reuters-21578. Los pasos del algoritmo se describen en el apartado cinco de esta sección.

3.4 Reestructurar los documentos que contienen las Secuencias de ADN

La estructura de las bases de datos con la información de las secuencias de ADN, es diferente a la estructura de la base de datos utilizada por los desarrolladores del algoritmo, por lo tanto se creó una función que se desarrolló en NetBeans IDE versión 8.1, implementada en el lenguaje de programación Java, para reestructurar los documentos que contienen las

secuencias de ADN. Los datos de entrada del algoritmo DIMASP, tienen los siguientes requerimientos: se integra en un solo archivo la colección de secuencias de ADN que se va analizar; el formato del documento debe estar en texto plano; cada secuencia de las cadenas de ADN se almacena en una sola línea; cada cadena tiene que identificarse con un número entero, seguido del signo igual, a continuación los nucleótidos representados por los caracteres A, C, G, T, que deben estar separados por una coma; el último nucleótido de cada secuencia de ADN no lleva coma y todo es sin espacios.

De los documentos en formato FASTA de cadenas de longitud considerable, se genera un archivo de entrada con sus longitudes correspondientes (Ver figura 9).

3.5 Diseñar el mecanismo para extraer las SFM

Se diseñó el mecanismo integrando los elementos computacionales. Para su representación se utilizó el software de Microsoft Word y Paint. El resultado se muestra en la figura 10.

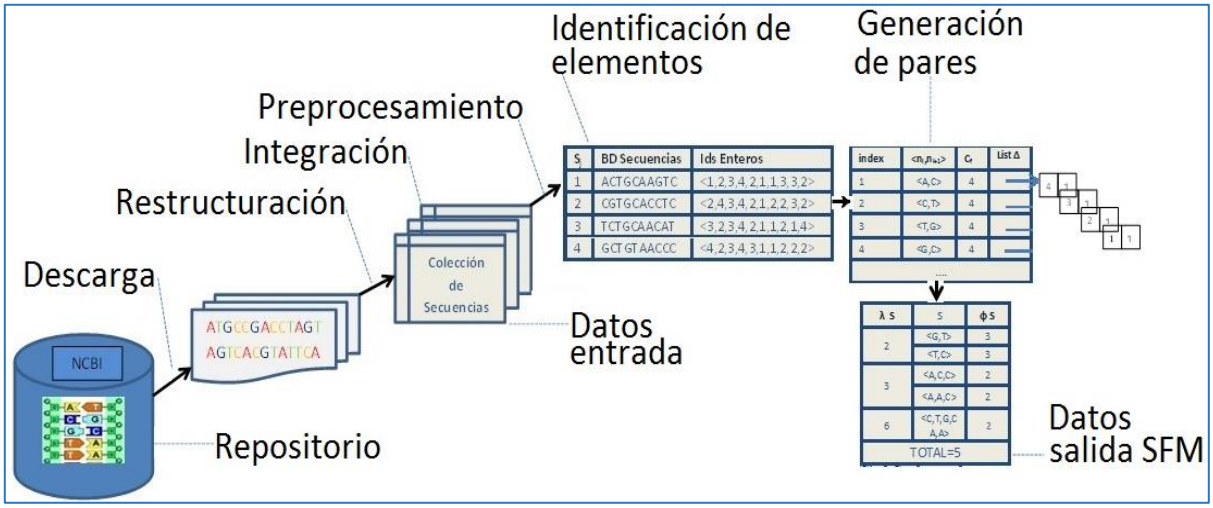

Figura 10. Diseño del Mecanismo computacional para la Extracción de las Secuencias Frecuentes Maximales de las cadenas de ADN.[11]

[11] Elaboración propia.

Las acciones de este mecanismo se describen a continuación de acuerdo al orden de sus elementos.

Repositorio: se accedió a la plataforma de NCBI, se descargaron de la base de datos GenBank, las secuencias de ADN en formato FASTA; para los experimentos se descargaron tres secuencias de diferentes organismos.

Reestructuración e Integración: luego se normalizaron los datos con la función que se creó para reestructurar los documentos que contienen las secuencias a analizar; en el apartado 3.4 se describió los requerimientos de los datos de entrada para el algoritmo. La función ejecuta un proceso con el que se integró en un único archivo la colección de secuencias de ADN, este documento es la entrada de datos del algoritmo.

Preprocesamiento: Esta acción se hace en 2 pasos.

Paso 1. En el primer paso se asigna un número natural 'N' como identificador, que representa cada uno de los elementos diferentes en la colección de secuencias. También, la frecuencia para cada identificador se almacena, es decir el número de secuencias donde aparece. Estos identificadores son usados en el algoritmo en lugar de los caracteres de la colección de secuencias. La Tabla 1, muestra un ejemplo de los identificadores.

Sj	Secuencias BD	Id N
1	A,C,T,G,C,A,A,G,T,C	<1,2,3,4,2,1,1,3,3,2>
2	C,G,T,G,C,A,C,C,T,C	<2,4,3,4,2,1,2,2,3,2>
3	T,C,T,G,C,A,A,C,A,T	<3,2,3,4,2,1,1,2,1,4>
4	G,C,T,G,T,A,A,C,C,C	<4,2,3,4,3,1,1,2,2,2>

Tabla 1. Ejemplo de una colección de secuencias y la representación de sus identificadores[12]

[12] Elaboración propia.

Paso 2. En el segundo paso, DIMASP construye una estructura de datos de la colección de secuencias almacenando todos los pares de caracteres contiguos $< n_i, n_{i+1}; n_{i+1}, n_{i+2;\ldots} n_{i+m} >$ que aparecen en un documento y también información adicional para preservar el orden secuencial. La estructura de datos es un array especial el cual contiene un identificador de cada par contiguo de los elementos. En cada celda se almacena un par de nucleótidos $C = < n_i, n_{i+1} >, < n_{i+1}, n_{i+2} >$, la frecuencia del par (C_f) también se almacena como referente para el umbral β, una marca booleana para agregar el nodo a la lista Δ de nodos δ, donde un nodo δ almacena un identificador del documento (δ.Id), un índex (δ.Index) de la celda donde el par aparece en el array, un link (δ.NextDoc) para mantener la lista Δ y un link (δ.NextNode) para preservar el orden de la secuencia de los pares con respecto al documento. Ver figura 1.

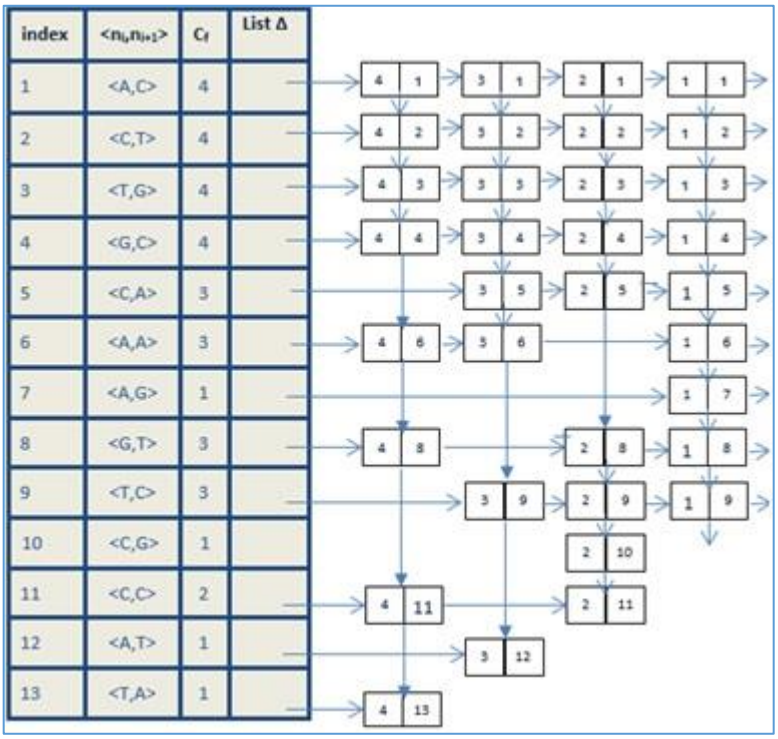

Figura 11. Estructura de datos del paso 2, construida a partir de la colección de secuencias de la tabla 1.[13]

[13] Elaboración propia.

Datos de Salida: En el último paso se extraen las SFM, el algoritmo DIMASP encuentra todas las Secuencias Frecuentes Maximales, usando la estructura del paso 2, extrae todas las secuencias β frecuentes, beta es el umbral requerido por el usuario. Con esto el algoritmo encuentra Posibles Secuencias Frecuentes Maximales (PSFM) tomando los primeros pares como inicio de la secuencia. Mientras que la secuencia del siguiente par sea mayor o igual a β sigue creciendo el patrón y lo agrega al conjunto de SFM, si no se encuentra una secuencia equivalente almacenada o si no es subsecuencia de ninguna otra SFM. Ver Tabla 2. Esta muestra las SFM extraídas de la colección de secuencias presentadas.

λS	*SFM*	C_f
2	< G,T >	3
	< T,C >	3
3	< A,C,C >	2
	< A,A,C >	2
6	< C,T,G,C,A,A >	2
TOTAL = 5		

Tabla 2. Ejemplo Extracción de las SFM[14]

La Tabla 2. Representa la extracción de SFM resultantes del algoritmo DIMASP. La columna representada por lambda y la letra s, '*λS'* muestra la longitud de las SFM extraídas, tenemos que extrajo secuencias de longitud 2, 3, y 6. El encabezado de la columna *SFM* almacena cada una de las SFM. Y la columna representada por C_f muestra la frecuencia, es decir; en cuántas secuencias se encuentra la SFM; la tabla refiere que las SFM de $\lambda = 2$ aparece en 3 de las 4 secuencias de la colección, $\lambda = 3$ en 2 secuencias, $\lambda = 6$ también en 2 secuencias. En este caso el umbral requerido fue de 2 y extrajo las secuencias con frecuencia 2

[14] Elaboración propia.

y 3, con esto se confirma que el algoritmo es independiente del umbral de soporte, es decir que si descubre SFM que tienen una frecuencia mayor al umbral también las extrae.

3.6 Prueba del mecanismo con secuencias propias de ADN

Las pruebas del mecanismo se desarrollan ampliamente en el capítulo 4. Para su valoración, se ejecutaron varios experimentos con secuencias de diferentes organismos, longitudes y umbrales;

3.7 Análisis de resultados

En el Capítulo 4 se muestran los resultados de cada experimento bioinformático. La información que se obtuvo de los resultados se muestra mediante tablas para su visualización.

CAPÍTULO 4.

EXPERIMENTOS Y ANÁLISIS DE RESULTADOS

4.1 Experimentos Bioinformáticos

Experimento 1

Para la primera prueba se descargó la secuencia identificada con los siguientes datos: >NC_000002.12 Homo sapiens chromosome 2, GRCh38.p7 Primary Assembly. Luego de forma manual se extrajeron los primeros 10,000 nucleótidos y de manera sintética se generó un documento con la colección de 15 secuencias de ADN de diferentes longitudes. Con esta misma colección se ejecutaron 3 corridas del algoritmo con diferente umbral β en cada ejecución. La primera se requirió un umbral de β=10, en la segunda un umbral de β=6 y en la tercera un umbral de β=2. A continuación se muestra en las tablas los resultados de las SFM extraídas. La información que se obtiene es la longitud de las Maximales, cuántas secuencias diferentes de esa misma longitud se obtuvieron y el total de SFM en toda la colección.

Extracción de Secuencias Frecuentes Maximales. β 10				
Longitud / Cantidad de SFM diferentes				
24/2	26/1	30/1	44/1	45/1
48/1	49/1	50/1	52/2	54/2
57/2	60/1	61/1	62/1	64/1
69/1	72/1	74/2	84/1	92/1
93/1	99/1	104/1	105/1	114/1
122/1	123/1	124/1	125/1	155/1
163/1	164/1	201/1	204/1	208/1
210/1	246/1	294/1	296/1	299/1
328/1	346/1	401/1	415/1	421/1
423/1	430/1	431/1	449/1	479/1
551/1	722/1	1075/1	1238/1	1707/1
TOTAL = 60 SFM				

Tabla 3. Resultados Experimento 1, Extracción SFM, β10[15]

[15] Elaboración propia.

Para verificar que los resultados fueron correctos se hicieron 3 corridas del algoritmo con los mismos datos de entrada y el mismo valor del umbral β del experimento 1, se obtuvieron los mismos resultados en cada una de las corridas, por lo tanto se considera un algoritmo determinista. También se utilizó el editor de texto Sublime, para comprobar los umbrales y la cantidad de SFM, en este caso sí coincidieron los umbrales a excepción de una SFM que aparecía una vez más que el umbral registrado, esto debido a que el algoritmo tiene la restricción de que si encuentra más de una vez una SFM en una misma secuencia solo la cuenta una vez.

Extracción de Secuencias Frecuentes Maximales. β 6				
Longitud / Cantidad de SFM diferentes				
5/1	87/1	88/1	102/1	111/1
119/1	132/2	174/1	228/1	237/1
305/1	324/1	341/1	353/1	416/1
427/1	451/1	467/1	493/1	518/1
524/1	1086/1	1456/1	1610/1	1728/1
1832/1				
TOTAL = 27 SFM				

Tabla 4. Resultados Experimento 1, Extracción SFM, β6[16]

Extracción de Secuencias Frecuentes Maximales. β 2				
Longitud / Cantidad de SFM diferentes				
5/1	6/13	7/41	8/34	9/18
10/12	11/6	12/4	13/3	14/6
15/6	16/1	17/5	18/3	19/1
20/1	21/1	22/1	25/1	26/1
27/2	31/1	33/1	38/1	42/1
43/1	54/1	62/1	64/1	68/1
73/1	79/1	82/1	89/1	105/1
133/1	141/1	353/1	769/1	4973/1
5733/1				
TOTAL = 181 SFM				

Tabla 5. Resultados Experimento 1, Extracción SFM, β2[17]

[16] Elaboración propia.
[17] Elaboración propia.

En la Tabla 6, se muestra una relación de los valores de las SFM extraídas de acuerdo a los datos que representan las tablas 3, 4 y 5 que se obtuvieron con diferentes umbrales. Se analizaron las Secuencias Frecuentes Maximales de manera global de acuerdo a su longitud mínima y máxima de cada umbral y la cantidad total de SFM extraídas.

Umbral	Longitud Mínima	Longitud Máxima	Total SFM
β2	5	5733	181
β6	5	1832	27
β10	24	1707	60

Tabla 6. Comparación con diferentes umbrales[18]

En este caso no se puede afirmar un comportamiento constante de la información extraída por que las secuencias utilizadas en este experimento 1 son sintéticas y la colección de secuencias es muy pequeña, no obstante se observa que la longitud mínima es menor para el umbral menor, y la longitud máxima es mayor también para el umbral menor, de igual manera con el umbral más pequeño se obtuvo el mayor número de SFM. Ver la relación de la Tabla 7. Cabe aclarar que no se trata de dar una interpretación a los resultados ya que en ningún momento se planteó esto como un objetivo, ni tampoco es parte de la solución a la problemática. Sin embargo puede ser relevante para otra problemática.

Umbral	β2	<	β6	<	β10
Longitud Mínima	5	=	5	<	24
Longitud Máxima	5733	>	1832	>	1707
SFM	181	>	27	<	60

Tabla 7. Relación de los valores de las SFM.[19]

La Tabla 7, muestra que no hay una dependencia directa o proporcional entre el umbral, longitud y cantidad de las SFM extraídas con respecto a esta colección de secuencias de ADN.

Experimento 2

[18] Elaboración propia.
[19] Elaboración propia.

En este apartado se presenta un experimento que tiene la misma finalidad de extraer las SFM, comunes de 3 cadenas de ADN, con los siguientes datos que las identifican: de la especie *Homo Sapiens* el cromosoma 21. De *Caenorhabditis elegans* el cromosoma II. De *Drosophila melanogaster* el cromosoma X. En cada caso se tomaron los primeros 213,000 nucleótidos. De acuerdo al mecanismo, en la reestructuración se integran las 3 cadenas en un solo documento, que es la entrada del algoritmo DIMASP. Luego se ejecuta el algoritmo para obtener las SFM.

En la Tabla 8, se muestra el resultado de las SFM extraídas de las 3 cadenas mencionadas, con un umbral igual a β2.

Longitud de las SFM	SFM Diferentes
7	147
8	4582
9	19298
10	23342
11	14629
12	6490
13	2535
14	892
15	286
16	120
17	45
18	23
19	7
20	1
21	4
23	2
27	1
β2	**71567**
β3	**1437**
Total SFM	**73004**

Tabla 8. Resultados de la Extracción de SFM de 3 cadenas, con umbral β2[20]

[20] Elaboración propia.

Una de las características que tiene el algoritmo es que es independiente del umbral de soporte. En la Tabla 8, se muestra la cantidad de SFM, correspondientes al umbral β2, y también extrae las secuencias que encuentre por arriba del umbral especificado, en este caso se obtuvieron secuencias con umbral de β3.

En la Tabla 9, se presentan los resultados de la misma colección de las 3 secuencias mencionadas antes. En este caso se especificó un umbral de 3.

Longitud de las SFM	SFM Diferentes
6	3
7	920
8	8466
9	12736
10	6312
11	1983
12	543
13	101
14	24
15	8
16	3
17	1
Total SFM	**31100**

Tabla 9. Resultados de la Extracción de SFM de 3 cadenas, con umbral β3[21]

[21] Elaboración propia.

Experimento 3

El tercer experimento se hizo con las mismas especies del experimento 2, aumentando la longitud de cada una de las secuencias a 419,930 nucleótidos. Se especificó un umbral de β2 para la primera corrida y un umbral de β3 para la segunda. Los resultados se muestran en las Tablas 10 y 11.

Longitud de las SFM	SFM Diferentes
7	3
8	2164
9	23342
10	50532
11	42422
12	22914
13	9943
14	3763
15	1284
16	535
17	178
18	82
19	26
20	17
21	8
22	4
23	2
24	7
25	3
26	3
27	1
37	1
β2	**154221**
β3	**3113**
Total SFM	**157334**

Tabla 10.Resultados del Experimento 3, con umbral β2 [22]

[22] Elaboración propia.

La Tabla 11. Muestra los resultados con las mismas especies y una longitud de cada una de las secuencias de 419,930 nucleótidos, con un umbral de β3.

Longitud de las SFM	SFM Diferentes
7	191
8	6677
9	26435
10	22000
11	9045
12	2681
13	658
14	182
15	54
16	13
17	10
19	1
20	1
22	1
Total SFM	**67949**

Tabla 11. Resultados del Experimento 3, con umbral β3 [23]

[23] Elaboración propia.

Experimento 4

El cuarto experimento se hizo con las mismas 3 especies del experimento 2 y 3. En esta prueba se aumentó la longitud de las secuencias a 699,930 nucleótidos y con umbrales de β2 y β3. Los resultados con un umbral de β2 se muestran en la Tabla 12.

Longitud de las SFM	SFM Diferentes
8	567
9	19833
10	76681
11	84450
12	51383
13	23547
14	9173
15	3419
16	1290
14	465
18	193
19	76
20	37
21	15
22	14
23	9
24	14
25	3
26	8
27	2
28	2
29	2
30	2
31	2
32	1
33	1
35	1
36	5
37	1
58	1
β2	**265842**
β3	**5355**
Total SFM	**271297**

Tabla 12. Resultados del Experimento 4, con umbral β2 [24]

[24] Elaboración propia.

En la Tabla 13, se muestran los resultados de la extracción de las SFM con un umbral β3.

Longitud de las SFM	SFM Diferentes
7	15
8	3720
9	33904
10	46094
11	22686
12	7230
13	1962
14	482
15	140
16	34
17	16
18	9
19	3
20	1
21	1
22	2
23	1
24	2
25	1
29	1
31	2
32	1
35	1
39	1
Total SFM	**116309**

Tabla 13. Resultados del Experimento 4, con un umbral β3 [25]

[25] Elaboración propia.

CONCLUSIONES

La presente tesis tuvo como objetivo desarrollar una propuesta de un mecanismo integrando elementos computacionales, para extraer Secuencias Frecuentes Maximales, basándose en el algoritmo DIMASP, el objeto de estudio es el análisis de cadenas de ADN. Para conocer cómo funciona el algoritmo se ejecutó con una colección de secuencias de datos tipo texto, referentes a un noticiero, esto permitió conocer la estructura del documento de entrada, los procesos y las salidas del algoritmo, para adaptar el documento con la colección de secuencias de ADN. Con la información que se obtuvo se integraron los elementos para el mecanismo.

Se diseñó el mecanismo de manera gráfica, el cual se puede apreciar en la figura 10, donde se muestran los elementos de acuerdo al orden en que se hace el proceso, para la obtención de los patrones. Para obtener los datos de las cadenas de ADN, se investigó qué instituciones se dedican al análisis de las secuencias de ADN y cuáles son los formatos de los documentos que almacenan las secuencias, después se decidió obtener los datos de la plataforma NCBI, en el formato FASTA, por ser una plataforma amigable y un formato más compatible con las entradas de datos del algoritmo DIMASP. No obstante, se tuvo la necesidad de reestructurar los documentos con una función programada en java, para que pudieran procesarse los datos con el algoritmo. Con las secuencias individuales se generó un documento de entrada que contiene la colección de secuencias a analizar. En esta etapa del proyecto, no es posible analizar genes completos por lo que en los experimentos 2, 3, y 4, se consideraron secuencias de longitudes entre los 213,000 y 699,930 nucleótidos. Para la implementación del mecanismo se hicieron 4 experimentos de secuencias de ADN con diferentes especies y como resultado se

obtuvieron las Secuencias Frecuentes Maximales de acuerdo al umbral requerido en cada experimento.

El mecanismo computacional propuesto para la extracción de SFM de las secuencias de ADN, da resultados satisfactorios. El algoritmo DIMASP, que se usa en minería de texto funciona eficazmente con minería de secuencias de nucleótidos. Parte importante es la reestructuración de los documentos originales de nucleótidos, para que pueda funcionar el algoritmo.

La aportación de este trabajo en el área computacional es la compresión de la información, reduciendo el espacio de almacenamiento y el espacio de búsqueda. Las SFM son una representación de todas las secuencias frecuentes implícitas en los datos. Esto facilitará el trabajo de investigación genómica y se avanzará en el desarrollo de fármacos enfocados a la medicina personalizada.

TRABAJO A FUTURO

El trabajo futuro de esta investigación se centra en la indexación de las SFM, que facilitaran las consultas, búsquedas y comparaciones. Con la indexación se obtendrá información como: la secuencia específica a la que pertenece la SFM; la posición en la secuencia; dónde empieza la SFM y dónde termina; la longitud de la secuencia; en qué otra secuencia está la SFM. Una ventaja del algoritmo es que si se quiere agregar una nueva secuencia a la colección de secuencias, no es necesario volver a ejecutar desde cero todas las secuencias, solo procesa los datos de la nueva secuencia y agrega la información necesaria. Queda pendiente analizar cómo se hace el anterior proceso, para hacer las adaptaciones correspondientes.

Para resolver la problemática del análisis de múltiples secuencias completas, es necesario la adaptación del algoritmo para la paralelización de los procesos. Contar con una infraestructura robusta, de hardware y de software, por la gran cantidad de información que se maneja. Se requiere hacer uso de tecnologías como sistemas distribuidos y tecnologías de paralelización.

Para la implementación del sistema distribuido es necesario la creación de un cluster de computadoras con elementos de Hardware, software, y protocolos, para la comunicación y coordinación entre los elementos. Se investigará sobre las tecnologías existentes para redes de computadoras y sobre protocolos que sean adecuados para los procesos de paso de mensajes.

Se harán las adecuaciones necesarias al algoritmo para la paralelización de los procesos. Se investigará sobre estrategias de paralelización, para elegir e implementar la más conveniente en relación al algoritmo DIMASP.

REFERENCIAS

Agrawal, R., & Srikant, R. (1995). Mining Sequential Patterns. *In Proc. of the 11th Int'l Conference on Data Engineering* (págs. 3-17). TaipeiTaiwan: IBM Almaden Research Center.

Alonso, A. e. (2005). Challenges of DNA Profiling in Mass Disaster Investigations. *Croat Med J*, *46*(4), 540-548. Recuperado el 20 de Octubre de 2016

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.*, 403-410.

Bergroth, L., Hakonen, H., & Raita , T. (2002). A Survey of Longest Common Subsequence Algorithms. *IEEE*, 39-48. doi:10.1109/SPIRE.2000.878178

Cañedo Andalia, R., & Arencibia Jorge, R. (2004). Bioinformática: en busca de los secretos moleculares de la vida. *Acimed, 12*(6). Recuperado el 9 de Septiembre de 2016, de http://bvs.sld.cu/revistas/aci/vol12_6_04/aci02604.htm

Dao-I, L. (1998). Fast Algorithms for Discovering the Maximum Frequent Set. *New York University,*. Recuperado el 5 de Noviembre de 2017

Edgar C, R. (1 de March de 2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*, 1792-

1797. Obtenido de https://academic.oup.com/nar/article/32/5/1792/2380623

Escobar, M., Murillo, R., & Soto, F. (2011). Tecnologías bioinformáticas para el análisis de secuencias de ADN. *Scientia Et Technica*, 116-121.

Fogglino, M. A. (2009). *Identificación biológica de repeticiones en secuencias de ADN del genoma humano.* Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires Argentina, Computación. Recuperado el 10 de Septiembre de 2016

García Hernández, R. A., Martínez Trinidad, J. F., & Carrasco Ochoa, J. A. (2006). A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. *Intenational Conference on Computational Linguistics and Text Processing* (págs. 514-523). Mexico: CICLing.

Ginsburg , G., & McCarthy, J. (1 de December de 2001). Personalized medicine revolutionizing drug discovery and patient care. *TRENDS in Biotechnology, Vol.19*(12), 491- 496. doi:https://doi.org/10.1016/S0167-7799(01)01814-5

Green, E. D. (21 de Octubre de 2015). *National Human Genome Research Institute*. Recuperado el 9 de Septiembre de 2016, de https://www.genome.gov/27562614/cido-desoxirribonucleico-adn/

Humphreys, B. L. (9 de Febrero de 2016). *National Library of Medicine*. Obtenido de https://www.nlm.nih.gov/about/2017CJ.html

Hutchinson, C. A. (2007). DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research, 35*(18), 6227-6237. doi:doi: 10.1093|nar|gkm688

Lesko, L. J. (2007). Personalized Medicine: Elusive Dream or Imminent Reality? *nature publishing group, 81*(6), 807-816. doi:DOI: 10.1038/sj.clpt.6100204

Liu, Q. (13 de Enero de 2000). ADN Computing on Surfaces. *Nature, 403*(13), 175-179. doi:http://dx.doi.org/10.1038/35003155

Mabroukeh, N. R., & Ezeife, C. I. (2010). A taxonomy of sequential pattern mining algorithms. ACM Comput. *ACM Computing Surveys, 43*(1), 41. doi:10.1145/1824795.1824798

Mabrouketh, N. R., & Ezeife, C. I. (Noviembre de 2010). A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys, 43*(1), 41 pages. doi:DOI=10.1145/1824795.1824798

Macip, S. (14 de Abril de 2010). *elmundo.es.* Obtenido de Salvador Macip. 2010. http://www.elmundo.es/elmundosalud/2010/04/14/investigacion/1271230336.html consultado 2016

Mckenna, A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 1297-1303. doi:http://doi.org/10.1101/gr.107524.110

Meneses Escobar, C. A., Rozo Murillo, L. V., & Franco Soto, J. (Diciembre de 2011). Tecnologías Bioinformáticas para el Análisis de Secuencias de ADN. *Scientia et Technica*(49), 116-120. Recuperado el 2017

Necochea Campion, R., & Canul Tec, J. C. (Junio de 2004). *Instituto de Biotecnología-UNAM.* Obtenido de http://www.ibt.unam.mx/computo/pdfs/met/secuenciacion_acidos_nucleicos.pdf

NIH. (21 de Octubre de 2015). *National Human Genome Research Institute*. Recuperado el 9 de Septiembre de 2016, de https://www.genome.gov/27562614/cido-desoxirribonucleico-adn/#al-3

Pei, J., Han, J., Mortazavi, B., & Pinto, H. (2001). PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. *IEEE*, 1063-6382.

Rucci, E. (Abril de 2013). *SEDICI*. doi: http://hdl.handle.net/10915/27737

Shendure, J., & Ji, H. (Octubre de 2008). Next-generation DNA sequencing. *Nature Biotechnology, 26*(10), 1135-1145. doi:doi:10.1038/nbt1486

Toledo, C., & Saltsman, K. (11 de June de 2012). *National Institute of General Medical Sciences*. Recuperado el 5 de Diciembre de 2016, de https://publications.nigms.nih.gov/insidelifescience/genetics-numbers.html

Witten, I. H., & Frank, E. (2005). *Data Mining. Practical Machine Learning Tool s and Techniques.* San Francisco, CA: Elsevier.

Zaki, M. J. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Kluwer Academic Publishers*, 31-61.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS