CARACTERIZACIÓN Y ANÁLISIS BIOINFORMÁTICO DE REGIONES PROMOTORAS DE *Mycobacterium tuberculosis*

Claudia Milena Rivera Trujillo

Universidad El Bosque

Facultad de Ciencias

Bogotá D.C, 2013

CARACTERIZACIÓN Y ANÁLISIS BIOINFORMÁTICO DE REGIONES PROMOTORAS DE *Mycobacterium tuberculosis*

Claudia Milena Rivera Trujillo

Director: Juan Manuel Anzola, Ph.D.

Corpogen

Corporación para la investigación y biotecnología

Universidad El Bosque

Facultad de Ciencias

Bogotá D.C, 2013

AGRADECIMIENTOS

Me gustaría agradecer a mi familia por apoyarme, a Rafa y José, bioinformáticos de CorpoGen, que me demostraron el arte de la programación y de manera muy especial a mi director, quien fue muy generoso y paciente conmigo. Me siento muy afortunada de haber realizado mi pasantía con él.

NOTA DE SALVEDAD INSTITUCIONAL

La Universidad El Bosque no se hace responsable de los conceptos emitidos por los investigadores en su trabajo, solo velará por el rigor científico, metodológico y ético del mismo en aras de la búsqueda de la verdad y la justicia.

Contenido

Lista de figuras

Lista de Tablas

Resumen

Este trabajo realizó la caracterización y el análisis bioinformático de las regiones promotoras en el agente causal de la tuberculosis en humanos, *Mycobacterium tuberculosis.* El éxito de *M. tuberculosis* como patógeno recae en su habilidad de adaptarse a distintas condiciones dentro del hospedero. Estas adaptaciones dependen en la coordinación de la expresión génica a través de la regulación de la transcripción.

Para entender mejor los promotores micobacterianos, se realizaron análisis estadísticos de la frecuencia de octámeros de DNA presentes en regiones promotoras de genes individuales y genes organizados en operones de las trece cepas de *M. tuberculosis*. Este tipo de análisis se basa en dos hechos, el primero tiene relación con el promedio de nucleótidos presentes por giro de DNA en el surco mayor es ocho nucleótidos y el segundo hecho con que las palabras, o motivos en el DNA de los organismos no siguen un patrón al azar, por el contrario, son conservadas debido a que el proceso evolutivo conserva las partes funcionales y las reutiliza para generar nuevas funciones.

El conocer los sitios unión a factores de transcripción en *M. tuberculosis* puede ayudar a mejorar los métodos para combatir la tuberculosis a través del desarrollo de drogas diseñada para inhibir la expresión de determinados genes clave como son los genes factores de virulencia y variabilidad antigénica.

Palabras clave: Promotores, bioinformática, *Mycobacterium tuberculosis,* frecuencia.

Abstract

This study made a bioinformatics characterization and analysis of promoter regions in the responsible agent of tuberculosis in the human beings *Mycobacterium tuberculosis.* The succeed of *M. tuberculosis* as a pathogen is due to its skills to adapt in different conditions within hospedge. These adaptations depend on the coordination in the genetic expression through regulation in the transcriptional regulation.

To understand better mycobacterial promoters, statistical analyses were performed in the octamer frequency of DNA that appears in promoters regions of individual gene and organized operons genes by thirteen strains of *M. tuberculosis*. This type of analysis is based on two facts; the first one has relation with the average of nucleotides that appear in DNA per spin in the main groove that is eight nucleotides and the second fact is that with words or patterns in the DNA of organisms do not follow a random pattern, On the contrary, they are conserved because of the evolutionary process maintains the functional parts and reuse it, to generate new functions.

Knowing the binding transcriptional factors in *M. tuberculosis* could help to improve the methods for fighting the tuberculosis through the development of drugs designed to inhibit the expression of certain key genes such as virulence genes and antigenic variability.

Key words: Promoters, bioinformatics, My*cobacterium tuberculosis*, frequency

1. Introducción

Uno de los conceptos más importantes en la biología molecular es el dogma central. El proceso mediante el cual el ADN es duplicado, transcrito en mRNA y este a su vez es traducido en proteína. En nuestro afán de entender el cómo funciona este proceso se desató una carrera para determinar la funcionalidad del ADN y cómo la información biológica está almacenada en la cadena de ADN.

La secuenciación genómica tuvo sus orígenes en 1972 cuando Walter Fiers y su equipo de la Universidad de Gante lograron secuenciar el gen de un bacteriófago MS2. Hoy en día, gracias a las nuevas tecnologías de secuencia disponemos de más de 6000 genomas secuenciados o en proceso de ser secuenciados (Lagasen *et al*.,2010). La información generada por los proyectos genómicos crece a un ritmo muy acelerado, y hoy en día se están obteniendo muchos más datos de los que se puede analizar. Es así que a medida que el volumen de datos genómicos crece, las herramientas informáticas son indispensables para manipular esta vasta información (Mount, 2004).

Al analizar distintos genomas, se ha podido encontrar que la complejidad organísmica no se correlaciona bien con el número de genes o el tamaño del genoma, sino con la forma en que el conjunto de genes de un organismo interactúan entre sí, acoplado a su activación/desactivación en diferentes condiciones celulares (Akatsu,1998).

En este orden de ideas, la regulación de genes es una importante fuerza motriz en la evolución de las especies y sus genomas, probablemente el factor más importante en la complejidad de los organismos (Jordan *et al*.,2005).

Es menester aclarar que la fuerza motriz, es decir la regulación genética, es un concepto que está intrínsecamente asociado con factores de transcripción. Estas son proteínas que

se unen a las regiones promotoras de los genes y permiten que el complejo de la ARN polimerasa inicie la transcripción de los mismos.

Son estas regiones promotoras las que determinan si un gen debe estar siendo expresado en determinadas condiciones ambientales, o de desarrollo, o de algún otro proceso biológico como virulencia, defensa, reproducción, división celular, morfogénesis, etc. Se propone un estudio detallado de regiones reguladoras, particularmente regiones promotoras en *M. tuberculosis*, el agente causal de la tuberculosis en humanos que anualmente provoca nueve millones de nuevos casos de tuberculosis (TB) y dos millones de muertes. Aunque su diagnostico, la quimioterapia y la vacuna están disponibles, la enfermedad está aún lejos de ser erradicada (Kaufman *et al.*, 2011).

Por este motivo, entender la micobacteria a nivel genético permitirá formular mejores mecanismos de diagnóstico, de prevención y terapéuticos para combatir la enfermedad, que en este momento es prevalente en la tercera parte de la población a nivel mundial.

Para entender mejor los promotores micobacterianos, se realizaron análisis estadísticos de la frecuencia de octámeros de ADN presentes en regiones promotoras de genes individuales y genes organizados en operones de *M. tuberculosis*. Este tipo de análisis se basa en dos hechos, el primero tiene relación con el promedio de nucleótidos presentes por giro de DNA en el surco mayor es ocho nucleótidos y segundo hecho con las palabras o motivos en el DNA de los organismos no siguen un patrón al azar, por el contrario, son conservadas debido a que el proceso evolutivo conserva las partes funcionales y las reutiliza para generar nuevas funciones (Jordan *et al*.,2002).

Como resultado de esto, el estudio de la distribución de palabras de ADN en regiones promotoras de los genes nos puede dar luces sobre elementos reguladores que hayan

sido conservados funcionalmente y nos puede permitir entender mejor la biología de la *M. tuberculosis.*

1.2. Justificación

El éxito de *M. tuberculosis* como patógeno recae en su habilidad de adaptarse a distintas condiciones dentro del hospedero. Estas adaptaciones dependen en la coordinación de la expresión génica a través de la regulación de la transcripción (Corbett *et al*.,2003).

En *M. tuberculosis* esta propiedad se consigue mediante la acción colectiva de 190 reguladores transcripcionales presentes en el genoma (Cole *et al*.,1998). La importancia de estos reguladores y su relación con la patogenicidad se encuentra reportada por varias observaciones de activación/desactivación de genes que son regulados por los factores sigma causando severas alteraciones in vivo (Ando *et. al*.,2003).

Por lo tanto, entender los mecanismos que regulan la expresión génica e identificar los elementos reguladores los cuales permiten la expresión, es tal vez el mayor desafío en la biología molecular (Xiong,2006). Recientes avances en la secuenciación genómica ha permitido contar con 1500 genomas procariotas completamente secuenciados (Münch *et al*.,2011), lo cual exige métodos computacionales eficientes y exactos para la identificación, anotación y tabulación de las regiones codificantes y de los elementos funcionales no codificantes como los promotores (Xiong,2006).

La identificación de la localización y función de estas regiones promotoras es compleja, debido a que estas regiones no están claramente definidas y son altamente diversas. (Das &Dai, 2007).

Las regiones promotoras pueden integrar varios signos de control de la tasas de transcripción, de los genes que regulan, como las etapas de desarrollo respuestas hormonales y señales fisiológicas y ambientales.

En este sentido, el estudio de la distribución las regiones promotoras de los genes permitiría entender mejor la biología de la *M. tuberculosis*.

1.3. Antecedentes

La tuberculosis ha estado presente a lo largo de la historia de la humanidad desde mucho tiempo, se han reportado registros de bacilos de la edad de piedra y época egipcia (Manterola, 2004). Sin embargo esta enfermedad no representaba un problema significativo sino hasta el siglo XVIII y el siglo XIX, cuando presento una mayor morbilidad, debido a las malas condiciones de trabajo y de vivienda que ayudaron a la proliferación de la infección (Farga,1999).

Fue hasta el año 1865 cuando el científico francés Villemin logró demostrar que esta enfermedad era transmisible y no hereditaria como se creía hasta ese momento. Años más tarde y conociendo el trabajo de Villemin, el bacteriólogo Robert Koch, en 1882, fue capaz de aislar y cultivar la bacteria, creando las bases del diagnostico mico-bacteriológico (Manterola, 2004).

En 1943 se descubrió el primer fármaco para combatir la tuberculosis, la isoniacida, después siguió el descubrimiento de la rifampicina en 1960 y la pirazinamida a finales de la década de 1970, este último permitió acortar la duración del tratamiento a seis meses (Farga,1999).

Como consecuencia del mal uso de las quimioterapias, ha surgido cepas inmunes a los medicamentos convencionales, estas cepas resistentes pueden dividirse en dos grupos, las cepas multirresistentes (MDR) y las extremadamente resistentes (XDR) afectando el control de la tuberculosis a nivel mundial (Sáenz, 2007).

En la búsqueda de diagnosticar con mayor rapidez la enfermedad se ha logrado un avance en el diagnostico de mico-bacterias, como el empleo de sistemas automatizados de cultivo en medio liquido, y el uso de la amplificación génica.

La determinación de la secuencia de los genomas de *M. tuberculosis* H37Rv (Cole *et al*. 1998), de *M. leprae* (Cole *et al.*, 2001) y de *M. bovis* (Gordon e*t al.*,2001) han hecho una contribución importante para el estudio de la virulencia, patogenicidad y resistencia a antibióticos de las micobacterias.

Por otra parte en el 2005 se creó una base de datos para el análisis de la regulación transcripcionales en *M. tuberculosis*, llamada MtbRegList, la cual contiene motivos reguladores predichos y caracterizados, además de reportes de sitios de inicio de transcripción identificados experimentalmente (Jacques *et al.*,2005).

1.4. Objetivos

Objetivo General

Analizar bioinformáticamente las regiones promotoras de *M. tuberculosis*.

Objetivos Específicos

• Analizar mediante métodos estadísticos de sobre-representación de octámeros en las regiones promotoras en *M. tuberculosis* al descomponer las regiones promotoras en octámeros palabras de ocho letras y ventanas móviles de un nucleótido.

•Comparar los resultados de estos análisis estadísticos en diferentes genomas de micobacterias y determinar la asociación de palabras sobre-representadas con respecto a regiones promotoras de función conocida en *M. tuberculosis*.

•Generar un nuevo conjunto de datos de posibles octámeros que puedan estar funcionando como sitios de unión de factores de transcripción con base en los resultados anteriores.

1.5. Hipótesis

Existen diferencias significativas entre las frecuencias de octámeros en regiones promotoras versus regiones al azar del genoma.

2. Marco Teórico

Las células pueden ser consideradas como ambientes cerrados encapsulados por una membrana, donde las proteínas asumen uno o más funciones específicas. El desarrollo, integración, comunicación y sinergia de las células dentro de un sistema u organismo, dependen de controlar y administrar la producción y acumulación de estas macromoléculas, al mantenerlas dentro de unos límites de concentración adecuados (Kozak *et al*., 1999).

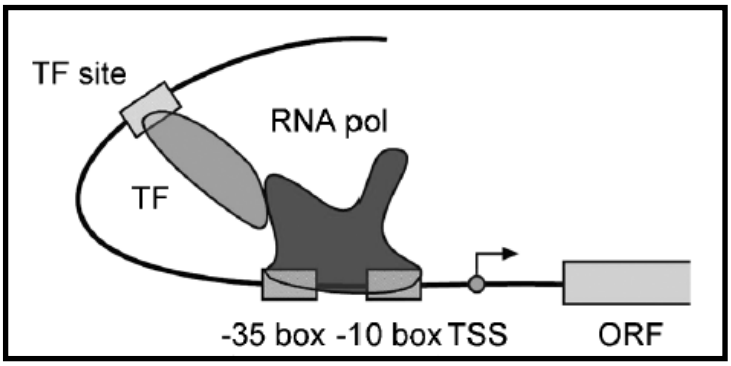
De acuerdo al dogma central de la biología, el cual postula el flujo genético en términos generales desde el ácidodesoxirribonucleico (DNA) a ácido ribonucleico (RNA) y este a proteínas, las instrucciones de la cantidad y sitio específico de producción están codificadas en el DNA (Walker, 2010).

La mayoría de los productos de las regiones codificantes del DNA están conservados entre las distintas especies de bacterias, es así como la increíble diversidad de estos organismos recae en las diferencias entre la cantidad relativa de los productos y de la coordinación temporal-espacial de estos (Sánchez *et al*., 2011).

En este sentido, las diferencias en la fuerza intrínseca de los promotores y la acción de los factores de transcripción, son elementos que inciden en la variación de los niveles de RNA, donde la transcripción inicial es probablemente el paso más frecuente para controlar la expresión (De Avila E Silva *et al*., 2011).

2.1 Promotores En Procariotas

En el proceso de transcripción se sintetiza el mRNA utilizando DNA como molde, que es polimerizado desde el nucleósido 5'-trifosfato y catalizado por la enzima RNA polimerasa (Helmann *et al*., 1999).



Tomado de Xiong,2006. Essential Bioinformatics

Figura 1. Representación esquemática de los distintos elementos que intervienen en el proceso inicial de la transcripción, RNA polimerasa (RNA pol); Sitio de inicio de transcripción (TSS); ORF, marco de lectura; pol, polimerasa; -35 box y -10 box, cajas -35 y -10 respectivamente; TF factores de transcripción y TF site, sitio de unión de factores de transcripción.

La transcripción empieza, como se muestra en la figura 1. cuando la RNA polimerasa, reconoce al promotor, una secuencia específica del DNA necesaria y suficiente para que la RNA polimerasa se una a esta región, la cual se encuentra localizada en la vecindad, corriente arriba, del sitio de inicio de la región codificante (Xiong,2006).

Sin embargo, para que la transcripción inicie se necesitan de factores de transcripción (Zhou & Yang, 2006). Los factores de transcripción, son proteínas que se adhieren a una secuencia de DNA específica para habilitar o inhabilitar la función de la RNA polimerasa, modulando los niveles de transcripción. Estas proteínas actúan aumentando

la unión o la actividad de la holoenzima, o pueden bien, disminuir la transcripción (López, 2001).

La mayoría de los factores de transcripción, poseen motivos proteicos independientes, es decir, el motivo que genera la interacción con la secuencia de DNA es independiente de aquel que activa o disminuye la transcripción, haciendo que el proceso de transcripción sea más complejo de lo que se pensó al principio (Newton & Gey, 2012).

Las interacciones entre los factores de transcripción y el segmento de DNA, se debe a las interacciones atómicas entre las proteínas y los nucleótidos, enlaces temporales de Van der Waals, puentes de hidrógeno e interacciones entre moléculas (López, 2001).

De hecho, en organismos procariotas la transcripción se inicia por la subunidad σ de RNA polimerasa, la cual es una proteína que reconoce la secuencia especifica corriente arriba de un gen que permite la unión del resto del complejo. La secuencia corriente arriba donde la proteína sigma se adhiere, constituye la secuencia de promotores (Hsu, 2002).

La transcripción incluye los segmentos de secuencia localizados a 35 y 10 pares de base (bp) corriente arriba del sitio de inicio de transcripción. También se les referencia como cajas -35 y -10 para la subunidad σ70 (Haugen et al., 2008). En *Escherichia coli*, por ejemplo, la caja −35 tiene una secuencia consenso TTGACA. La caja -10, en cambio, tiene un consenso TATAAT.

Todas las secuencias promotoras pueden determinar la expresión de un gen o de un número de genes enlazados corriente abajo. En este último caso, los genes co-expresados forman un operón, el cual es controlado por un solo promotor. (Browning & Busby, 2004).

Para el típico ejemplo de *E.coli*, mientras que la mayoría de los promotores son reconocidas por el factor σ70, otros factores de transcripción sigma son requeridos para iniciar la transcripción de algunos promotores, que son activados por condiciones de estrés (Sachdeva *et al*., 2010).

Es necesario un número de factores de transcripción para la activación o inhibición del inicio del proceso de transcripción. Durante el reconocimiento de la RNA polimerasa forma un complejo cerrado, donde el DNA se encuentra como doble cadena, luego la doble hebra es desnaturalizada para formar un complejo abierto.(Newton & Gey,2012). Después de la síntesis de cerca de 9 nucleótidos de largo de RNA, la RNA polimerasa se presenta en el paso de elongación del mRNA y los factores de transcripción dejan de ser necesarios (Pérez-Martín *et al*., 1994).

Es necesario anotar que la importancia de la caracterización de la red regulatoria génica se debe a la posibilidad de identificar los sitios de unión del DNA reconocidos por los factores de transcripción. Estos últimos generalmente activan o reprimen la expresión génica, por la asociación especifica con el promotor (Bauer *et al*., 2010).

Sin embargo, hay otros factores, tales como los metabólitos de unión y los de interacción proteína – proteína, segundos factores de transcripción, los cuales pueden afectar la expresión génica (Walker *et al.*,2010). Al tener un mayor entendimiento de la regulación génica, la cual juega un papel central e indiscutible en la respuesta celular a cambios ambientales, se puede manipular el comportamiento celular con una variedad de propósitos, entre los que se encuentran aplicaciones metabólicas e ingeniería genética.

Un ejemplo de la aplicación se evidencia en el trabajo de Van Oijen y colaboradores en el 2012, donde demuestra, al conocer los promotores, de *Ostreoco taurii* que regulan el

nivel de fosfato exógeno en el medio, factor limitante en la etapa estacionaria, puede modificar las rutas metabólicas y anular expresiones proteicas conocidas, afectando de sobre manera el ciclo normal de este organismo.

2.2 Algoritmos De Predicción

Los promotores como elementos de DNA, que están directamente relacionados con la regulación de expresión génica, han sido tradicionalmente determinados por análisis experimentales que requieren mucho tiempo y dinero (Gelfand, 1999).

Con base en lo mencionado anteriormente, se aumento el desarrollo en la construcción de algoritmos que permitiera ayudar en la predicción de promotores, comprobando estas predicciones con experimentos en el laboratorio. Sin embargo, la identificación de promotores *in silico* es una tarea ardua, por tres motivos: el primero radica en la naturaleza de los promotores, ya que estos no están claramente definidos y son altamente diversos. Cada gen parece tener una única combinación de motivos regulatorios que determinan su expresión espacio-temporal. Segundo, las regiones reguladoras no pueden ser traducidas a secuencias proteicas para incrementar la sensibilidad en su detección. Tercero, las regiones reguladoras son cortas, de seis a ocho nucleótidos, aumentando su probabilidad de ser encontradas al azar y generando así muchos falsos positivos (Vanet *et al*., 1999).

Algunas algoritmos proveen identificaciones preliminares de estos elementos promotores y son una combinación de distintas características y uso sofisticado de algoritmos basados en *ab initio* e información o datos experimentales (Mount, 2004).

Los algoritmos *ab initio* hacen predicciones mediante un barrido individual de las secuencias y algoritmos basados en similitud, basados en alineamientos de secuencias

homólogas. Este tipo de predicción que utiliza la similitud también es llamado *phylogenetic footprinting* (huella filogenética, en español) (Jong *et al*., 2012).

Cabe destacar, que el continuo desarrollo de sofisticadas bases de datos ha hecho asequible una vasta cantidad de datos biológicos a los investigadores. Adicionalmente, los avances en biología molecular y técnicas computacionales han permitido la investigación sistemática de los complejos moleculares en sistemas biológicos. Numerosos algoritmos han sido desarrollados para la detección en promotores de genomas procariotas (Pitarque *et al*.,2004).

Por ejemplo, Askary, Masoudi y Sharafi (2009) desarrollaron un flujo de trabajo para la predicción de promotores basados en la diferencia de la estabilidad entre la vecindad corriente arriba y corriente abajo del sitio de inicio de transcripción (TSSs). Por su parte, Mann y Chen (2007), usaron una técnica híbrida combinando modelos ocultos de Markov (HMMs) y redes neuronales artificiales (ANN), métodos que han alcanzado una precisión del 70.5%. (Qiu, 2003).

Aunque estos intentos emplearon sofisticados métodos de aprendizaje maquinal para identificar promotores, ofreciendo un aumento en la precisión en determinadas circunstancias, no justifican los inmensos requerimientos computacionales basados en el entrenamiento de estos algoritmos. La selección y optimización de parámetros como el número de capas y nodos ocultos, necesitan suficiente conocimiento *a priori* de las propiedades estadísticas de las observaciones, lo que hace impráctico para el análisis de secuencias de nuevos genomas (Gelfand, 1999).

Por otra parte, existe un método estadístico denominado, la curva Z, originalmente ideado por Zhang, la cuales una poderosa herramienta útil para visualizar y analizar secuencias de DNA (Zhang, 1997 y Song, 2012).

El método está basado en una técnica lineal, adaptable para cualquier computador ya que así se logra reducir la complejidad del procesamiento de los datos, y como no requiriere información previa, es un método práctico para la predicción de promotores de especies de las que no se posee información de las propiedades estadísticas de las observaciones.

2.3 Algoritmos Ab Initio

Estas clases de algoritmos predicen promotores eucariotas, procariotas y elementos reguladores basados en patrones de las secuencias caracterizadas como promotores. Algunos programas *ab initio* están basados en características de las secuencias promotoras como la caja TATA, mientras que otros dependen de la frecuencia de hexámeros u octámeros. La ventaja de los métodos *ab initio* se basa en que la secuencia puede ser aplicada sin tener información experimental previa, sin embargo la limitación es que necesitan entrenamiento, lo cual convierte en específico a cada programa de predicción de promotores (Browning *et al.*, 2007).

La aproximación convencional para detectar los promotores es a través de la coincidencia de una secuencia consenso, representado por expresiones reguladoras o coincidencias de posiciones específicas en una matriz de puntaje construida por sitios de unión conocidos. En cualquier caso, las secuencias consenso o las matrices son cortas, cubriendo entre seis y diez pares de bases (Mount, 2004).

Los puntajes de coincidencias y no coincidencias en toda la matriz son sumadas para dar un puntaje total, el cual es evaluado para determinar su significancia estadística. Esta simple aproximación presenta problemas al obtener falsos positivos en secuencias aleatorias (Mount, 2004).

Para discriminar verdaderos palabras del ruido de fondo, una nueva generación de algoritmos ha sido desarrollados, los cuales tienen en cuenta el alto orden de correlación de múltiples características sutiles, usando funciones discriminantes, redes neuronales, o modelos ocultos de Markov, los cuales son capaces de incorporar información de secuencias vecinas (Mardone et al., 2007).

Para mejorar la especificidad de la predicción algunos algoritmos, se excluyen las regiones codificantes y se enfocan en las regiones corriente arriba entre 0.5 y 2.0 kb únicamente, regiones que tienen la más alta probabilidad de contener estas estructuras promotoras (De Avila E Silva *et al.*, 2011).

2.4 Predicción De Sitios Reguladores En Procariotes.

A pesar de la disponibilidad de investigaciones validadas con biología experimental de laboratorio, los algoritmos de predicción son muy utilizados a gran escala para la identificación y caracterización de distintas secuencias. Un aspecto en la predicción de promotores que lo hacen único, para organismos procariotas, es poder determinar las estructuras operón, ya que estos genes operónicos son coexpresados por un solo promotor (De Avila E Silva *et al.*, 2011). Una vez las estructuras operón son conocidas, sólo la región corriente arriba del primer gen es utilizada como promotor putativo, ya que los otros genes dentro del operón no poseen tales elementos de DNA.

Existe un número considerable de métodos disponibles para la predicción de operones. Uno de los métodos más precisos es un conjunto de simples reglas desarrollado por Wang, Trawick, Yamamoto y Zamudio (2004). Este método cuenta con dos tipos de información: la orientación génica y las distancias intergénicas de un par de genes de interés (Xiong, 2006).

La mayoría de los algoritmos que hacen predicciones de regiones promotoras, tienen en común una hipótesis, la cual establece las diferentes características que poseen las regiones promotoras a las que no son promotoras, como son la concentración de Guanina y Citosina, islotes de CpG, la densidad de factores de transcripción, la composición de palabras y elementos promotores. Algunos algoritmos, como PePPER estudian una, dos o más características, combinadas entre sí, para realizar la predicción más precisa posible de promotores (Wei & You, 2007).

2.5 Modelos de los sitios de unión con los factores de transcripción

NTTS y FTTS, como varios programas, usan un análisis lineal discriminante (LDA) que consiste, en combinar tres aspectos, uno el puntaje de la caja TATA, el segundo preferencias de los sitios de inicio de transcripción y finalmente el puntaje en hexámeros en tres ventanas, sin sobrelapamiento, de 100 pares de bases corriente arriba del sitio de inicio de la transcripción (Vanet *et al.*, 1999).

Otros programas, usan la sobre representación de los sitios de unión de los factores de transcripción, en los que estas secuencias conservadas, reciben puntajes más altos y recuperadas por su relevancia, en distintas estructuras de datos (Vanet *et al.*, 1999).

Algoritmos, como Core Promoter, que comparan la presencia relativa de cada palabra, de un conjunto de secuencias, contra otro conjunto de secuencias, permiten identificar con efectividad motivos cortos y poco degenerados (Nardone, 2004).

Por otra parte, otros programas utilizados, para la predicción de promotores, como Promoter Inspector, tienen en cuenta el descubrimiento, del alto contenido de Guanina y Citosina en estas regiones blanco, para la predicción de promotores (Wei & You,2007).

Este último aspecto, el contenido de Guanina - Citosina y la flexibilidad del DNA son propiedades físicas inherentes de regiones regulatorias, las cuales se comparan con

diferentes zonas, como, corriente arriba, corriente abajo de regiones codificantes, para establecer igualdades y diferencias, información la cual se suministra a redes neurales y modelos de Markov, con el fin de obtener una diferenciación de los promotores, modelos algorítmicos que reducen los falsos positivos (Tavares *et al.,* 2008 y Wei & Yu, 2007).

2.6 Motivos De Representación: Consenso O Matriz

Existe una variedad en las secuencias de unión a los factores de transcripción, variaciones las cuales han sido reunidas por el estudio de genes blanco, mutagénesis, huellas filogenéticas (sitios de unión que son ortólogos en diferentes especies) y microarreglos de proteínas y sitios de unión, elementos que han permitido determinar la especificidad de estos sitios de unión con un alto desempeño (Baldi *et al.,* 2000).

Estos motivos de sitio de unión o perfiles pueden describirse con una secuencia consenso, por un alineamiento de secuencia. La posición de las bases, está en concordancia con un gradiente determinado, por la representatividad de las secuencias, las bases con mayor afinidad o las más frecuentes bases son las que reciben más peso en el resultado del análisis (Liu & Jiao, 2010).

Como el consenso no se puede cuantificar, los grados de afinidad de los sitios de unión, no son características eficaces para la predicción de la frecuencia de nuevos sitios.

Resulta ser mejor opción, para estos motivos, la matriz de posición y peso (PWM) (Iliopoulos *et al*., 2007 y Liu & Jiao, 2010).

2.7 *Método Huella Filogénetica (Phylogenetic Footprinting)*

Debido al continuo crecimiento de genomas secuenciados disponibles, el análisis comparativo de regiones no codificantes se ha convertido en un acercamiento importante en la detección de promotores (Thomas *et al.*, 2003).

La conservación de la secuencias a lo largo de distintas especies es un importante indicador de la funcionalidad. La técnica conocida como *Phylogenetic footprinting* se

refiere a la identificación de estas regiones funcionales, a través de la comparación de genes ortólogos.

Con respecto a lo anterior, se ha observado que los promotores y los elementos reguladores de organismos relacionados como el ratón y el humano están altamente conservados. La conservación es a nivel de secuencia o en la organización de estos elementos. A partir de esto, es posible obtener secuencias promotores para un gen particular a través del análisis comparativo (Xiong, 2006).

Este método puede ser aplicado tanto a secuencias eucariotas como procariotas. La selección del organismo para la comparación es de vital consideración en este tipo de análisis si la pareja a comparar está muy cercana, como el pongo y el humano, las diferencias entre secuencias no son suficientes para filtrar estos elementos funcionales. En cambio, si el organismo se encuentra a gran distancia evolutiva, como el humano y el pez, la larga divergencia evolutiva interpreta demasiadas diferencias para detectar cualquier elemento promotor (Thomas, 2003). El valor predictivo de este método también depende la calidad de la subsiguiente secuencia de alineamiento.

La topología de regulación, la cual controla el desarrollo de los sistemas en metazoarios, tiende a ser compleja, involucrando promotores y potenciadores distales. Los cambios sutiles de algunos de estos reguladores, tienen efectos significativos en el sistema, por lo tanto estos sitios presentan con frecuencia una alta conservación (Xiong,2006).

En contraste, en organismos con células que pertenecen a tejidos diferenciados, los promotores tienden a estar cerca de los sitios de inicio de la transcripción y son menos conservados con especies distantes. Para identificar los sitios de factores de unión. La mejora obtenida por implementar estos algoritmos es que se evita el entrenamiento probabilístico previo, haciéndolo más generalista. Existe también una mayor probabilidad de descubrir nuevos motivos reguladores que comparten los demás

organismos. La obvia limitación es la obligación en las distancias evolutivas entre las secuencias ortólogas (Bork *et al*., 1998).

Con el alineamiento múltiple de secuencias, es posible detectar pequeños motivos de factores de transcripción, que se encuentran más conservados, que las secuencias extraídas al azar. Con este concepto se han hecho múltiples detecciones de elementos reguladores en distintas especies de levaduras y mamíferos (Noureen, 2009).

2.8 *M. tuberculosis*

Junto con el sida y la malaria, la tuberculosis comprende la triada de infecciones que más afectan a la humanidad hoy en día. La tuberculosis es producida por el patógeno *M. tuberculosis*, descubierto por Roberto Koch en 1882, este microorganismo es un bacilo recto alargado y mide aproximadamente 0.4 x 3 micras, su tiempo de duplicación es lento con una duración aproximada de 12 horas o más.

Basados en los estudios de 16 rRNA, se rectifico a la especie de *M. tuberculosis* dentro de las bacterias Gram positivas del orden Atinomicetae. *M. tuberculosis* tiene características generales y especificas, entre las características generales que comparte con el resto de su grupo, se encuentra el alto contenido de Guanina y Citosina, metabolismo aeróbico y tendencia al crecimiento a través de micelios (Kaufmann, 2003). Una de las características especiales más sobresalientes es la estructura de pared celular.

Su pared celular posee un 40% de lípidos, además de proteínas y polisacáridos, es rica en ácido micólico, esta gruesa pared celular es separada de la membrana celular por cuatro capas de peptidoglicano y glicopeptido con moléculas de acetilglucosamina y glucolilmurámico, y cadenas cortas de alanina, sustancias que crean una pared celular permeable y resistente, protegiendo al organismo de factores ambientales adversos y contribuyendo a la inefectividad de los antibióticos convencionales (Ramírez *et al*.,2002).

Además de esta pared celular, *M. tuberculosis* cuenta con distintos antígenos capaces de evadir la respuesta inmune, los principales antígenos se dividen entre dos grupos, en el primer grupo se encuentran los antígenos solubles o citoplasmáticos y el segundo grupo se encuentran los antígenos ligados a la pared celular o insolubles (Ramírez *et al*.,2002).

Uno de los estudios más relevantes en la investigación de tuberculosis se alcanzo en 1998 cuando el genoma de la cepa de referencia H37Rv fue secuenciado por el centro Sanger en Cambridge en colaboración con el instituto Pasteur en Paris. En el estudio anterior se encontró que *M. tuberculosis* tiene un genoma circular con 4,411,529 bp, además se ha logrado identificar 3,924 marcos abiertos de lectura de los cuales el 40% posee función asignada, basado en la similaridad de los genes conocidos, y hay un 44% de genes con probable función, dejando un 16% de genes huérfanos (Kaufmann, 2003).

Al hacer un análisis del genoma, se encontraron 250 genes para el metabolismo de lípidos, los cuales permiten, al organismo adaptarse ante cambios en la fuente de carbono, además hay muchos genes que codifican para enzimas de respiración anaerobia, a pesar de su naturaleza aerobia. Cerca de un 10% del genoma codifica para dos familias proteicas ricas en glicina Pro-Glu (PE) y Pro-Pro-Glu (PPE). Los genes que codifican para estas familias, los principales factores de virulencia y variabilidad antigénica contienen polimorfismos repetitivos conocidos como PGRS y tándem polimórfico (MPTR), los cuales son frecuentemente usados en los estudios de fingerprinting (Ramírez *et al*.,2002).

3. METODOLOGÍA

3.1. Descargar genomas con anotación.

A partir de la base de datos curada RefSeq, se accedió a los 13 distintos genomas disponibles de *M. tuberculosis* en formato GenBank.

Con referencia de las secuencias en NCBI, NC_000962 para la cepa de referencia Hr37V; NC_017528, RGTB423; NC_017026, RGTB327; CTRI-2, NC_017524; para las cuatro cepas de origen Surafricano KZN 605, NC_018078; KZN 4207, NC_016768; KZN 1435, NC_012943; F11, NC_009565.

Para H37Ra referencia NC_009525.1; UT205 la identificación NC_016934.1; CCDC5079 con código NC_017523.1, la cepa CDC5180, NC_017522.1; y por último CDC1551 identificado con NC_002755.

Por medio del algoritmo uno (Figura 2), utilizando BioPerl se extrajo la información del formato GenBank, determinando las coordenadas de las regiones codificantes. Esto se realizó para identificar operones según el método de Salgado y colaboradores, 2000, quienes reportan como la distancia existente entre genes co-expresados, en la misma cadena es menor que la distancia que se halla entre genes que no están regulados por el mismo promotor, es decir si existen 40pb o menos entre el final del gen y el principio del siguiente, codificados en la misma cadena, se identifica como operón.

(Anexo1 y 2, para cadena líder y cadena complementaria respectivamente).

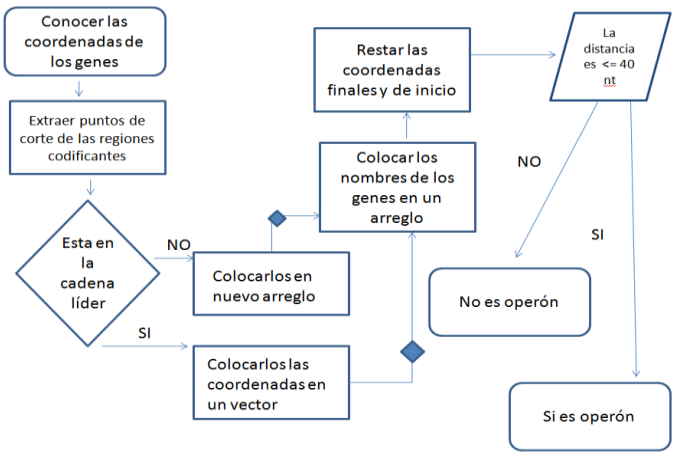


Figura. 2 Flujo de trabajo. En el cual se uso Perl y Bioperl, modulo Bio::SeqIO; el cual permitió extraer las coordenadas de las regiones génicas, con sus respectivos nombres y distancias

3.2 Con base a la información registrada en la base de datos DOOR (Database of prokaryotic Operons) (Mao *et al.*,2009) se logró establecer, estrictamente para *M. tuberculosis*, cual es la distancia discriminatoria entre los genes policistrónicos y monocistrónicos (Salgado *et al.*, 2000).

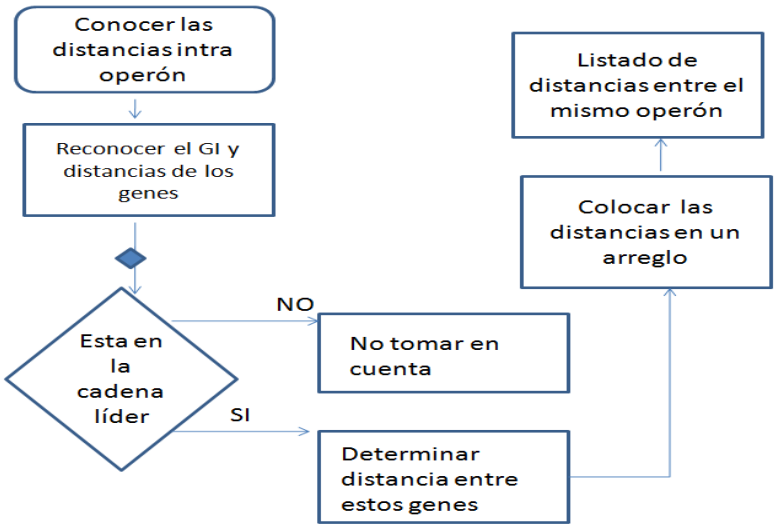


Figura 3. Se Diagrama de flujo, del algoritmo 3, el cual estableció, las distancias que existen entre los genes de un mismo operón(Anexo3).

Esta base de datos DOOR, contiene una predicción de operones, de todos los genomas procarióticos secuenciados; la cual utiliza información de distintas bases de datos de

operones que son curados manualmente, y su programa utiliza características que incluyen la distancia intergénicas, la conservación del orden la distancia filogenética, información de motivos cortos de DNA, y la longitud entre el par de genes. Características que en conjunto permiten una sensibilidad y especificidad del 90%; en sus registros se encuentra información de los genes que están organizados en operones, con sus respectivas coordenadas de inicio y final, la longitud y cadena en la cual está ubicado el operón.

En DOOR se encuentran los genomas de cinco cepas bacterianas las cuales son *M. tuberculosis* CDC1551 (NC_002755), *M. tuberculosis* F11 (NC_009565), *M. tuberculosis* H37Ra (NC_009525), *M. tuberculosis* H37Rv (NC_000962) y *M. tuberculosis* KZN 1435 (NC_012943). Una vez descargados los archivos de texto, se halló la distancia inter-operonica e intra-operonica, de las distintas cepas.

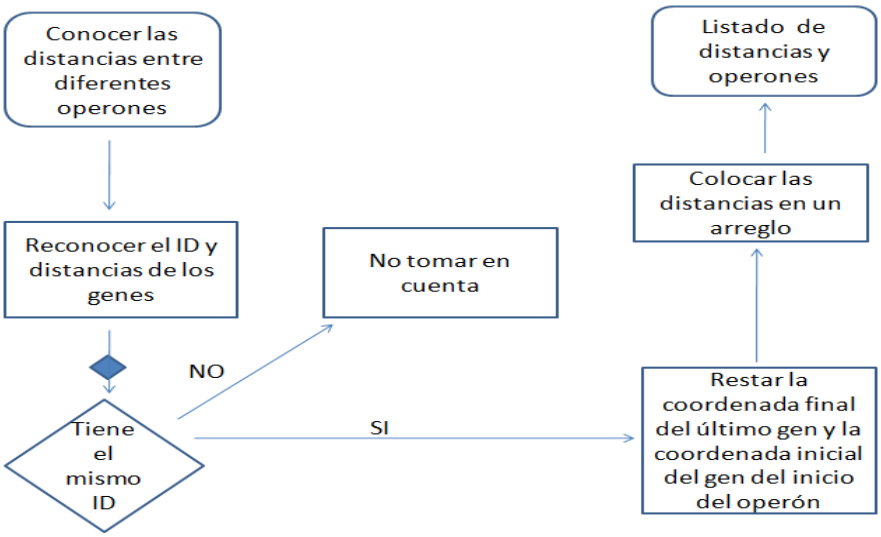


Figura 4. Flujo de trabajo, algoritmo 4, se determinó las distancias que existentes entre distintos operones, usando el modulo de CPAN, Tie::IxHash(Anexo 4)

3.3. Definir de todos los genes cuales están organizados en operones y cuales están organizados de manera individual (Anexo 5).

Una vez, confirmada que la distancia de 40 pb, era adecuada para *M. tuberculosis*, se procedió a realizar el algoritmo (Anexo 5), que ordena las coordenadas iníciales y finales de las regiones codificantes, según si era o no operones.

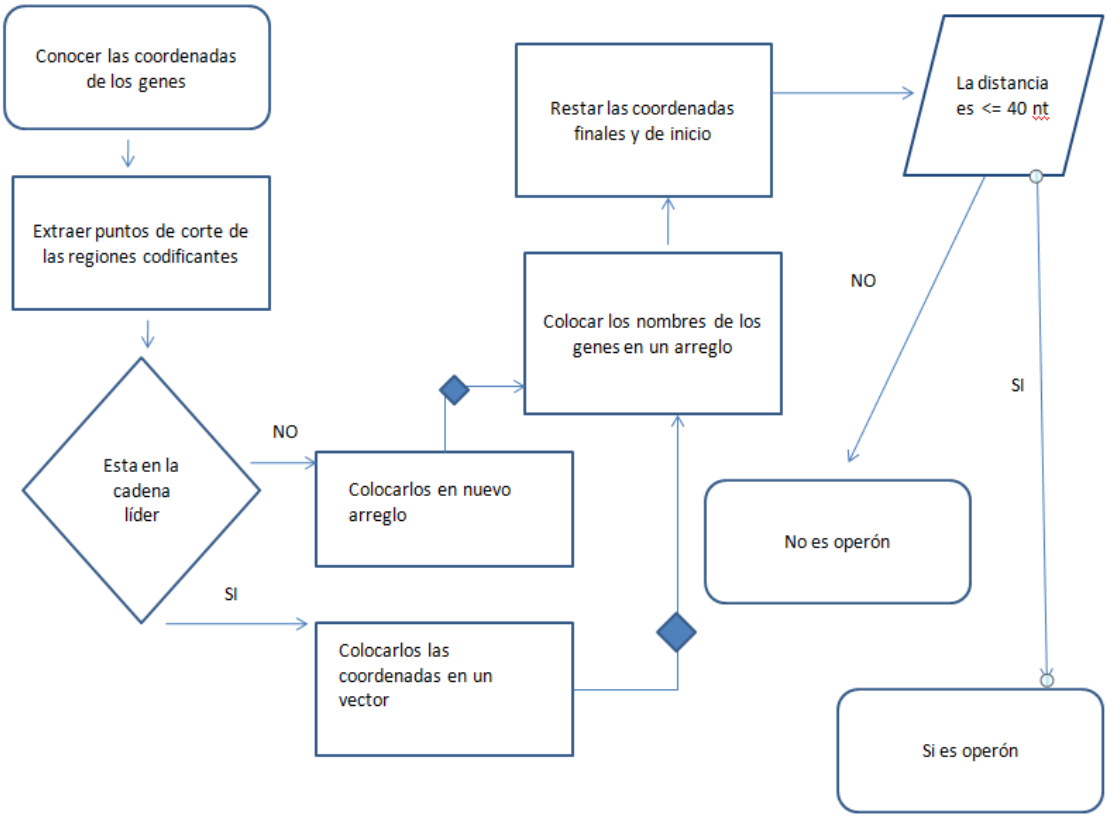


Figura 5. .Flujo de trabajo algoritmo 5. La entrada de este algoritmo, fue una modificación del primer flujo de trabajo, en cambio de mostrar el nombre de los genes, se modifico para que mostrara las coordenadas de inicio y final de cada gen, y así dependiendo de las distancias, encontrar las coordenadas de inicio y final para los genes individuales y los operones.

3.4 Generar un listado de operones y genes individuales.(Anexo 6)

3.5 Con base en esta información se tomó la región corriente arriba de cada gen u operón y dependiendo de la distancia que separa el gen u operón, se tomó hasta 500nt corriente arriba o la distancia corriente arriba, que existe entre genes.

La ubicación de las regiones codificantes en cada cadena, anotadas en los genomas de GenBank, hace posible extraer las posibles regiones promotoras, corriente arriba, de la

cadena líder y complementaria, esto es: para la cadena líder, la región promotora, corriente arriba hacia 5', está ubicado restando la coordenada de inicio, de cada unidad codificante. En cambio para la cadena complementaria, la región promotora, corriente arriba 5'se encuentra sumando a la coordenada final de la región codificante.

Para obtener estas regiones promotoras, se realizó un script con lenguaje Python, donde se utilizó BioPython, y su modulo, BioSeqIO, el cual permitió, la extracción de las secuencias, corriente arriba, de las unidades codificantes. (Anexo 7)

Con base en esta información se computo la frecuencia de octámeros en estas regiones. (anexo 8)

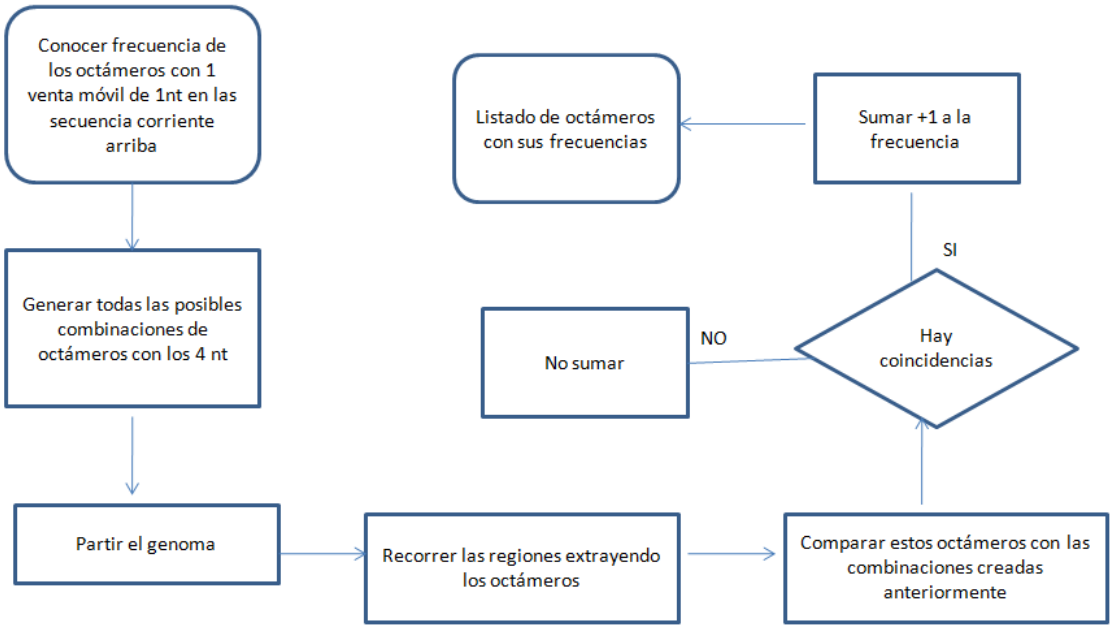


Figura 6. Flujo de trabajo, para el conteo de la frecuencia de los distintos octámeros, en las regiones promotoras extraídas corriente arriba. De cada gen y genes organizados en operón.

Tratamiento Estadístico

3.6 Obtener las frecuencias de lo octámeros

3.6.1 Sumar y calcular el total de nucleótidos de todas las secuencias obtenidas. Esto para tener una muestra de igual tamaño a la obtenida en regiones promotoras.

3.6.2 Se computó lo mismo sobre las regiones al azar obtenidas del genoma y obtener 1000 regiones en total (Anexo 8).

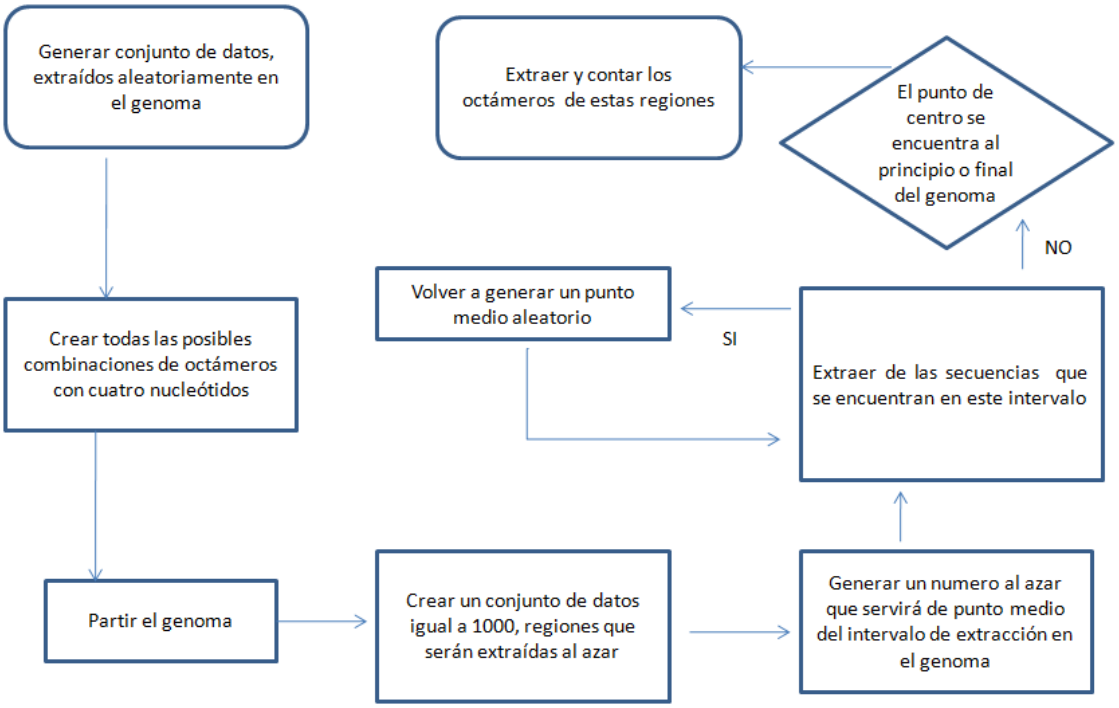


Figura 7. Flujo de trabajo, para hallar el conjunto de datos obtenidas al azar, las cuales deben ser comparadas con las frecuencias de los octámeros obtenidos en las regiones promotoras.

3.7 Con base en los conteos se halló el valor Z y el valor p de cada palabra con un nivel de significancia de 0.05. Es este resultado se le aplicó la corrección de Bonferroni, prueba de comparación múltiple para la corrección del error tipo 1, dado el número de repeticiones (65536) que puede arrojar, un alto número de falsos positivos.

3.8 Agrupamiento de los octámeros, de cada cepa, con una distancia de Hamming de 1. (Anexo 9)

3.9 Se construyó un algoritmo en python (Anexo 10) donde se estableció cuales octámeros obtenidos a través del procedimiento estadístico, se encontraban corriente arriba de las regiones codificantes de *M. tuberculosis* HR37v.

Estos datos se usaron para construir el estudio de enriquecimiento funcional, con ayuda de GeneMerge (Castillo-Davis y Hartl,2003) herramienta que permite calcular los

términos que son estadísticamente sobre representado en dicho conjunto, por medio de la distribución hipergeométrica y una corrección de Bonferroni basada en el número de términos que son examinados en cada análisis (Castillo-Davis y Hartl,2003).

Para el primer test se usaron los recursos de Gene Ontology (Ashburner *et al*.,2000) donde dos genes pueden estar anotados al mismo término, o pueden estar relacionados por medio de un término compartido, como archivo de asociación, para el segundo test se utilizo los recursos expuestos en Tuberculist (Lew,2011), en donde cada gen pertenece a una sola categoría.

4. Resultados y Análisis

Como resultado del algoritmo uno (Anexo 1), permite reconocer y analizar patrones de comportamiento de la variación de las longitudes de las regiones genómicas, en ambas cadenas. Lo anterior debido que si la distancia de los genes es menor de 150 pb, no se puede utilizar el método de identificación de operones distancia de 40 pb para genes policistrónicos. (Vanet *et al*.,1999).

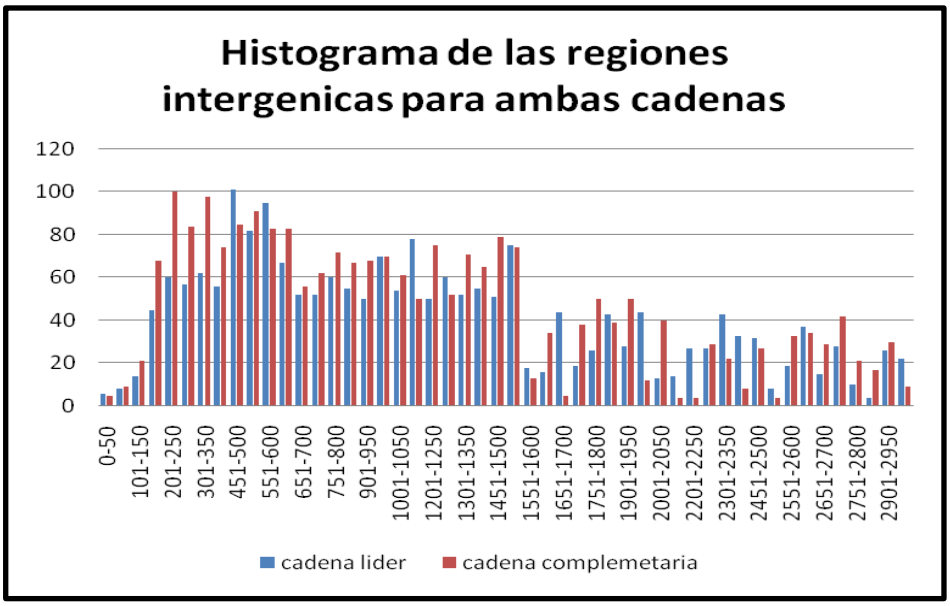


Figura 8. Histograma de las distancias de las regiones intergénicas, de la cepa *M. tuberculosis* HR37v. En color rojo se encuentra la representación de las distancias génicas en la cadena líder, en azul, la cadena complementaria.

Como se puede observar en la figura 7, la mayoría de los genes presentan una distancia mayor a 150 pb, por lo que el método de distancias mayores a 40 pb se clasifica como gen monocistrónicos.

En especial para *M. tuberculosis* las distancias intra-operónicas se encuentran por debajo de 33 nucleótidos, como se muestra en la figura 8.

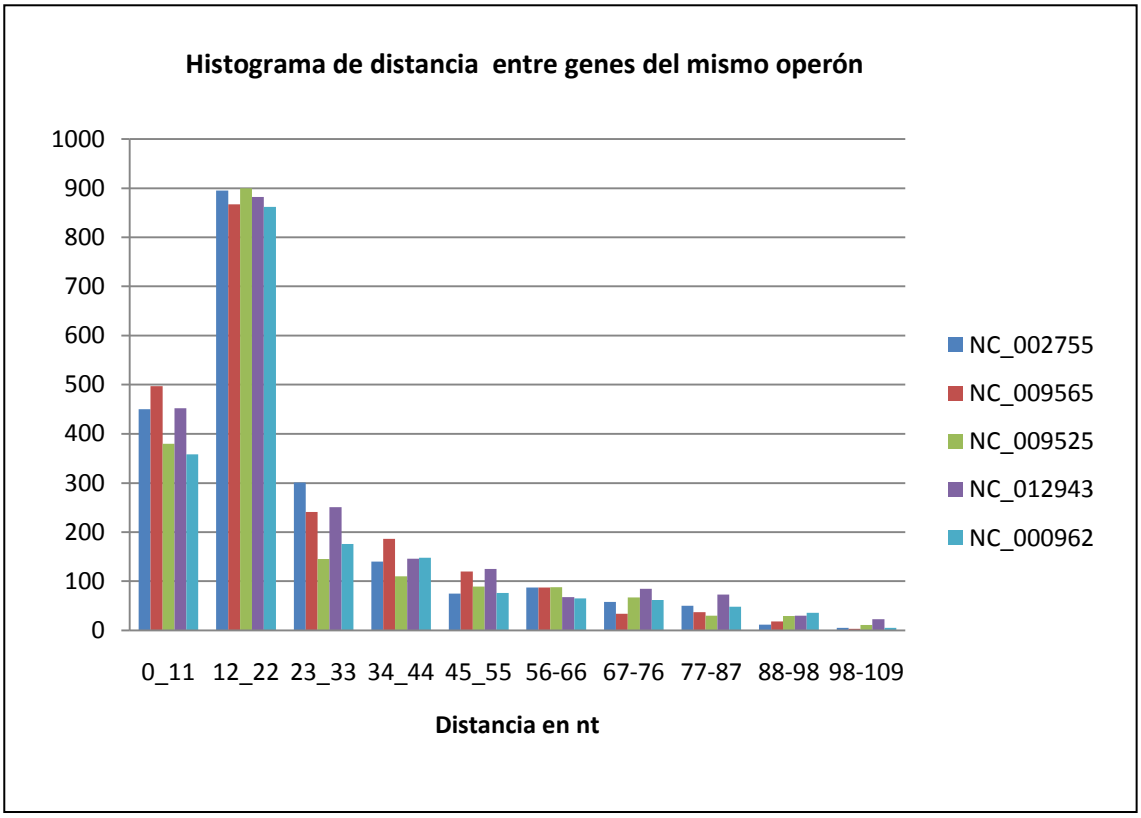


Figura 9. Distancia entre genes del mismo operón.

Resultado de los algoritmo dos (Anexo 2) se realizaron diferentes histogramas de frecuencias, de los cinco genomas que se encuentra en DOOR, lo que permitió establecer las diferentes distancias que existen entre un mismo operón, para las distintas cepas de *M. tuberculosis.*

Al comparar las distintas frecuencias de las cinco cepas bacterianas, sé puede observar que todas tienen un sesgo truncado hacia al eje y, donde la mayor frecuencia de las distancias se encuentran bajo los 40pb, por lo que esta distancia discriminatoria es eficiente para poder diferenciar los genes organizados en operones de aquellos organizados de manera individual en *M. tuberculosis.*

Para evitar sesgo en los datos por el de operones, se realizó el algoritmo tres que permitió calcular la distancia que existe entre dos unidades operónicas consecutivas.

En el histograma de la figura 10. Se puede observar que las cinco cepas bacterianas presentan una distancia no inferior a 2327 nucleótidos, entre distintos operones, por lo cual permite asegurar, que con una distancia de 40pb, no existe riesgo de solapamiento de los distintos operones y cometer errores futuros, en tomar genes de diferente operón, dentro de un mismo grupo de genes co-expresados.

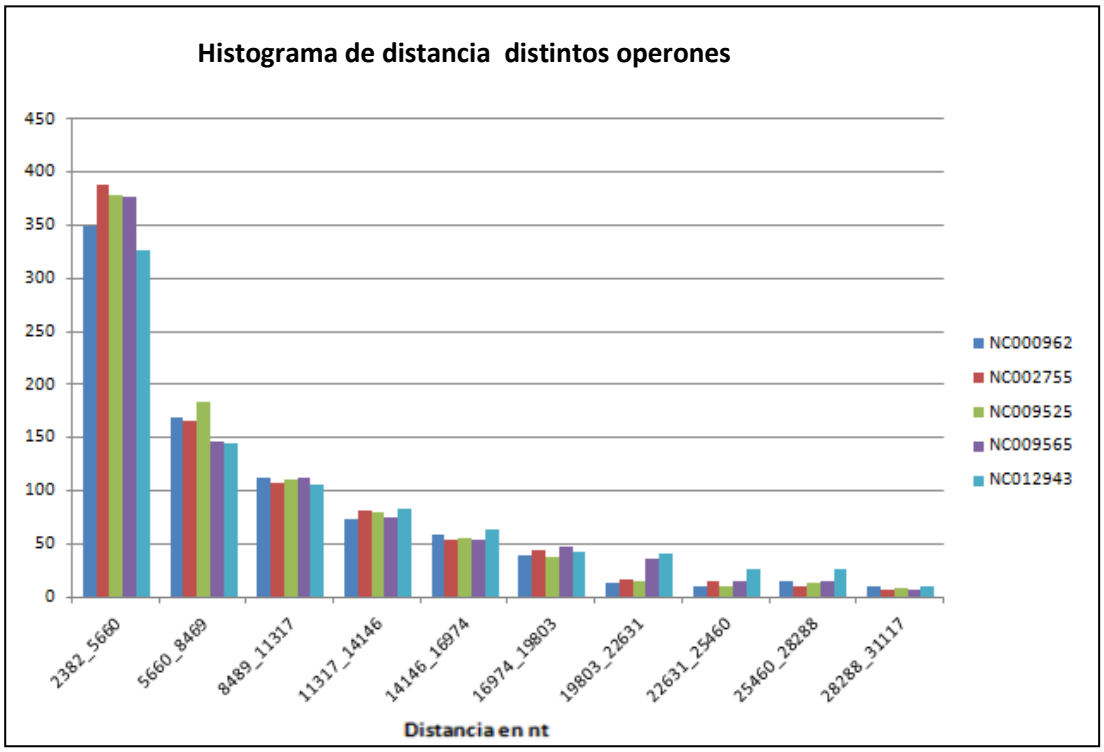


Figura 10. Frecuencias de las cinco cepas de la distancia inter-operónica

Octámeros sobre y sub representados

El análisis estadístico de las palabras sobre representadas usando la curva Z produjo una lista en función de las palabras con un valor $p<0.05$; categorizadas como palabras sobre representadas, las cuales son octámeros que posiblemente estén actuando como sitios de unión con factores de transcripción. Algunas de estas palabras poseen en su mayoría la base Adenina, con un promedio en todas las cepas alrededor del 29.31%, esta base pareada débilmente por un puente doble de hidrogeno, posibilita la apertura de la doble hélice de DNA, para la transcripción (Mariño *et al*.,2004).

Los octámeros, semi homopolimeros, como AAAAAAAT, TTTTTTTA, se encontraron siempre entre las palabras sub representadas de todas las cepas.

Se ha demostrado como el resbalamiento de la polimerasa tiene lugar dentro de las repeticiones cortas y directas, así como en las repeticiones tándem, de hecho en zonas de horquilla de replicación son sitios conocidos como 'hot spots' de deleción, pues una alteración en sitios de regulación afecta productos génicos como, proteínas de membrana de transporte, proteínas de factores de virulencia y de pared celular entre otras (Viguera *et al.*,2001).Sin embargo estos residuos repetitivos en *M. tuberculosis* se han convertido en herramientas de tipificación genómica. Su función biológica es relativamente rara y se ha reportado un rol de estas secuencias en las zonas intergénicas corriente abajo (Tantivitayakul *et al.*,2010).

Octámeros sobre representados

Para las trece cepas, se conto el número de palabras representadas y el número de clústeres que se formaron, bajo el criterio de distancia de Hamming igual a uno, los resultados se muestran en la tabla 4.1.

Tabla 4.1. Número de palabras significativamente representadas con un $p<0.05$

#de registro de cepa	# octámeros sobre representados	#clústeres de octámeros sobre representados con distancia de Hamming 1	Diversidad
NC_000962	1768	1278	1,38
NC_002755	1473	1095	1,34
NC_009525	1747	1261	1,38
NC_009565	1662	1195	1,39
NC_012943	1756	1258	1,39
NC_016768	404	344	1,17
NC_016934	1625	1190	1,36
NC_017026	1135	890	1,27
NC_017522	1497	1089	1,37
NC_017523	1515	1111	1,36
NC_017524	1694	1220	1,38
NC_017528	1037	840	1,23
NC_018078	1727	1250	1,38

Como se puede apreciar, el número de palabras sobre representadas es similar, con excepción de NC_016768 correspondiente a la cepa KZN_4203, la cual posee hasta un tercio menos de octámeros sobre representados, con respecto a la cepa NC_012943. Lo anterior puede estar relacionado con inversiones detectadas en este genoma (Garcia - Betancurt *et al*.,2012) posiblemente debido a errores en el ensamblaje y en la anotación de este genoma.

Diversidad de los octámeros sobre representados

Se desarrollo el algoritmo 7 (anexo 7), script de agrupamiento que usa la distancia de hamming igual a 1, esta distancia entre dos cadenas de igual tamaño hace referencia a una sustitución requerida para cambiar de una cadena a la otra.

Con el objetivo de determinar la diversidad entre los octámeros sobre representados de cada cepa, se relaciono el numero de clústeres, resultado del algoritmo 7 y el número de palabras sobre representadas (tabla 4.1), existe una divergencia similar en las palabras de las distintas cepas, la cepa con menor diversidad es NC_012943 con un índice de 1.39 y la cepa con mayor diversidad NC_016768 con un índice 1.17, sin embargo esta cepa presento el menor número de octámeros sobre representados.

Diferencia de GC entre regiones promotoras y regiones codificantes

Entre las características de las regiones promotoras, se encuentra las islas de CpG, para conocer la representación de estas bases en las secuencias corriente arriba de *M. tuberculosis*, se realizo el script en perl, (anexo 11), el cual lee las secuencias corriente arriba y traduce en 1 cuando se encuentra con GC o CG y 0 para las demás posibilidades, luego en R se leen estos archivos, que permiten, mostrar la cantidad de Citosina y guanina entre 150 nucleótidos corriente arriba y 150 nucleótidos en las regiones codificantes respectivas.

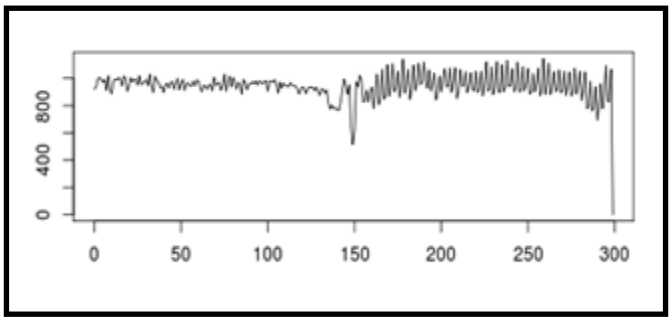


Figura 11. Guanina y Citosina, presente de *M. tuberculosis*H37rv en 300 nucleótidos, 150 nucleótidos corriente arriba y 150 nucleótidos codificantes de todo el genoma. En el eje X la posición sobre la secuencia. En el eje Y la cantidad de GC.

Se puede apreciar que en *M. tuberculosis* al igual que en *E.coli* no existe unas islas de CpG diferenciadas, por lo cual esta característica no ayuda en la predicción de promotores (Wei Huang,2010).

Sin embargo la diferencia en el comportamiento de la gráfica antes y después de la región codificante es notoria, la región promotora presenta un cambio suave en la cantidad de GC, en el principio de la región codificante, 150 nt, existe el mayor descenso en GC por la presencia de ATG, triada característica del inicio de los genes. En la región codificante, presenta un comportamiento de mayores fluctuaciones que las presentes en regiones promotoras. Al realizar las gráficas de las demás cepas

representando la Guanina y Citosina en el genoma, muestran un comportamiento similar al presentado en la cepa H37rv(Anexo 13).

Enriquecimiento Funcional

La ontología tiene como propósito resolver uno de los mayores problemas computacionales en la biología, la ambigüedad en el vocabulario biológico, esto se resuelve con un vocabulario consenso humano y para maquinas, que permitan un comunicación más eficaz (Blair,2010).

La prueba de enriquecimiento funcional, donde se caracteriza los atributos biológicos dado un conjunto de genes, utiliza la base de datos de *Gene Ontology* (GO). En este análisis se encontraron 66 motivos que fueron significativamente enriquecidos de las regiones codificantes en *M. tuberculosis* Hr37v (Anexo 14). Estos incluyen: 6 términos asociados a deshidratasa y fosfatasa; 2 asociados a ureasa, membrana plasmática y regulación de transcripción; 4 términos relacionados con isomerasa, transaminasa y proceso biológico; 3 asociados a transporte de oxigeno y actividad de ligasa; además se encontró 12 términos asociados a sintasa; 3 términos relacionados con reductosa; 19 asociados con transferasa; 9 motivos relacionados a liasa y únicos términos Unión DNA, proceso de síntesis glicolipídico, nucleasa, reacción a ion zinc, topoisomerasa y unión a la superficie celular del hospedero.

Tabla 4.2. Los primeros diez términos Gene Ontology significativamente enriquecidos (valor p<0.05) en *M. tuberculosis*Hr37v

Motivo	Valor P prueba Z	Valor P enriquecimiento	Descripción del Termino GO
CGGCGGCG	0,000	0,027	Fosfatasa
GCCGCCGG	0,000	0,050	Sintasa
CCACCAGC	0,000	0,003	Reductasa
CGTTGCCG	0,000	0,053	Ureasa, Proceso Biológico, Reductasa
CGGTGCCG	0,000	0,004	Fosfatasa,Reductasa
CGCCGGTG	0,000	0,047	Isomerasa
TGCCGCCG	0,000	0,024	Transferasa
GCCGGCGG	0,001	0,043	Liasa
GGCGCCGG	0,001	0,039	Trasnsferasa
CGGCAACG	0,002	0,040	Sintasa

Tabla 4.3. Términos Tuberculist significativamente enriquecidos (valor p<0.05) en *M. tuberculosis*Hr37v

Motivo	Valor p prueba Z	Valor-p	Descripción del Termino Tuberculist
CCGCCGGT	0,049	0,028	Inserción de secuencias y bacteriófagos
CCGCCGGG	0,45	0,044	Proteínas PE/PPE
GCGGCAAC	0,05	0,007	Proteínas PE/PPE, Inserción de secuencias y bacteriófagos
GGCACCGA	0,021	0,0049	Se desconoce
CGTTGATC	0,05	0,007	Hipotético conservado
AGGCCACC	0,05	0,012	Pared celular y procesos celulares

La segunda prueba de enriquecimiento funcional en la cual se utilizó Tuberculist se encontraron 6 motivos que fueron significativamente enriquecidos en los genes de *M.*

tuberculosis Hr37v (Tabla 2). Estos incluyen: 2 términos asociados a inserción de secuencias y genomas de bacteriófagos, 2 octámeros a proteínas PE/PPE, las cuales son proteínas Pro-Glu (PE) y Pro-Pro-Glu (PPE) reportadas por su responsabilidad en los factores de virulencia y en la variabilidad antigénica (Akhter et al.,2012).

Existen tres motivos en común significativamente enriquecidos, CCGCCGGT tiene descripción de sintasa e inserción de secuencias y genomas de bacteriófagos, CCGCCGGG asociado con actividad de proteínas PE/PPE, transaminasa y sintasa, por último el octámeros AGGCCACC relacionado con actividad de liasa, pared celular y procesos celulares.

Algunos de estos posibles sitios de unión a factores de transcripción significativamente enriquecidos se manifiestan, en *M. tuberculosis* en el inicio de su proceso de infección primaria y en la adaptación del patógeno dentro del granuloma.

Cuando el fagolisosoma, crea un ambiente ácido, a través de sus enzimas intraliosomales, el gen Rv1848 con sitio de unión a factor de transcripción CGTTGCCG, está asociado a la enzima ureasa, se expresa como respuesta de la micobacteria a esta condición de bajo pH, permitiendo inhibir o disminuir la producción enzimática intraliosomal, al segregar moléculas básicas neutralizando el pH del ambiente (Zahrt,2003).

Además, ante el estrés oxidativo, se decodifica una serie de enzimas protectoras, de los intermediarios de oxigeno reactivo, y los intermediarios de nitrógeno reactivos, estas enzimas incluyen catalasa, peroxidasa y superoxido dimutasa (Kauffman et al.,2005).

Una vez el sistema inmune, empieza con la respuesta específica, macrófagos y linfocitos se movilizan hasta el sitio de la infección, formando el granuloma inicial, al que se adhieren células multinucleares y epiteloidales o leucocitos modificados (Zahrt,2003).

Este granuloma limita el crecimiento bacteriano al crear condiciones adversas para el microorganismo, entre dichas condiciones se encuentran hipoxia, altas concentraciones de dióxido de carbono, bajo pH, poca accesibilidad a nutrientes, entre otras (Zahrt,2003).

M. tuberculosis sobrevive en el granuloma, al bajar su metabolismo disminuye la expresión del gen ATP sintasa o Rv1306 con sitio de unión a factor de transcripción GCCGCCGG, y cambiando su fuente de carbono, glucosa, a ácidos grasos, gracias a un conjunto de 250 genes que codifican estructuras para el metabolismo de lípidos.

El proceso de catabolismo de ácidos grasos, la ß-oxidación, empieza con la expresión del gen Rv1306 con posible sitio de unión a factores de transcripción GCCGCCGG, convierte el acido graso, ya sea par o impar en AcylCoA, en los siguientes pasos se activa las siguientes enzimas acil-coA deshidrogenasa, enoil –CoA hidratasa y la enol-CoA hidratasa esta ultima produce Ketoacil CoA, el cual puede convertirse en AcetylCoa si el acido graso es par o en PropionilCoa si el acido graso es impar, completando el ciclo de la ß-oxidación (McKinney *et al.,* 2000).

A través de la acetil-CoA transferasa, expresado por el gen fadA5, se transfiere los metabólitos productos de la ß-oxidación al ciclo del Glioxilato, el cual se encarga de metabolizar estos productos para la producción de oxalacetato el cual reacciona con el fosfoenolpiruvato para entrar a la gluconeogénesis y así producir glucosa a partir de ácidos grasos, permitiendo a la micobacteria sobrevivir por largo tiempo el periodo de latencia, condición en la cual la quimioterapia es inefectiva (Tan *et al.*,2010).

Aunque algunos excesos de Acetil-Coa y Propionil-Coa pueden ser transformados en lípidos asociados a virulencia de la pared bacteriana, la mayoría de estos productos

tienen que ser metabolizados por el ciclo de glioxilato, debido a que Acetil-Coa y Propionil-Coa son sustancias bactericidas (Layre,2008).

5. CONCLUSIONES

- En el mecanismo de transcripción de *M. tuberculosis*, los octámeros que están actuando como sitios de unión con los factores de transcripción, no son generados aleatoriamente, estos se conservan, ya que al proceso evolutivo tiende a mantener los motivos funcionales.

- El método estadístico, con la prueba Z, permitió establecer las posibles palabras que están sirviendo como sitios de unión con factores de transcripción, en genomas nuevos como el secuenciado con la referencia RefSeq UT205.

- Al realizar el enriquecimiento funcional, algunos genes significativamente enriquecidos se encontraron asociados a términos de regiones de regulación, ión de Zinc, y sitios de unión transcripcional. Siendo estos genes un blanco opcional en el diseño de drogas debido a su función en la regulación.

- Al conocer los sitios unión a factores de transcripción en *M. tuberculosis* puede ayudar a mejorar los métodos para combatir la tuberculosis a través del desarrollo de drogas diseñada para inhibir la expresión de determinados genes clave como fadA5, el cual permite el catabolismo al transferir el Acetil-Coa y el proponilCoA al ciclo del glioxilato, evitando que la micobacteria se intoxique. El gen ATPF o Rv1306, se expresa aún en condición de latencia, estado en el cual la quimioterapia es inefectiva.

6. BIBLIOGRAFIA

Ando M, Yoshimatsu T, Ko C, Converse PJ & Bishai WR (2003) Deletion I of M. tuberculosis sigma factor E results in delayed time to death with bacterial persistence in the lungs of aerosol-infected mice. Infect Immun 71: 7170–7172.

Akutsu T, Kuhara S, Maruyama O, Miyano S (1998) Identification of gene regulatory networks by strategic gene disruptions and gene overexpression. Proc Noveno ACM-SIAM Symp Discrete Algorithms: 695–70.

Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE. (2012)The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more? Biochimie. 94(1):110-6.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G.(2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. May;25(1):25-9.

Askary,A., Masoudi-Nejad,A., Sharafi,R.(2009). A precise and highly sensitive promoter predictor using neural network fed by nearest neighbors. Genes Genet. Syst., 84, 425–430.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.(2000). Assessing the Accuracy of Prediction Algorithms for Classification: an Overview. Bioinformatics 16, 412–424.

Bauer, A. L., Hlavacek, W. S., Unkefer, P. J., & Mu, F. (2010). Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. PLoS computational biology, 6(11).

Blair Stuart.(2010). A review of the Gene Ontology:past developments, present roles, and future possibilities. BioC 218. recuperado el día 27 de diciembre del 2012 en http://biochem218.stanford.edu/Projects%202010/Blair%202010.pdf

Brameier M, Wiuf C. (2007)Ab initio identification of human microRNAs based on structure motifs. BMC Bioinformatics. Dec 18;8:478.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., & Yuan, Y. (1998). Predicting function: from genes to genomes and back. Journal of molecular biology, 283(4), 707–25.

Browning, D. F. & Busby, S. J. (2004).The regulation of bacterial transcription initiation. Nature Rev. Microbiol. 2, 57–65.

Castillo-Davis, C. I., and D. L. Hartl, (2003). GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics 19(7): 891-892.

Cole, S. T., Brosch, R., Parkhill, J. y otros 39 autores. (1998). Deciphering the biology of M. tuberculosis from the complete genome sequence. Nature 393, 537–544.

Cole, S. T., Eiglmeier, K., Parkhill, J. y otros 42 otros autores. (2001). Massive genedecay in the leprosy bacillus. Nature 409, 1007– 1011.

Corbett, E. L., C. J. Watt, N. Walker, D. Maher, B. G. Williams, M. C. Raviglione, and C. Dye. (2003). The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Arch Intern Med 163:1009-21.

Cooper, G. M., Brudno, M., Green, E. D. (2003). Quantitative estimates of sequencedivergence for comparative analyses of mammalian genomes. Genome Res 13, 813–820.

Das, M. K., and H. K. Dai. (2007).Asurvey ofDNAmotif finding algorithms. BMC Bioinformatics 8 (Suppl. 7).

De Avila E Silva, S., Echeverrigaray, S., & Gerhardt, G. J. L. (2011). BacPP: bacterial promoter prediction--a tool for accurate sigma-factor specific assignment in enterobacteria. Journal of theoretical biology, 287, 92–9.

Farga, V; (1999)The origins of DOTS; Int J Tuberc Lung Dis; 3:175-176.

Garcia-Betancur JC, Menendez MC, Del Portillo P, Garcia MJ. (2012). Alignment of multiple complete genomes suggests that gene rearrangements may contributetowards the speciation of Mycobacteria. Infect Genet Evol. 12(4):819-26.

Gelfand, M. S. (1999). Recognition of regulatory sites by genomic comparison. Research in microbiology, 150(9-10), 755–71.

Gordon, S. V., Eiglmeier, K., Garnier, T., Brosch, R., Parkhill, J., Barrell, B., Cole, S. T. & Hewinson, R. G. (2001a). Genomics of Mycobacterium bovis. Tuberculosis (Edinb) 81, 157–163.

Haugen, S. P., Ross, W., & Gourse, R. L. (2008). Advances in bacterial promoter recognition and its control by factors that do not bind DNA. Nature reviews. Microbiology, 6(7), 507–19.

Helmann, J. D. & deHaseth, P. L. (1999). Protein–nucleic acid interactions during open complex formation investigated by systematic alteration of the protein and DNA binding partners. Biochemistry 38, 5959–5967.

Hsu, L. M. (2002). Promoter clearance and escape in prokaryotes. Biochimica et biophysica acta, 1577(2), 191–207.

Iliopoulos, C., Perdikuri, K., Theodoridis, E., Tsakalidis, a., & Tsichlas, K. (2007). Algorithms for extracting motifs from biological weighted sequences. Journal of Discrete Algorithms, 5(2), 229–242.

Jacques PE, Gervais AL, Cantin M, Lucier JF, Dallaire G, Drouin G, Gaudreau L, Goulet J, Brzezinski R. (2005), MtbRegList, a database dedicated to the analysis of transcriptional regulation in M. tuberculosis. Bioinformatics. Mayo 15;21(10):2563-5.

Jong A, Pietersma H, Cordes M, Kuipers OP, Kok J.(2012) PePPER: a webserver forprediction of prokaryote promoter elements and regulons. BMC Genomics. 2;13:299.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV. (2002)Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. Jun;12(6):962-8.

Jordan IK, Mariño-Ramírez L, Koonin EV. (2005) Evolutionary significance of gene expression divergence. Gene. Jan 17;345(1):119-26.

Kaufmann SHE, Hahn H,(2003), Mycobacteria and TB. Issues Infect Dis. Basel, Karger, , vol 2, pp 97–111

Kaufmann SH, Cole ST, Mizrahi V, Rubin E, Nathan C. M. tuberculosis and the host response.(2005) J Exp Med. (11):1693-7.

Kozak M. (1999) Initiation of translation in prokaryotes and eukaryotes. Gene. 8;234(2):187-208.

Kozak M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. Nov 21;361:13-37.

Lagesen K, Ussery DW, Wassenaar TM.(2010) Genome update: the 1000th genome—a cautionary tale. Microbiology. Mar;156(Pt 3):603-8.

Layre E, Collmann A, Bastian M, Mariotti S, Czaplicki J, Prandi J, Mori L, Stenger S, De Libero G, Puzo G, Gilleron M. (2009). Mycolic acids constitute a scaffoldfor mycobacterial lipid antigens stimulating CD1-restricted T cells. Chem Biol. Jan 30;16(1):82-92.

Lew JM, Kapopoulou A, Jones LM, Cole ST.TubercuList (2011)- 10 years after.Tuberculosis (Edinb). Jan 91(1):1-7.

Liu, L., & Jiao, L. (2010). Detection of over-represented motifs corresponding to known TFBSs via motif clustering and matching. Computers & Mathematics with Applications, 59(2), 779–786.

López López Carlos (2001) Factor de transcripción acii (activator of classii). Formas proteicas, estructura y unión al ADN,Tesis de Maestria, Universidad de Barcelona, Facultad de Biología.

McKinney JD, Höner zu Bentrup K, Muñoz-Elías EJ, Miczak A, Chen B, Chan WT, Swenson D, Sacchettini JC, Jacobs WR Jr, Russell DG. (2000). Persistence of Mycobacteriumtuberculosis in macrophages and mice requires the glyoxylate shunt enzymeisocitrate lyase. Nature. Aug 17;406(6797):735-8.

Mann S, Li J, Chen YP. (2007)A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts. Nucleic Acids Res.35(2):e12

Mariño-Ramírez, L., Spouge, J. L., Kanga, G. C. and Landsman, D. (2004) Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Research. 32:949-958.

Manterola Mantija Jose Maria, (2004) Nuevas aportaciones al diagnostico de las enfermedades causadas por las micobacterias, Tesis doctoral, Hospital General Badalona.

Mao F, Dam P, Chou J, Olman V, Xu Y.(2009)DOOR: a database for prokaryotic operons.Nucleic Acids Res. 37:D459-63.

Mount David, (2004) Bioinformatics Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press.

Münch Richard, Kelin Johannes y Jahn Dieter.(2011). Prediction and Analysis of Gene Regulatory Networks in Prokaryotic Genomes,Systems and Computational Biology – Molecular and Cellular Experimental Systems, Ed. Intech.

Nardone, J., Le, D. U., Ansel, K. M. et al. (2004) Bioinformatics for the 'bench biologist': howto find regulatory regions in genomic DNA. Nature 5(8), 768–774.

Newton-Foot M, Gey van Pittius NC. The complex architecture of mycobacterial promoters. Tuberculosis (Edinb). 2012 Sep 24.

Noureen, N., Kulsoom, N., de la Fuente, A., Fazal, S., & Malik, S. I. (2009). Functional and promoter enrichment based analysis of biclustering algorithms using gene expression data of yeast. 2009 IEEE 13va conferencia internacional, 1-6.

Pérez-Martín, J., Rojo, F., & de Lorenzo, V. (1994). Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. Microbiological reviews, 58(2), 268–90.

Pitarque, M., von Richter, O., Oke, B. et al. (2001). Identification of a single nucleotide polymorphism in the TATA box of the CYP2A6 gene: impairment of its promoter activity.Biochem Biophys Res Commun 284(2), 455–460.

Qiu P.(2003) Recent advances in computational promoter analysis in understanding thetranscriptional regulatory network. Biochem Biophys Res Commun. Sep 26;309(3):495-501.

Ramírez Alejandra, Cocotle Elvia, Méndez Armando, Arenas José (2002) Mycobacterium tuberculosis: Su pared celular y la utilidad diagnóstica de las proteínas 16 y 38 kDa, Revista Médica de la Universidad Veracruzana/Vol. 2.

Sachdeva, P., Misra, R., Tyagi, A. K., & Singh, Y. (2010). The sigma factors of M. tuberculosis: regulation of the regulators. The FEBS journal, 277(3), 605–26.

Saenz Belén. (2007). Situación actual de las resistencias de M. tuberculosis en la población inmigrante de la Comunidad de Madrid, ALAT, Vol. 43, Nº. 6, 2007, págs. 324-333.

Salgado H, Martínez-Flores I, López-Fuentes A, García-Sotelo JS, Porrón-SoteloL, Solano H, Muñiz-Rascado L, Collado-Vides J. (2012). Extracting regulatory networks of Escherichia coli from RegulonDB. Methods Mol Biol. 804:179-95.

Sanchez, A., Garcia, H. G., Jones, D., Phillips, R., & Kondev, J. (2011). Effect of promoter architecture on the cell-to-cell variability in gene expression. PLoS computational biology, 7(3): e1001100.

Song, K. (2012). Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. Nucleic acids research, 40(3), 963–71.

Tan MP, Sequeira P, Lin WW, Phong WY, Cliff P, et al. (2010) Nitrate Respiration Protects Hypoxic *M. tuberculosis* Against Acid- and Reactive Nitrogen Species Stresses. PLoS ONE 5(10).

Tantivitayakul P, Panapruksachat S, Billamas P, Palittapongarnpim P. (2010)Variable number of tandem repeat sequences act as regulatory elements in Mycobacteriumtuberculosis. Tuberculosis (Edinb). Sep;90(5):311-8.

Tavares, L. G., Lopes, H. S., & Lima, C. R. E. (2008). A Comparative Study of Machine Learning Methods for Detecting Promoters in Bacterial DNA Sequences, 959–966.

Thomas, J.W., Touchman, J.W., Blakesley, R.W.et al. (2003). Comparative analyses of multispecies sequences from targeted genomic regions. Nature 424, 788–793.

Vanet, A., Marsan, L., & Sagot, M.-F. (1999). Promoter sequences and algorithmical methods for identifying them. Research in Microbiology, 150(9-10), 779–799.

Van Hijum SA, Medema MH, Kuipers OP(2009) Mechanisms and evolution of control logicin prokaryotic transcriptional regulation. Microbiol Mol Biol Rev. Sep;73(3):481-509.

Van Ooijen G, Knox K, Kis K, Bouget FY, Millar AJ.(2012) Genomic transformation ofthe picoeukaryote Ostreococcus tauri. J Vis Exp. 2012 Jul 13;(65)

Viguera Enrique, Canceill Danielle, and S.Dusko Ehrlich (2001)Replication slippage involves DNA polymerase pausing and dissociation EMBO J. May 15; 20(10): 2587–2595.

Walker, J. M. (2010).Computational Biology of the transcription binding sites. (H. Springer link , primera edición. Londres

Wang, L., Trawick, J. D., Yamamoto, R., and Zamudio, C. (2004). Genome-wide operon predictionin Staphylococcus aureus. Nucleic Acids Res. 32:3689–702.

Wei, W., & Yu, X.-D. (2007). Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. Genomics, proteomics & bioinformatics, 5(2), 131–42.

Wei Huang,(2003), Promoter prediction in DNA sequences, Inpartial Fulfillment of requirements Degreee, National Sun Yat Sen University.

Xiong Jin, (2006), Essential Bioinformatics, Cambridge University Press

Zahrt TC. (2003)Molecular mechanisms regulating persistent Mycobacteriumtuberculosis infection. Microbes Infect. Feb;5(2):159-67.

Zhou, D., & Yang, R. (2006). Global analysis of gene transcription regulation in prokaryotes. Cellular and molecular life sciences : CMLS, 63(19-20), 2260–90.

Zhang,C.T. (1997) A symmetrical theory of DNA sequences and its applications. J. Theor. Biol., 187,297–306.