UACM

Universidad Autónoma de la Ciudad de México

Nada humano me es ajeno

COLEGIO DE CIENCIAS Y HUMANIDADES

Maestría en Dinámica no Lineal y Sistemas Complejos

"Identificación de transferencia horizontal de genes por medio de herramientas de detección de anomalías"

TESIS

Para obtener el grado de Maestro en Dinámica no Lineal y Sistemas Complejos

Presenta:
Leticia Hernández Rodríguez

Director de tesis:
Dr. José Antonio Neme Castillo

MÉXICO DF. ABRIL DEL 2013

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

Dedicatoria

A mi familia y amigos

Especialmente a mis padres, mi hijo Santiago y Héctor.
Mi mayor motivación en la vida

Agradecimientos

No hubiera sido posible realizar esta tesis de maestría sin el apoyo siempre generoso de mi tutor, director de tesis y amigo, el Dr. José Antonio Nemc Castillo, quien mediante su comprensión, motivación, orientación, paciencia y amistad me ayudó a alcanzar mis metas, demostrándome confianza y un apoyo incondicional, realmente es un ejemplo a seguir.

Agradezco la colaboración de los revisores en este trabajo:

Al Dr. Pedro Miramontes Vidal, a quién le debo gran parte de mi inclinación en este tipo de investigación, recordando las palabras de aliento y orientación profesional en mi formación como matemática aplicada.

Igualmente al M. C José luis Gutiérrez quién cotribuyó en mi formación como docente, al brindarme la oportunidad de colaborar con él en distintos cursos de ciencia, promoviendo un significativo aprendizaje de la biología y matemática.

Al Dr. Felipe Contreras, quien con gran dedicación y paciencia promovió una mejor presentación de la tesis, mil gracias por sus observaciones tan detalladas.

Finalmente, al Dr. Pablo Padilla, le agradezco sus consejos útiles y recomendaciones en la presentación final de la tesis.

Mi gratitud es total para la Universidad Autonóma de la Ciudad de México por abrirme sus puertas y mostrarme que un proyecto con causas nobles siempre es bien recibido, valorando el apoyo económico en impresión y empastado de este trabajo. De igual forma agradezco al ICyT/179/2011 por sus estímulos monetario para terminar el proyecto de tesis, reconociendo la paciencia que brindaron sus directivos.

Índice general

Índice de figuras

Índice de cuadros

Introducción

En el afán de descubrir el proceso de evolución, ha sido importante comprender cómo es que los organismos van cambiando al transcurrir el tiempo, superficialmente podemos observar cómo generación tras generación algunas poblaciones cambian modificando su apariencia, pero pocas veces nos percatamos que estos cambios se generan desde la relación entre genotipo-fenotipo y medio ambiente; la propuesta de los evolucionistas para describir la historia de la vida y el descubrimiento de las leyes de la herencia en los seres vivos son el fundamento de la biología moderna.

Es conocido, que desde el descubrimiento del ADN(ácido desoxirribonucleico) muchas de las investigaciones se han basado en descubrir el origen de las especies desde un enfoque informático, pues el manejo de gran cantidad de datos complica su estudio. Actualmente, las técnicas de investigación utilizan un análisis computacional identificando los aspectos importantes sin perder el significado biológico, para después tratar de generalizar los resultados en teorías matemáticas que permitan comprender los procesos biológicos en forma analítica. De hecho, en el campo de la genética, los avances se han desarrollado en temas como: algoritmos genéticos, correlación de proteínas, redes neuronales, redes genéticas, por decir algunos.

Específicamente, en este trabajo se estudia la tan cuestionada, **transferencia horizontal de genes** (genes que no son transferidos desde los ancestros), considerada por algunos autores como un paradigma de la evolución[35], mediante el método de detección de anomalías, el cual consiste en identificar datos anormales en un sistema, buscando relaciones no lineales entre diferentes estructuras. Para nuestro caso, dichas relaciones dentro de una secuencia de genes nos permitirán identificar aquéllos genes transferidos horizontales mediante la propuesta de un algoritmo basado en éste método.

Lo interesante, es definir a la vida no completamente en términos del genotipo ni en la capacidad que tiene de copiarse a sí mismo y ser sujeto de evolución para estar vivo, hay que considerar la producción del fenotipo: es decir, algo desarrollado desde el genotipo y que interactúa con el medio ambiente como lo establece Waddington en palabras de J. L. Gutierrez [21, pág 32-34]. Tomando como referencia esta idea se supone que la evolución de los organismos no sólo se realiza de acuerdo a la transmisión de genes en cada generación, sino también mediante la transmisión de genes en forma horizontal. Es decir, mediante perturbaciones desde el medio ambiente ocasionadas tal vez por la interacción con otros virus o bacterias.

Por lo anterior, surge la pregunta ¿De qué sirve identificar los genes que no son nativos (los que no provienen de los ancestros) de un organismo? se supone que al reconocer estos genes se puede construir filogenias más acertadas entre organismos, aunque todavía existe discusión sobre este tema[35] , además de lograr una mejor comprensión sobre la resistencia a antibióticos por parte de algunas bacterias. Por ello en este trabajo se pretende discutir algunas ideas presentadas en la literatura sobre la funcionalidad de los genes que se consideran transferidos horizontalmente.

Algunos metódos propuestos para este tipo de fenomenos, es decir, para la detección de la transferencia horizonal pueden ser clasificados en cuatro categorias: Desvío de la composición, distribución filogenetica anómala, similitud de secuencias anormales y árboles filogenéticos incongruentes [36], cuyas características se detallarán a lo largo del trabajo. En estos métodos se utilizan varias herramientas matemáticas y computacionales para su análisis, algunos modelos que podemos mencionar son: son los modelos bayesianos, modelos markovianos y metódos de la teoría de la información[25], los cuales toman como referencia algún espacio de características para la identificación de genes horizontales, mostrando diversos resultados un tanto acertados de acuerdo al espacio de características empleado.

El trabajo se divide de la siguiente forma: en el capítulo I se presentan varios aspectos biológicos, mostrando el comportamiento de los genes dentro del ácido desoxirribonucleico (frecuentemente abreviado como ADN o DNA por sus siglas en inglés) y en el ácido ribonucleico (ARN o RNA por sus siglas en inglés) además de algunos conceptos biológicos involucrados en el proceso de sintesís de proteínas con el fin de comprender el modelado. En el capítulo II se detalla la utilidad de las firmas genómicas al analizar secuencias genómicas de diversos organismos. Una vez conocidos los fundamentos generales nos centramos en los antecedentes, métodos para la detección de genes anómales e investigaciones realizadas para la identificación de genes transferidos horizontalmente en el capitulo IV, posteriormente en el capítulo V se exponen avances sobre la utilidad de técnicas de anomalías y se presenta nuestra propuesta del algoritmo para la identificación de los genes no nativos, al igual que se muestra la metodología empleada, para finalmente discutir los resultados realizados en las simulaciones considerando 29 genomas de algunos organismos en el capitulo VI.

Capítulo 1

Conceptos Biológicos

Al parecer el vínculo entre las matemáticas y la biología hace posible resolver muchos de los problemas actuales, pero es sumamente importante que los especialistas en estas disciplinas conozcan los aspectos generales de ambas ramas, por tanto, es pertinente introducir en éste primer capítulo, un panorama general de los conceptos y teorías biológicas que ayudarán a comprender nuestra interpretación y aplicación del algoritmo propuesto.

1.1. ADN y ARN

A tráves de los años nos hemos percatado de la gran diversidad biológica en nuestro planeta, conocemos varias clasificaciones de organismos en reinos, géneros, ordenes, etc, esencialmente mediante la observación de características similares. Sin embargo molecularmente estas discrepancias no son tan notorias inclusive en organismos de una misma especie. Los avances del conocimiento y los experimentos realizados por Avery McLeod y McCarty[23] sugirieron que las discrepancias se debían al material genético contenido en la sustancia llamada ácido nucléico. Desde ahí comenzaron las investigaciones hasta el descubrimiento de la estructura tridimensional del ADN realizado por Watson, Crick y R. Franklin en 1953[1].

El **ADN** es un polímero lineal largo, semejante a un collar, construido por los llamados **nucleótidos** que contienen información de manera que pueda ser transmitida de una generación a la siguiente. Estas macromoléculas están formadas por un gran número de **nucleótidos**, unidos cada uno de ellos y compuestos por un azúcar, un grupo fosfato y una base nitrogenada. Cada ácido nucleico está formado por una de las cuatro posibles bases **A**(adenina), **G** (guanina), **C**(citosina) y **T**(timina) unidas a un eje (esqueleto) de azúcar-fosfato. Las bases tienen la propiedad de unirse con enlaces de hidrógeno formando pares específicos(A con T y C con G), dicho emparejamiento ocasiona la estructura doble hélice, estos pares de bases aportan un mecanismo para copiar la información genética de una cadena de ácido nucléico existente a una nueva cadena, en [3, pág 117] pueden estudiarse más detalles. Algunas de las características mencionadas se identifican en la figura 1.1.

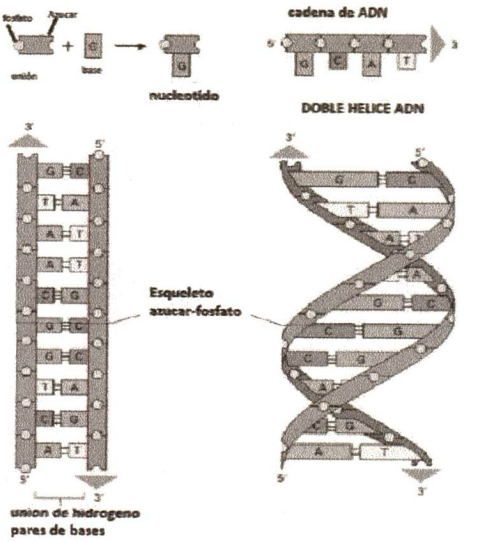

Figura 1.1: Estructura esquemática del ADN, emparejamiento de bases y unión al esqueleto azúcar-fosfato

La complementariedad de las bases (A frente a T y G frente a C) constituyentes de los ácidos nucléicos permiten almacenar y transferir información. La duplicación de estas moléculas se basa precisamente en dicha propiedad.

Por otra parte, el **ácido ribonucléico o ARN** es similar al ADN, pero estructuralmente difiere en tres aspectos [3]:

1. El ARN está constituido normalmente de una sola cadena.
2. El ARN tiene azúcar ribosa en vez de desoxirribosa en su esqueleto
3. El ARN tiene la base de uracilo en vez de la base timina del ADN

1.1.1. ¿Cómo codifica el ADN la información?

Si observamos las características de un organismo, resulta interesante conocer cómo es que el ADN porta la información genética, pues no sólo se debe a la cantidad de subunidades contenidas en él, sino a la secuencia de bases que hay dentro, generadas por el orden de las 4 bases nitrogenadas. Tal como un idioma permite formar una cantidad ilimitada de palabras a partir de un número reducido de letras al variar la secuencia y la cantidad de éstas en cada palabra. Lo mismo hace el ADN para codificar grandes cantidades de información con diversas secuencias y cantidades de nucleótidos en diferentes genes.

El ADN constituye la base físico-química de las instrucciones, es decir, la maquinaria celular(en la cual intervienen diversas enzimas) transcribe y traduce la información del ADN en

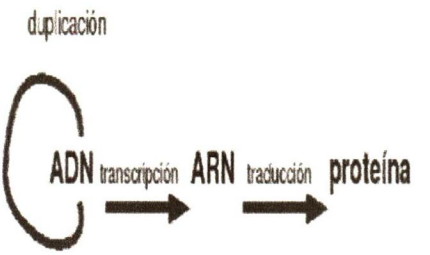

Figura 1.2: Relación entre el ADN y ARN

proteínas específicas. Por tanto, los genes especifican los tipos de proteínas que fabricarán las células. Pero el ADN no es el molde directo para la síntesis de proteínas. Estos moldes son las moléculas de ARN(ácido ribonucléico).

Existen procesos fundamentales que involucran al ADN y ARN para sintetizar proteínas, es decir procesos donde se generan proteínas a partir de los veinte aminoacidos, generados éstos a su vez por secuencias de genes, según [2, pág 158]. Dichos procesos son: la duplicación, la transcripción y la traducción. En la figura 1.2 podemos notar la relación existente.

Duplicación del ADN. El ADN tiene la importante propiedad de reproducirse idénticamente, permitiendo transmitir la información genétcia desde una célula madre a las células hijas, ocasionando la herencia del material genético. Este proceso se realiza en tres sencillos pasos[2]:

1. La doble hélice del ADN debe abrirse en forma que se pueda leer la secuencia de las bases. Esta tarea la hacen un tipo de enzimas llamadas ADN helicasas.

2. Deben sintetizarse las nuevas cadenas del ADN con las secuencias de las bases complementarias respecto de las bases de las dos cadenas parentales. Esto lo realizan otras enzimas llamadas ADN polimerasas (encargadas de la duplicación), las cuales avanzan a lo largo de cada cadena separada de ADN parental combinando las bases de la cadena con los nucleótidos libres correspondientes.

3. Después la ADN polimerasa debe unir los fragmentos para formar una cadena continua de ADN.

Así, al final la cadena ADN parental y su cadena hija de ADN recién sintetizada y complementaria se enrollan una alrededor de otra y forman una molécula de ADN, regresando a su conformación espacial original.

En la figura 1.3 se observa las dos cadenas nuevas de ADN obtenidas de la cadena parental.

Por otra parte, dado que las células viven en un entorno complejo, pueden percibir muchas señales diferentes, incluyendo parámetros físicos como la temperatura, nutrientes beneficiosos y productos químicos nocivos. También pueden recibir información sobre el estado interno de la

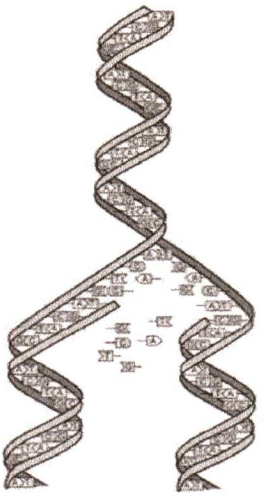

Figura 1.3: Duplicación de la molécula de ADN

misma célula; como daños al ADN, membrana, o proteínas[34]. Por tanto, las células responden a estas señales mediante la producción de proteínas apropiadas que actúan sobre el medio interno o externo, es decir, las células encuentran diferentes situaciones en las que se requieren diferentes proteínas y para satisfacer dicha demanda se realiza el proceso de la **transcripción**.

Detalladamente en el proceso de la **transcripción** se pueden ubicar tres etapas: *iniciación, alargamiento y terminación.*

- En la iniciación, la ARN polimerasa se une a la región del promotor(donde se encuentra un triplete de ácidos nucléicos determinando el inicio) de ADN cerca del principio de un gen, separando la doble hélice del ADN próxima al promotor.

- En el alargamiento, la ARN polimerasa viaja a lo largo de la cadena molde del ADN catalizando la incorporación de los nucleótidos de ribosa a la molécula de ARN. Los nucleótidos en el ARN son complementarios a la cadena molde del ADN. En este proceso se realiza la sustitución de la base nitrogenada timida por uracilo

- En la terminación, al final de un gen la ARN polimerasa encuentra una secuencia de ADN llamada señal de terminación(tres posibles tripletes de bases indicando paro), la ARN polimerasa se desprende del ADN y libera la molécula del ARN.

Al final de la transcripción, la molécula de ADN se enrolla de nuevo y por completo en una nueva hélice. Ahora la molécula de ARN está libre para desplazarse del núcleo al citoplasma

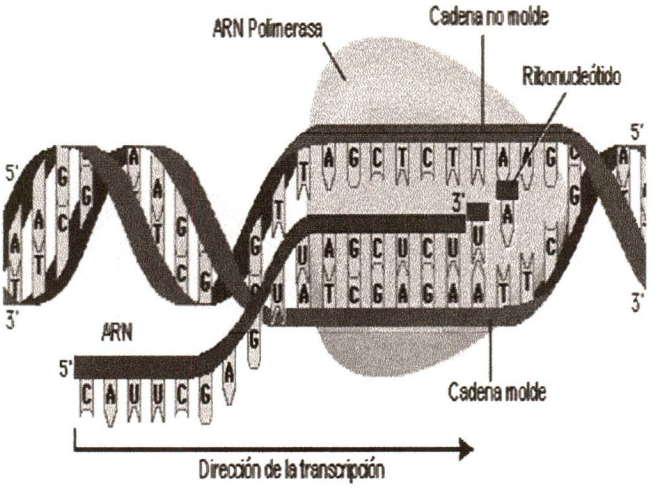

Figura 1.4: Proceso de la transcripción del ARN.

para la traducción, y la ARN polimerasa puede desplazarse a otro gen y comenzar de nuevo la transcripción [2, pág 173]. Todo este proceso se puede observar en la figura 1.4.

Resumiendo, la transcripción del ADN es el primer proceso de la expresión génica, mediante el cual se transfiere la información contenida en la secuencia del ADN hacia la secuencia de proteína utilizando diversos ARN como intermediarios. Durante la transcripción genética, las secuencias de ADN son copiadas a ARN mediante una enzima llamada ARN polimerasa que sintetiza un ARN mensajero(ARNm) el cual mantiene la información de la secuencia del ADN. De esta manera, la transcripción del ADN también podría llamarse síntesis del ARN mensajero.

Finalmente, el proceso de **la traducción** ocurre después de ser liberado el ARN mensajero en la transcripción, y consiste en transformar una secuencia de ARN mensajero en una secuencia de aminoácidos para formar una proteína. Aquí el ARN mensajero se decodifica para producir un polipéptido(unión de varios aminoácidos) específico de acuerdo con las reglas especificadas por el código genético.

El proceso de traducción tiene cuatro fases: activación, iniciación, elongación y terminación (entre todos describen el crecimiento de la cadena de aminoácidos, o polipéptido, que es el producto de la traducción). Dicho proceso ocurre de la siguiente manera:

En las células procariotas todos los nucleótidos de un gen codificador de proteína codifican los aminoácidos así que el ARN que se transcribe es el ARNm que se traducirá en ribosoma.

En los eucariontes el ARNm maduro transporta la información genética del núcleo al citoplasma, donde los ribosomas la utilizan para sintetizar una proteína. Los ribosomas contienen

ARNr(ARN ribosomal) y proteínas que se organizan en subunidades grandes y pequeñas. Estas subunidades se reúnen en el primer codón(compuesto por tres nucleótidos) AUG de la molécula de ARNm para formar la maquinaria completa de síntesis de proteínas. Los ARNt(ARN de transferencia) llevan los aminoácidos correctos a los ribosomas para su incorporación a la proteína en crecimiento. El ARNt que se une y, por siguiente, el aminoácido que se entrega, dependen del apareamiento de bases entre el anticodón(secuencia de tres nucleótidos complementaria al codón de ARNm) del ARNt y el codón del ARNm.

Dos ARN de traducción, cada uno con un aminoácido, se une simultáneamente al ribosoma; la subunidad mayor cataliza la formación de enlaces peptídicos entre los aminoácidos, conforme se acopla cada nuevo aminoácido, se descopla el ARNt y el ARNr avanza un codón para unirse a otro ARNt que lleve el siguiente aminoácido especificado por el ARNm. La adición de aminoácidos a la proteína en crecimiento prosigue hasta que se alcanza un codón de paro, el cual indica al ribosoma que deberá desintegrarse y liberar tanto el ARNm como la proteína recién formada [2, pág 187].

En resumen, una célula logra decifrar la información genética almacenada en su ADN para sintetizar una proteína de la manera siguiente: En cada etapa hay un apareamiento de bases complementarias requiriendo diversas proteínas y enzimas.

1. Salvo algunas excepciones como los genes que codifican para una molécula de ARNt o ARNr, cada gen contiene el código de la secuencia de aminoácidos de una proteína.

2. La transcripción de un gen que codifica para una proteína produce una molécula de ARNm, que es complementaria respecto a una de las cadenas de ADN del gen. A partir del primer codón del ARNm, una secuencia de tres bases que especifican un aminoácido representan una señal de "alto".

3. Las enzimas del citoplasma enlazan el aminoácido correcto a cada ARNt , con base en el anticodón ARNt.

4. Durante la traducción, los ARNt trasladan al ribosoma los aminoácidos que llevan consigo. El aminoácido correcto se elige de acuerdo con los pares de bases complementarias que se forman entre las bases del codón de ARNm y del anticodón ARNt. A continuación, el ribosoma enlaza los aminoácidos unos con otros en secuencia para formar una proteína.

Todo el proceso de transcripción y traducción se aprecia en la figura 1.5.

Esta "cadena decodificadora", que pasa de las bases del ADN a los codones de ARNm, y luego a los anticodones del tARN y finalmente a los aminoácidos, dá por resultado la síntesis de una proteína con una secuencia específica de aminoácidos. La secuencia de aminoácidos está determinada, en última instancia, por la secuencia de bases que tienen un gen.

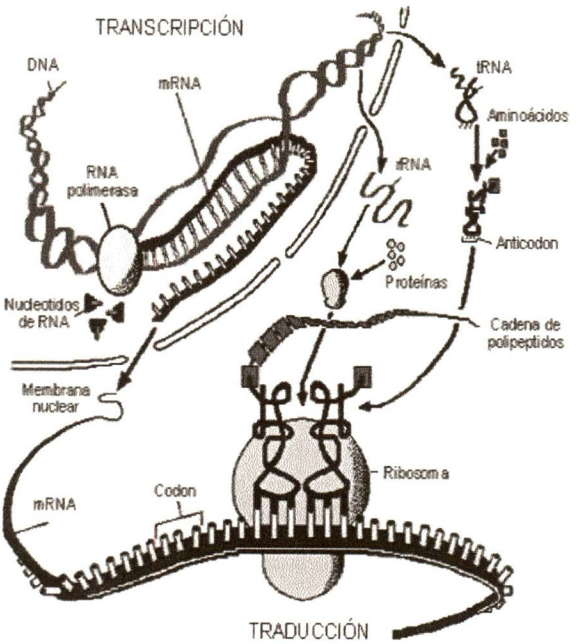

Figura 1.5: Síntesis de proteínas

1.2. Genes, genomas y código genético

A finales de los años setenta en el campo de la ingeniería genética, la estructura y función del gen se fue clarificando y cobrando detalles, prácticamente se define a los **genes** como segmentos de moléculas del ADN, constituidos por diversas partes o secuencias, generalmente contiguas dentro de una molécula de ADN, que constan de una región regulatoria, las cuales establecen cuándo y en qué cantidad se expresará el gen[34]. Después se encuentra el gen estructural que está formado por la secuencia que, al traducirse, determina la secuencia de aminoácidos de la proteína codificada. Al fin, hay señales(secuencias de paro), también constituidas por secuencias de ADN que determinan el término de proceso de transcripción evitando que continúe hacia otros genes que se encuentran más adelante. Todas estas señales son "leídas" y decodificadas por interacciones establecidas por proteínas específicas.

El código genético es la relación entre la secuencia de bases en el ADN (o de su ARN transcrito) y la secuencia de aminoácidos en las proteínas. Este código relaciona el idioma de cuatro letras del ADN tomando grupos de tres en tres, en el idioma de las proteínas constituido por 20 letras o monómeros. En 1961, los experimentos de Francis Crick, Sydney Brenner [37] y otros establecieron las siguientes características del código genético:

1. Tres nucleotidos codifican un aminoácido. Las proteínas se construyen a partir de un conjunto básico de 20 aminoácidos, pero tan sólo hay 4 bases diferentes, es decir que se requiere un mínimo de tres bases para codificar al menos 20 aminoacidos, experimentalmente se comprobó que un aminoácido esta codificado por un grupo de tres bases o codón.

2. El código no se encima. Consideremos una secuencia de bases ABCDEF, en un código con solapamiento, ABC especificará el primer aminoácido BCD el siguiente, CDE el siguiente y así sucesivamente, en un codigo sin solapamiento ABC designará al primer aminoácido , DEF al segundo y lo mismo en adelante.

3. El código no tiene puntuación. En principio una base (denotada por Q) podría servir como una "coma" entre grupos de tres bases.

4. El código genético es degenerado. O sea algunos aminoácidos están codificados por más de un codón, pues que existen 64 posibles tripletes de bases y sólo 20 aminoácidos.

Visto abstractamente el código genético es un lenguaje en el cual 64 posibles combinaciones de 4 bases Uracilo (U), Citosina (C), Adenina (A) y Guanina (G), tomando 3 a la vez, especifican un solo aminoácido o terminactión de la secuencia de la proteína. Con 64 posibles "palabras" y 21 posibles "significados" , esto es claramente el potencial para diferentes codificaciones de codones para idénticos aminoácidos. Éstas combinaciones se pueden obtener siguiendo el orden del disco mostrado en la figura 1.6.

1.3. Secuencias del genoma

La información contenida en el ADN determina cuándo, cómo y dónde las células crecen y se dividen, siendo el responsable de los procesos de evolución que generan los millones de formas

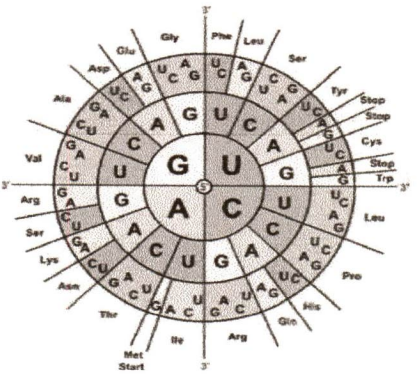

Figura 1.6: Combinaciones de las bases nitrogenadas para obtener los aminoácidos que generan diversas proteínas

de vida que existen en la tierra, por ello muchos científicos se han dedicado a descubrir las reglas que hacen posible tales procesos, realizando análisis de secuencias de ADN, tomando ventaja del enfoque multi-escala, el cual consiste en usar pequeñas escalas para regiones codificantes de proteínas y usar grandes escalas para grandes regiones, algunos métodos utilizados son las transformada de Morlet, función de información mutua, espectro de Fourier, entre otros[4].

1.3.1. Homología de secuencias

En genética y biología molecular, la homología de secuencias se refiere a la situación en la que las secuencias de dos o más proteínas o ácidos nucleicos son similares entre sí, debido a que presentan un mismo origen evolutivo. Generalmente se puede concluir que dos secuencias son homólogas sobre la base de la alta similitud que las mismas presentan.

La diferencia entre similitud y homología consiste en que la primera se basa en la observación, mientras la segunda se obtiene de la interpretación de una alta similitud entre dos secuencias, debido a su origen común, así que una alta similitud de secuencias puede deberse simplemente al azar.

De acuerdo a la variación de genes en cada especie las secuencias homólogas pueden clasificarse en dos tipos, ortólogas y parálogas.[23]

Tipos de secuencias homólogas

- Ortólogas: son aquellas secuencias homólogas que se han separado por especiación, al diverger la especie en dos especies difrentes. La evidencia más concluyente de que dos genes similares son ortólogos es el resultado del análisis filogenético sobre el linaje de ese gen.
- Parálogas: si las mismas se hallan separadas por un evento de duplicación, es decir, si el

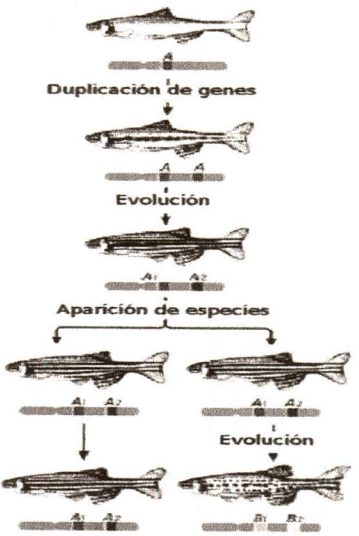

Figura 1.7: Visualización de genes parálogos y ortólogos

gen de un organismo se duplica para ocupar dos posiciones diferentes en el mismo genoma, entonces las dos copias son parálogas.

En la figura 1.7 se puede observar cuales son las diferencias entre los genes ortólogos y parálogos. de acuerdo a la definición anterior tenemos que el gen A1 comparado con A2 es parálogo, lo mismo para el gen B1 comparado con B2 pues se han realizado copias por eventos de duplicación dentro del mismo genoma, mientras que los genes A1 con B1 y A2 con B2, son genes ortólogos, pues cada uno pertenecia a una especie diferente separada por la especiación.

Capítulo 2

Firmas genómicas

El desarrollo de herramientas estadísticas para el análisis de secuencias biológicas ha sido útil para capturar el efecto de la evolución en los genomas. Un descubrimiento importante en este sentido es que las características de composición de un genoma llevan información sobre la historia evolutiva de una especie. Estas características de composición hacen posible transportar señales específicas permitiendo, a los organismos distinguirse en base a un género y especie. Esta especificación puede ser interpretada como el resultado del proceso de adaptación de especies al medio ambiente. Inclusive los parámetros ambientales y estructurales son algunos de los factores que determinan la composición del ADN, ARN y proteínas[4]. Así las características específicas de especies en las secuencias biólogicas se describen a menudo por firmas genéticas a largo plazo.

Debido a la estructura que presentan las secuencias del genoma (vistas como una cadena lineal formada por 4 bases nitrogenadas), existen modelos matemáticos llamados **firmas genómicas** que representan en forma matemática y compacta una secuencia de ADN, con el objetivo de resumir las características específicas de un organismo desde su ADN. Un ejemplo de éstos son los modelos paramétricos que hacen uso de estadísticas recolectadas de los fragmentos de la secuencia de ADN [22]. Y dado que los parámetros estimados son únicos para una especie, las firmas genómicas constituyen una caracterización específica de cada especie.

La firma del genoma es una herramienta potencialmente importante para las aplicaciones de la bioinformática, pues utiliza dos propiedades importantes: *la especificidad de especie* mencionada en el párrafo anterior y la *invarianza*. La invarianza se refiere a que no importa el fragmento que se tome de la secuencia siempre conservará similares caracterizaciones matemáticas incluso en diferentes longitudes de los fragmentos.

Las dos propiedades de especificidad de especie e invarianza determinan la fuerza de una firma genómica. Una firma fuerte, que es altamente específica de la especie e invariante puede caracterizar un genoma usando sólo una pequeña parte aleatoria del genoma. Para muchas aplicaciones bioinformáticas, se requiere detectar el origen de las especies mediante selección al azar de pequeños fragmentos del genoma, aquí las firmas del genoma son buenos candidatos para dichas tareas.[22, pág 3]

En la práctica, la especificación de especies y la generalización de las firmas están limitadas por muchos factores, pues poca especificidad y problemas de capacidad de invarianza obligan a utilizar firmas genómicas simples que no requieran estimar un gran número de parámetros. Pero, el uso de estructuras simples también puede dar lugar a caracterizaciones pobres. Este fenómeno es un obstáculo importante para el uso de la firma del genoma en diversas aplicaciones de la bioinformática.

Una observación considerable es que; la especificación de especies e invarianza de una firma genómica no sólo depende de la estructura de la caracterización, sino también de cuánta distancia o similitud hay entre las firmas medidas, de hecho muchos investigadores consideran que el uso de una distancia métrica apropiada explota mejor la información de la firma, para ello se utilizan en algunos casos la frecuencia de occurrencia de cortos oligonucleótidos[20]. Por lo tanto, es necesario mostrar la siguiente definición:

Una **firma genómica computacional** es una estructura matemática que puede ser generada desde un fragmento arbitrario del genoma. Es decir, dado un fragmento aletatorio de cualquier genoma con suficiente longitud, uno puede generar la misma o similar caracterización matemática para el genoma dado. Así, para encontrar similitudes entre las firmas, es necesario utilizar una métrica en un espacio donde la firma este definida, con el objetivo de describir la relación filogenética entre organismos a partir de sus segmentos de ADN. Oskan en [22] lo expresa matemáticamente como:

$$d_S(S(G_{x_i}), S(G_{x_j})) < d_S(S(G_{x_i}), S(G_{y_k}))$$

donde G_{x_i} y G_{x_j} son secuencias aleatorias del genoma G_x, y G_{y_k} es una secuencia aleatoria del genoma G_y, para $i, j, k \in N^+$. $S(.)$ es una operación sobre el dominio de la posible secuencia del ADN y el rango de $S(.)$ existe en el espacio métrico de la firma. La distancia en el espacio de la firma es mostrado con la métrica $d_S(.,.)$ originando las discrepancias entre cada secuencia, algunas veces éste proceso necesita realizarse de acuerdo a las características composicionales del genoma, por ello es necesario conocer con base en que contenido de la secuencia se puede hacer esta tarea.

2.1. Características composicionales para las firmas genómicas.

Los atributos determinan la calidad de una firma del genoma: la especificidad de especie e invarianza. Las firmas del genoma tienen invarianza, pues estos atributos aparecen por todo el genoma, y la especificación de especie, al ser diferentes para los distintos organismos. Nuevamente basándonos en [22, pág 8-13] podemos mencionar algunas de las características composicionales.

2.1.1. Contenido GC

El contenido de G-C, se considera como una caracterización composicional del genoma, pues mide el contenido de guanina y citosina, es decir el número de pares de G y C dentro de la molécula de un fragmento de ADN o ARN, ocasionando una variación filogenética (la determinación de la historia evolutiva de los organismos). Debido a su variación con diferentes parámetros ambientales los valores de contenido GC parecen ser específicos de cada especie.

Los diferentes valores de contenido GC de varias especies, y su relativa invarianza dentro del mismo genoma se observó en la década de 1960. El contenido de GC es también la forma más simple de tener una firma, ya que sólo utiliza un parámetro y la distancia métrica es simplemente la diferencia aritmética de estos valores. Además, como las bases están distribuidas a lo largo de un genoma en proporciones similares, el contenido de GC satisface el atributo de invarianza de una firma genómica.[22]

2.1.2. Contenido de aminoácidos

Recordemos que los aminoácidos son moléculas orgánicas que contienen átomos de carbono, hidrógeno, oxígeno y nitrógeno en su composición, siendo el resultado de la traducción de agrupamientos funcionales de moléculas de ADN. Los aminoácidos forman pequeñas cadenas de polímeros llamados polipéptidos o monómeros de proteínas.

El contenido de aminoácidos representa las frecuencias relativas de los aminoácidos utilizados en una proteína o un proteoma (conjunto de proteínas expresado por un genoma) con un vector de 20 dimensiones. Se trata de una simple función que mide el nivel de proteoma, análogo al nivel de contenido de GC del genoma.

Ciertos organismos prefieren diferentes aminoácidos en sus proteínas. Se ha notado que el uso de aminoácidos en la especificación de especies es el resultado de ciertos procesos evolutivos. La respuesta a distintas condiciones de temperatura ambientales, la cantidad de nutrientes, la susceptibilidad a la oxidación, se encuentran entre los factores que configuran el contenido de aminoácidos.

La preferencia por ciertos aminoácidos también se conserva durante un genoma. Debido a que los genes no difieren de forma significativa en la preferencia de los códigos de aminoácidos, esta preferencia es un fenómeno generalizado a través del genoma. Esta propiedad de la firma fue utilizada por Sandberg para la clasificación de las proteínas con base en su contenido de aminoácidos [22].

2.1.3. Codones sinónimos

Los codones sinonimos a veces estan representado por vectores de 64 dimensiones que reflejan la frecuencia relativa de cada codón que codifica para un aminoácido, recordemos que hay 64 codones diferentes (61 codones que codifican para los aminoácidos y 3 codones de paro), pero

sólo 20 diferentes traducciones en aminoácidos.

El exceso en el número de codones permite a muchos aminoácidos ser codificados por más de un codón. En la década de 1980, se señaló que cada especie de manera sistemática prefiere ciertos codones para codificar un aminoácido; este fenómeno es cierto para la mayoría de los genes de un organismo. La proposición de que el uso de sesgos de codones es específico de la especie es conocida como hipótesis genoma Grantham[12]. La variación de sesgo del codón entre los genes de un organismo se atribuye con frecuencia a los niveles de expresión génica y la abundancia relativa del ARNt(ARN de transferencia) que está en una célula.

La variación entre los genomas es más importante que la variación intergenómica. Aunque el uso de sesgo de codón no hace cambiar la composición de proteínas, también se ha relacionado con la composición de aminoácidos, la estructura de proteínas, sesgos mutacionales direccionales, y la estructura secundaria ARNm(ARN mensajero). La relación directa de uso de codones sinónimos para los factores ambientales pueden ser vistos por el hecho de que el uso de codones sinónimos lleva la información que revela las señales sobre el comportamiento térmico y las vías respiratorias de un organismo. Siguiendo una metodología similar estadística utilizada para el uso de aminoácidos se ha demostrado que los codones sinónimos exhiben características del genoma de la firma.[22]

2.2. Características organizacionales para las firmas genómicas.

En ocasiones es muy importante conocer la organización que presentan las bases dentro del gen, para saber qué es lo que codificarán, pues no basta con la frecuencia de algun atributo, por ello se considera importante caracterizar una firma genómica con base a su organización interna, para lo cual se han desarrollado nuevas técnicas, donde intervienen ideas de correlación y hacen uso de herramienta de teoría de la información, por ejemplo; los modelos bayesianos, los modelos de Markov [25] y los métodos de la teoría de información [26].

Hay que mencionar que la detección de las relaciones entre dos o más variables no se limita al análisis de la expresión génica, pero es de gran interés en muchas áreas de la ciencia. Las variables que no son estadísticamente independientes sugieren la existencia de alguna relación funcional entre ellas. Si bien existen varios métodos para cuantificar la dependencia lineal entre las variables, el marco de la teoría de la información (Shannon, 1948) proporciona una medida general de las dependencias entre las variables. En particular, una fuga de correlación de Pearson no implica que dos variables son independientes. La información mutua por lo tanto proporciona un criterio mejor y más general para investigar las relaciones entre las variables, pues no sólo encuentra relaciones lineales como es el caso de Fisher y Spearman[31]

2.2.1. Teoría de la información, uso de la función de información mutua.

Un método de detección de correlaciones de largo alcance en secuencias de ADN y que además realiza correlacciones no lineales es el uso de la Función de Información Mutua (IFM). La función de información mutua promedio fue introducidá por primera vez por Claude Shannon, para el estudio de las señales de los canales en condiciones ruidosas [27].

Detalladamente, la teoría de la información fue inventada en respuesta a los problemas prácticos que enfrentan los diseñadores de sistemas de comunicación, tales como teléfonos y módems de datos. El problema básico es encontrar una manera eficaz de transmitir información de un lugar a otro.

Básicamente la idea que propuso Shanon en su artículo de investigación, para resolver este problema consiste en un proceso de dos pasos: en primer lugar, eliminar la redundancia existente mediante la compresión en el mensaje enviado y a continuación, introducir el tipo adecuado de redundancia para transmitir el mensaje a través del canal en cuestión. Estos dos conceptos importantes se conocen como "codificación de la fuente" y "codificación de canal", respectivamente. Para realizar el proceso anterior, es importante conocer la cantidad máxima de compresión posible, para ello se utiliza la entropía y la información mutua.

Se puede suponer que los mensajes a comprimir se generan de forma aleatoria y no hay una correlación entre el mensaje que se nos pide comprimir ahora y los mensajes ya comprimidos. Matemáticamente, cada x_k puede verser como un símbolo del mensaje y solicitar el número de bits por símbolo promedio que debe ser utilizado para comprimir la secuencia infinitamente larga de símbolos independientes e idénticamente distribuidos $x_1, x_2, \ldots$. Si cada símbolo tiene una probabilidad $p(X)$ de ocurrencia. Por tanto, en [38] menciona que cuando X es una variable aleatoria y su distribución es $p(X)$, su entropía se define como:

$$H(X) = E_{p(X)}[logp(x)],$$

donde $E_{p(X)}[.]$ denota la esperanza con respecto a $p(X)$.

En la práctica la esperanza se calcula sumando cuando X es discreta e integrando cuando X es continua. Cuando la base es 2 del logaritmo, es decir, log_2, las unidades de la entropía son los bits y $H(X)$ es precisamente el número de bits por símbolo requeridos en promedio para comprimir una secuencia infinitamente larga de símbolos cuando cada símbolo tiene probabilidad de $p(X)$ de ocurrencia.

Es por esta razón que se explica en muchas ocasiones a la entropía como la cuantificación de la "ambigüedad" o "incertidumbre" sobre la variable aleatoria X. Cuando X tiene un único estado posible (con probabilidad de 1), no hay ambigüedad acerca de X y la entropía es 0, como se observa en la figura 2.1. Sin embargo, si X tiene uno de dos estados con probabilidad p y $1 - p$, respectivamente, $(0 \leq p \leq 1)$, la entropía se maximiza cuando $p = 0,5$ y $H(X) = log_2 2$. Esto es exactamente 1 bit, e implica que una secuencia de ceros y unos igualmente probables

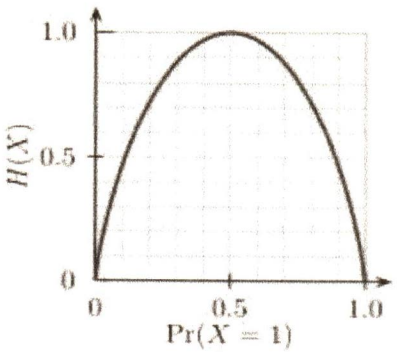

Figura 2.1: visualización de entropá de acuerdo a distintos valores de X

no se puede comprimir. Tenga en cuenta también que, si $p = 0$, $0log_20 = 0$.

Ahora supongamos que hay dos variables aleatorias X e Y, y que de alguna manera están relacionados entre sí. Por ejemplo, X podría denotar la temperatura, mientras que Y denota humedad. Nuevamene por [38] podemos realizar el siguiente análisis.

El hecho de que Y indica una información parcial sobre X se refleja en el hecho de que si Y es conocido, entonces X se puede comprimir más que si Y no es conocida. En el ejemplo anterior, si Y no se conoce, entonces es imposible comprimir X porque cada uno de los resultados es igualmente probable. La cantidad adicional de compresión posible, $I(X,Y) = H(X) - H(X|Y)$, se llama la información mutua y mide la cantidad de información Y proporcionada alrededor de X. Resulta que la información mutua es simétrica $-I(x,y) = I(Y,X)-$ por lo tanto, no hay necesidad de especificar el orden de X e Y.

Para nuestro interés, ésta teoría atraé la atención de los biólogos computacionales como un medio para observar las mutaciones correlacionadas en los sitios no contiguos y estructuras secundarias y las correlaciones en las secuencias de proteínas.

En el análisis de moléculas se supone que x es una secuencia aleatoria generada desde el ADN, donde x_i y x_j son casos correspondientes al ADN, entonces la información acerca de x_i dado el conocimiento de x_j se calcula mediante:

$$I(x_i, x_j) = H(x_i) - H(x_i|x_j)$$
$$I(x_i, x_j) = H(x_i) - (H(x_i, x_j - H(x_j))$$

Donde, $H(.)$ es la entropía de Shannon, considerada como una medida del desorden o la encargada de medir la cantidad de información promedio que contienen los símbolos usados en un sistema[27]. En este caso una secuencia de ADN se ve de nuevo como un proceso estocástico, es decir, con comportamiento no determinista, en la medida que el subsiguiente estado del

sistema está determinado tanto por las acciones predecibles del proceso como por elementos aleatorios. También se supone que el proceso es en amplio sentido estacionario y ergódico, en otras palabras, no cambia al pasar el tiempo y las correlación entre las observaciones tienden a cero, para más detalles puede consultar el anexo 3.

Por lo tanto, la función de información mutua (IFM) queda como:

$$I(x_i, x_j) = \sum_i P(x_i, x_{i+k}) log_2(\frac{P(x_i, x_{i+k})}{P(x_i)P(x_{i+k})})$$

La estimación de las probabilidades pueden hacerse simplemente mediante el recuento de frecuencias relativas de los pares de bases situadas a una distancia k. La estimación de la información mutua promedio da una medida estadística de la cantidad de información compartida entre los nucleótidos y las bases a una distancia k. Por lo tanto, constituye una medida de la correlación dentro de una secuencia de ADN.

Así lo realiza Matthew[28] comparando esta técnica con una medida de fractal usando mutaciones en Silicio con secuencias reales, virtuales y aleatorias. También se ha utilizado para explorar la existencia de patrones estadísticos que difieren en la codificación o no codificación del ADN y que pueden ser encontrados en todos los organismos vivos para el origen filogenético como lo hizo Ivo Grosse [29].

Capítulo 3

Transferencia horizontal de genes

Ciertamente en nuestro entorno podemos observar gran variedad de organismos, Colin[4] señala que sí toda la biodiversidad es difícil de comprender, la magnitud de la información genética es aún más difícil.

Pareciera que nunca tendremos una idea clara sobre el número de genes que evolucionaron a lo largo del tiempo, pues la mayoría han desaparecido sin dejar rastro. Ahora nos enfrentamos a la perspectiva de extinciones masivas, el cambio climático, el empeoramiento de la contaminación, etc. Muchos de los organismos descubiertos son microbios, y la gran mayoría de los genes están presentes en los virus desconocidos que ni siquiera tienen una clasificación como organismos, es difícil llegar a un acuerdo con el concepto de protección de la biodiversidad, sin embargo, se supone que una de las causas de diversidad se debe al flujo de genes entre organismos.

Al parecer existen dos flujos de genes entre organismos: *flujo de genes vertical*, el cual ocurre de padres a hijos y el *flujo de genes horizontal* entre diferentes organismos[23]. En todos los organismos hay un flujo vertical de genes de generación en generación, en los microbios unicelulares, este proceso es simplemente la división celular el cual no hace ninguna diferencia, sí el genoma es un único cromosoma circular o varios cromosomas lineales. En cada caso, el ADN se duplica para hacer dos cromosomas idénticos. El ADN se replica y forma dos cromosomas iguales que se separan en células hijas por mitosis. En los organismos con reproducción sexual, hay machos y hembras donde cada célula tiene dos copias de cada cromosoma: una paterna y una materna.

Para el caso del flujo horizontal de genes Colin J. Sanderson[4], menciona cuatro principales mecanismos de transferencia:

1. Algunos organismos toman el ADN del ambiente. La principal limitación para esta forma de transferencia es la rápida descomposición del ADN en la mayoría de las situaciones de la naturaleza. Esto se ha mejorado en el laboratorio mediante el uso de ADN purificado y con procedimientos de electroporación para aumentar la eficiencia de la transferencia.

2. La transferencia de plásmidos entre las bacterias durante una fase de conjugación sexual. Este mecanismo de transferir se limita a la transferencia de plásmidos entre las bacte-

rias estrechamente relacionadas que pueden sufrir conjugación. Es claramente un medio importante de transferencia natural y una parte importante de la ingeniería genética.

3. Transferencia por virus, se limita a las especies estrechamente relacionadas, debido a la limitada gama de huéspedes de la mayoría de virus. Los virus pueden transferir ADN en los genomas de animales. Gran parte de éste ADN termina como basura, aunque algunos genes funcionales se han encontrado en los genomas eucariotas.

4. Transferencia entre especies estrechamente relacionadas con la fertilización cruzada para formar un híbrido. Híbridos de animales, como la mula es estéril no juegan ningún papel en la evolución. Las plantas híbridas son a veces fértiles y forman nuevas especies. Por eso los híbridos son importantes en la filogenética convencional.

En resumen, una **Transferencia horizontal de genes** o **HGT** por sus siglas en inglés abarca cualquier mecanismo de pasaje de información biológica o cultural entre organismos parentalmente no relacionados, es decir, involucra a individuos pertenecientes a especies distintas. La transferencia horizontal es un proceso importante que afecta a los linajes biológicos y culturales en varias escalas, distorsionando la señal filogenética debido a la discordancia de caracteres que se producen entre las entidades implicadas [32].

El acceso inmediato del público a los datos de la secuencia completa del genoma abre una nueva edad biológica, ocasionando implicaciones sorprendentes para la medicina, a sabiendas de la teoría microbiana, la secuencia completa del genoma también proporciona una gran cantidad de información para rastrear las redes evolutivas. Así, la mayoría de genes de los genomas completos de las arqueas (grupo de microorganismos unicelulares pertenecientes al dominio Archaea) se asemejan más a las contrapartes entre las eubacterias y los no eucariotes. Por supuesto, esto pone en cuestión el arraigado "Arbol universal de la vida" determinado a partir de los análisis comparativos de las secuencias de nucleótidos de genes que codifican ARN ribosomal y varias proteínas. Al mismo tiempo, la relación genética entre bacterias y arqueas apoya los puertos de transferencia horizontal de genes como un importante factor en la especiación y la evolución molecular de microorganismos[6].

3.1. Antecedentes de la transferencia horizontal de genes

A medida que el siglo XX llegó a su fin, había mayor apreciación del hecho de que los genes se encuentran en las mitocondrias y cloroplastos incorporandose con frecuencia en el núcleo central del genoma de su organismo huésped. No obstante, han sido intensos los debates a través de los años sobre la posibilidad de que la transferencia de material genético entre las diferentes especies pueden desempeñar un papel significativo en la evolución. Antes de la era genómica, pocos eventos de transferencia de genes se documentaban en la literatura pero en los últimos años muchos trabajos se han desarrollado.

Al principio, los fundadores de la filogenia molecular utilizaban información molecular de diferentes proteínas y genes para reconstruir la filogenia y la relación entre los organismos. Uno

de estos marcadores moleculares, la secuencia de genes del ARN, propuso ser una buena herramienta para la reconstrucción de las viejas relaciones filogenéticas, sin embargo, pronto se dieron cuenta que los diferentes genes podrían conducir a incongruencias filogenéticas y filogenias conflictivas, sobre todo en los microorganismos, mediante la agrupación de las especies o grupos de especies que se dividen por otros marcadores morfológicos, fisiológicos o moleculares[30].

Por otro lado, los mecanismos para la transferencia de material genético entre los microorganismos eran bien conocidos desde el principio de la biología molecular e investigaciones de la ingeniería molecular. En este contexto, el concepto de la transferencia horizontal de genes entre organismos surgió a principios de la década de 1990 como una explicación alternativa para estos eventos filogenéticos conflictivos.

Desde entonces, los nuevos y abundantes datos de secuencias de organismos eucariotas y procariotas, han reforzado esta idea, sobre todo con el auge de la era genómica, lo que ha permitido la comparación de series completas de genes entre organismos. Esto ocasionó que la tradicional evolución basada en el modelo del árbol fuera cuestionada, considerando la posibilidad de un gen sustancial de cambio. Primero se observó que algunos genes de E. coli muestran sesgo en la frecuencia de codones significativamente de la mayoría de los genes. Estudios posteriores filogenéticos han demostrado que la proteína arquea se puede clasificar en dos grupos distintos: homólogos bacterianos y eucariotas. Este último comprende a los denomi-nados genes informativos (involucrados en la traducción, la transcripción y la reproducción), y su existencia puede ser explicada en el contexto del modelo de la evolución temprana el cual dicta que eucariotas y arqueas descienden de un ancestro común, mientras que, el primero parece ser el resultado de numerosas transferencias de genes entre bacterias y arqueas [5, pág 423]

Desde la aparición de la transferencia horizontal de genes como una manera de explicar la incongruencia filogenética entre los diferentes árboles de genes, un número considerable de estudios se han publicado acerca de los genes que han sido adquiridos por transferencia horizontal tanto en bacterias y arqueas, así como en las células eucariotas. Los estudios muestran que la transferencia puede ocurrir desde las bacterias hasta arqueas, de arqueas a bacterias, desde las bacterias hasta eucariotas, de eucariotas a bacterias e incluso dentro de los mismos eucariotas. Sin embargo, es en la evolución bacteriana y arqueas que la transferencia horizontal de genes ha sido más ampliamente documentada y aceptada[30].

La importancia de la transferencia horizontal de genes va más allá de ayudar a interpretar las incongruencias filogenéticas en la historia evolutiva de los genes. De hecho, existe una fuerte evidencia que las bacterias patógenas(bacterias que originan las intoxicaciones y toxiinfecciones alimentarias) pueden desarrollar resistencia a múltiples fármacos simplemente mediante la adquisición de genes de resistencia a antibióticos de otras bacterias[5].

Recientemente, varios estudios se han propuesto para una nueva síntesis evolutiva que abarca otros mecanismos de mutación. Por ejemplo; la seleción natural explica los cambios evolutivos mediante las limitaciones en el desarrollo epigenético (factores no genéticos que intervienen en el desarrollo de un organismo), entre otras modificaciones.

Hoy parece evidente que la transferencia horizontal o lateral de genes, es una fuerza importante que impulsa la evolución de bacterias, arqueas y eucariotas unicelulares. Por lo tanto, debe considerarse también como parte de la estructura de cualquier síntesis evolutiva[30].

Se puede decir, entonces que: **La transmisión no genealógica de material genético de un organismo a otro** conocida como transferencia horizontal de genes es un mecanismo que permite la adquisición de novedades evolutivas. Sin embargo, estas adquisiciones son principalmente no genealógicas, cuestionando como lo establece el autor en [30], la concepción neoDarwinista de un proceso gradual de aparición de nuevas características y funciones. Una controversia se presenta al tratar de discutir la importancia de la HGT en la evolución y el reto que le implica reconstruir la relación filogenética entre organismos, además de saber qué genes se han transferido y mantenido desde entonces, todo esto intenta resolver el paradigma de la evolución.

3.2. Avances y características de transferencia horizontal de genes.

Estudios recientes han demostrado que los genomas de las plantas han sido objeto de transferencia horizontal de genes. En los sistemas de parásitos de plantas, la HGT parece ser facilitada por la íntima asociación física entre el parásito y su huésped. La HGT en estos sistemas se ha observado cuando una secuencia de ADN obtenida a partir de un parásito se coloca filogenéticamente muy cerca de su anfitrión en lugar de con sus parientes más cercanos [39].

Científicos de Singapur, Malasia y Estados Unidos investigaron la transferencia entre dos plantas: Rafflesia y su planta anfitriona la Tetrastigma rafflesiae, véase la figura 3.1, y al analizar el transcriptoma (los productos transcritos de genes activados) encontraron 49 genes transcritos por la parásita, respondiendo por el 2% de su transcriptoma entero, los que originalmente pertenecían al hospedero. Tres cuartos de esas transcripciones parecen haber reemplazado la propia versión de la planta parásita.[39]

Lo verdaderamente sorprendente de este estudio es que la tasa de transferencia de genes entre la Tetrastigma rafflesia y su flor parasitaria sea tan alta como las tasas de transferencia horizontal visto en los genes de las bacterias, planteando la posibilidad de que HGT puede proporcionar un beneficio a Rafflesia para el mantenimiento de estos genes.

Una pregunta recurrente frente a la importancia de la transferencia horizontal de genes en la evolución es saber ¿cuántos genes de un organismo han sido adquiridos por transferencia horizontal?. Es evidente que en bacterias y arqueas, la transferencia de uno sólo o unos pocos genes pueden dar a los organismos receptores la oportunidad de ejercer una nueva función. Sin embargo, su importancia como un mecanismo evolutivo puede ser limitado si pocos eventos de transferencia horizontal de genes han tenido lugar en la historia de su vida.

Para resolver la cuestión anterior varios criterios se han propuesto y discutido: el sesgo

Figura 3.1: La Rafflesia es un género de plantas parásitas que habitan en el sureste asiático

en el uso de codones, las diferentes composiciones de bases en relación con otros genes en el genoma y la incongruencia filogenética. Siendo algunos de éstos sometidos a varias críticas. Bajo la hipótesis de que los genes extraños se asemejan a los genes nativos después de muchas generaciones, porque los dos conjuntos de genes, residentes y transferidos, apoyan a los sesgos mutacionales mismos una vez que forman parte del mismo genoma, algunos criterios podrían permitir la identificación de los recientes eventos de transferencia de genes, pero no son eficientes en la detección de eventos que ocurrieron hace mucho tiempo. En relación con la incongruencia filogenética, algunas investigaciones señalan problemática de diferenciar la transferencia horizontal de genes entre la pérdida de genes en la evolución del árbol filogenético. Una comparación cuidadosa de los diferentes métodos filogenéticos empleados junto con el uso de una prueba de compatibilidad entre los árboles pueden proporcionar algunas pistas sobre el proceso involucrado [30, 35].

A pesar de estas advertencias, numerosos trabajos han sido llevados a cabo en los últimos años para tratar de hacer frente a la importancia de la transferencia horizontal de genes en bacterias y arqueas. La evolución y los resultados obtenidos son controvertidos dependiendo de si se hace hincapié en la transferencia de genes o la pérdida de genes. Sin embargo, el panorama emergente es que la transferencia horizontal de genes juega un papel más importante en la evolución microbiana que se pensaba.

Por ejemplo, Aaron [40] comenta algunas ventajas y desventajas de la transferencia horizontal de genes. Al parecer la HGT es beneficiosa para una célula si el gen adquirido confiere una función útil, pero es perjudicial si el gen no tiene ninguna función, si es incompatible con los genes existentes, o si es un elemento móvil egoísta replicante. Propiciando así una tasa de transferencia de los genes. También se ha propuesto que la tasa de HGT fue muy alta en las primeras etapas de evolución procariota, y por lo tanto no hubo diferentes linajes de organismos. Sío cuando la tasa de HGT comenzó a caer, los linajes comienzan a surgir con sus propios conjuntos de distintos de genes.

Hoy en día se estima que entre 1,6 y 32,6 por ciento de los genes de cada genoma microbiano han sido adquiridos por transferencia horizontal de genes. Por otra parte, un estudio reciente muestra que esto aumenta drásticamente a 81,15 por ciento si se considera el efecto acumulativo de la transferencia horizontal de genes hacia linajes, lo que refuerza la importancia de este mecanismo en la evolución microbiana[30].

En la actualidad, ya se sabe que la HGT es importante en la evolución eucariota unicelular y es ampliamente aceptado que los genomas eucariotas contienen varios genes nucleares de origen microbiano que se han transferido de los antepasados a través de la mitocondria y el plastidio por antiguos eventos de endosimbiosis. Pero estudios recientes reconocen el papel de la HGT en la modulación de la evolución de otros genomas eucarióticos de una manera diferente: por la adquisición de nuevos genes de endosimbiontes de vida. Además, se ha demostrado que la adquisición es importante en los hongos y la evolución de las plantas, en genes extraños de bacterias y otros eucariotas, finalmente, también ha contribuido a la evolución de los rotíferos de Bdelloidea[30]

Al igual que L. Boto[30, 40], consideramos relevante saber el tipo de genes que estan involucrados en la transferencia de eventos y la manera en la cual estos genes son mantenidos después de dicho proceso, resaltando el impacto de la HGT al tratar de responder si la transferencia horizontal de genes ha sido igualmente prevalente en la evolución, y si dicha transferencia es mayor entre organismos cercanamente relacionados o entre organismos distantemente relacionados.

Algunas de las investigación sostienen las hipótesis de que existen dos tipos de genes involucrados en el proceso de HGT [10]:

- **genes informacionales**, implicados en la replicación de ADN, transcripción y traducción con participación en muchas iteraciones moleculares.
- **genes operacionales**, involucrados en el mantemimiento de células y con poca actividad molecular.

Los primeros son menos propensos para la HGT comparados con los genes operacionales, como lo describe Ravi [10] y otros autores, al mostrar que éstos últimos genes han sido continuamente transferidos horizontalmente dada la divergencia de los procariotas. Aunque otro trabajo muestra que no todos los genes participan igualemente en la HGT, pues un estudio muestra que algunos de los genes informacionales representan una fracción de genes inclonables.

En general se apoya la hipotesis de que, es más importante la participación de los genes en múltiples interacciones moleculares para la transferencia y el mantenimiento de los genes en el mundo procariótico, que la clase funcional a la que pertenezcan los genes transferidos.

Por otra parte, estudios sugieren que la transferencia de genes puede ser más frecuente entre distancias cortas e intermedias pero incomún entre organismos que son separados por un largo tiempo de evolución, es decir, apesar de que la transferencia puede ser mas frecuente en organismos relativamente cercanos, esta transferencia también se puede dar entre organismos distantes, contribuyendo en la evolución de arqueas y bacterias. También se supone que mediante la HGT se puede hacer una reconstrucción de un árbol filogénetico en el mundo microbiano, aunque esto todavía tiene mucha crítica.

3.3. Métodos para la identificación de genes por transferencia horizontal

Con el tiempo, una serie de métodos han sido concebidos para la identificación de genes horizontalmente adquiridos. Tradicionalmente, los métodos filogenéticos se han utilizado para demostrar que un gen ha sido horizontalmente transferido. Estos métodos funcionan bien cuando se cuenta con una cantidad suficiente de datos para la construcción de árboles con un buen soporte, pero frecuentemente ésto no ocurre, por tanto, otros enfoques deben ser explotados con el fin de identificar los genes transferidos horizontalmente en el genoma en cuestión.

Algunos ejemplos de estos enfoques incluyen la clasificación inesperada de similitud de secuencias entre los genes homólogos de un organismo en particular, mostrando una fuerte similitud con un homólogo de una taxonomía distante, para la conservación de genes en un operón de taxones distantes, y la composición de nucleótidos atípicos.

Muchos de los métodos anteriormente publicados para la detección de transferencia horizontal de genes, se basan en el contenido de genes y operan bajo el supuesto de que en un organismo dado, existen rasgos compositivos que permanecen relativamente constantes a través de su secuencia genómica. De acuerdo a ésto, los genes que muestran composición atípica de nucleótidos en comparación con las características prevalentes de composición de su genoma probablemente han sido adquiridos a través de un proceso horizontal. En consecuencia, un número de características se han propuesto para la definición de las firmas de un genoma, pues un gen que se aparta de la firma se pueden marcar como un candidato de transferencia horizontal.

Ragan. M[11] menciona cuatro métodos que han sido usados para la detección de HGT, incluyendo las características arriba mencionadas.

3.3.1. Con base en la distribución de genes o genotipos

Bajo la hipotesis de que los genes han alcanzado su distribución actual en los genomas, exclusivamente a través de un proceso de descenso vertical, todas las familias de genes deben estar representados en todos los genomas. Para estar seguro, la homología en una familia dada puede diferir entre genomas debido al mayor o menor nivel de duplicación de genes y al cambio de secuencias oscureciendo las relaciones de linaje entre ellos, aunque la pérdia de genes complique esta regularidad.

Existen factores que pueden indeterminar una explicación nativamente vertical. Dado que las secuencias de los genes no son simples caracteres y pueden en la alineación y análisis implicar una incongruencia en el árbol, existen perdidas de genes que hacen posible mejor un análisis por HGT.

El argumento de este método depende de manera decisiva en el reconocimiento preciso de secuencias ortólogas divergentes (secuencias homológas que se han separado por el proceso de especiación, cuando la especie diverge en dos especies). No es fácil hacer el esboso completo de

una familia, especialmente entre proteínas cuya estructura es compleja, pero en los genes microbianos que son a menudo estructuralmente simples es posible, el trazado de familias ortólogas se puede realizar mediante el uso de herramientas basicas de alinación de secuencias, realizando una comparación entre las proteínas traducidas conceptualmente[11].

Resumiendo, éste método se basa en la identificación de genes homólogos en genomas compartidos por linajes filogenéticos disjuntos y su ausencia en los parientes cercanos (en uno o ambos linajes). Reconociendo que la divergencia rápida de una secuencia puede inducir a error a la identificación de HGT[36].

Para la tarea de arriba, la herramienta básica Local Alignment Search (**BLAST**) encuentra regiones de similitud entre las secuencias locales. El programa compara las secuencias de nucleótidos o proteínas a bases de datos de secuencias y calcula la significación estadística de los partidos.[35]

3.3.2. Con base en la composición atípica

Los métodos basados en la composición atípica se basan en las características fenotípicas entre diferentes genomas divergentes. Ellos se centran principalmente en el sesgo del codón, en el desvio del contenido GC y sesgo en la composición de nucleótidos[36].

Composicionalmente los genes atípicos pueden ser reconocidos fácilmente y asociarse a elementos de transposición o a una potencial ventaja selectiva, después de la transferencia horizontal. Esto ha dado lugar a reclamaciones que a pesar de las complicaciones del sesgo de la cadena, hay "mejora" con el tiempo en la selección para el nivel de expresión. El análisis de composición evita la necesidad de utilizar la alineción de secuencias, inferencia filogenética y busqueda de base de datos, aunque los supuestos donantes permanecen en el anonimato.[11]

En genética se llama marco abierto de lectura (siglas ORF del inglés open reading frame) a la secuencia de ADN comprendida entre un codón de inicio (AUG) de la traducción y un codón de terminación, descontando las secuencias que corresponden a los intrones en caso de haberlas. Entonces, dado que un ORF es una ventana de lectura de la secuencia dada, entonces los ORF pueden diferir de manera estadísticamente descriptible de acuerdo a su composición de bases, radio de purina y pirimidina, frecuencia de oligonucleótidos y uso de codón, desde regiones no codificantes o desde otros ORFS en el mismo o diferente genoma.

Koski[12] evaluó la evidencia filogenética de los orígenes de los genes compositivamente atípicos en E. coli K12. Se identificaron 2728 genes de E. coli que tienen "posiciones ortólogas" en genomas alineados esencialmente de Salmonella typhi, y se encontró que el 18 % de los genes compositivamente atípicos en las primeras tienen una posición ortóloga en S. typhi, mientras que el 22 % de los típicos ORFs, no. Los genes altamente expresados son evolutivamente estables a esta profundidad, pero 15 de 24 posicionales ortólogos cuyos árboles no muestran evidencia de HGT, son compositivamente atípicos, mientras que 12 de 25 ORFs son filogenéticamente de composición típica.

En este caso al menos, modelos simples basados en estadísticas de rendimiento de composición, inaceptablemente presentan altas tasas de falsos positivos y falsos negativos, evaluadas contra los árboles.

3.3.3. Con base en las diferencias del contenido genómico entre parientes cercanos

Parece que los microbios pueden estar estrechamente relacionados, debido al grado de divergencia en las secuencias ortologas. Aunque, difieran significativamente en el contenido ORF. Esto quiere decir que cada ancestro común del genoma debe tener todos los contenidos ORF de sus descendientes. En otras palabras, el origen del arbol genealógico (el último antepasado común universal LUCA) ha sido 'monolítico Leviathan', un enorme genoma. Jeffrey Lawrence ha argumentado en contra de esta posibilidad, debido a que un número grande de genes no puede ser mantenidos de forma selectiva en una población de tamaño razonable [13].

Tal vez el ejemplo más notable hasta la fecha se produjo por los genomas secuenciados E. coli que representan cepas O157: H7 [11] y K12, los cuales divergieron de un común antepasado $4,5 \times 10^6$ años. El primero contiene 1387 genes no se encuentran en K12, mientras que recíprocamente K12 tiene 528 genes que no están en O157: H7.

Muchos de los genes de cepas específicas se producen en racimos que presentan la base de la composición atípica. Por el contrario, dos cepas de Helicobacter muestran un 6-7 % de diferencias del contenido de genes a pesar de un grado cuatro veces mayor de divergencia en las secuencias, mientras que los genomas de cuatro Clamidias (parásitos intracelulares obligados que presumiblemente tienen pocas oportunidades para el contacto con otros procariotas) son tanto como 12 veces tan divergentes como las dos cepas de E.Coli, pero muestran poca evidencia de HGT.

En este método podemos concluir que el éxito de la HGT puede correlacionarse con la proximidad filogenética o el grado de relación genética, aunque la importancia y generalidad posible de los grados observados de especificidad siguen sin estar claros. De hecho, hay que reconocer que si bien estos métodos funcionan rápido, no son totalmente fiables, pues la similitud entre un gen en dos especies diferentes se pueden explicar por una serie de fenómenos extras de la HGT. Por ejemplo, la variación de la tasa de evolución puede conducir a resultados engañosos en la identificación de genes HGT, tanto como falsos positivos y falsos negativos [36]

3.3.4. Con base en incongruencias entre árboles filogenéticos

Las opiniones están divididas sobre si la HGT debe evaluarse contra un árbol, o si los métodos "sustitutos" pueden, individualmente o conjuntamente proporcionar una investigación más segura.

Recordemos que la ortología es un concepto basado en arboles. Si se infiere correctamente, un árbol estadísticamente bien soportado de una familia de ortólogos es topológicamente incongruente con la de un segundo árbol, entonces las familias tienen historias genealógicas incompatibles.

El caso de la HGT se fortalece en medida que podamos descartar explicaciones alternativas, incluyendo paralogía, artefacto metodológico, y sobre interpretación de subárboles estadísticamente débiles. Si la HGT no es demasiado penetrante, los genes parálogos deberían ser más fácil de identificar en el crecimiento de bases de datos genómicos. Los métodos filogenéticos han madurado hasta el punto que, uno podría legítimamente dudar de si el grado de incongruencia realmente observada entre los árboles de un sólo gen podría ser completamente artificial.

La falta de métodos indirectos para encontrar un conjunto común HGT proporciona otra justificación para confiar en los árboles. De los cuatro métodos aplicados para el ORF del genoma de E. coli, sólo dos (composición atípica y modelos de Markov) identificaron el mismo ORFs con más frecuencia que por casualidad, la mayoría coincidió en menos de las veces por casualidad[11].

La inferencia filogenética de conjuntos de datos genómicos (phylogenomics) puede ser automatizada, produciendo el "phylome" de cada análisis en el organismo, es decir el conjunto completo de filogenias de genes de un individuo o un organismo. Argumentos en contra de un enfoque puramente filogenético incluyen la dificultad de inferir árboles de alta calidad, de hecho, existe preocupación sobre los modelos y la posibilidad de que la HGT ha sido tan penetrante que la señal vertical se ha perdido. Por ejemplo en E. coli sus secuencias registran una historia de pérdida recurrente. Por lo tanto, estos linajes han sufrido episodios de mutación alta y las tasas de recombinación.

Se puede decir que el análisis filogenético es una buena opción para investigar la ocurrencia de HGT porque sigue siendo la única manera confiable para inferir eventos históricos de secuencias de genes; En consecuencia, los árboles filogenéticos incongruentes entre diferentes familias de genes será causado por HGT, sin embargo filogenias conflicto puede ser el resultado de cualquiera de los artefactos de reconstrucción filogenética, HGT o paralogía no reconocida. [36]

Capítulo 4

Genes transferidos horizontalmente y su detección como anomalías

En el afán de mostrar cuáles genes han sido transferidos horizontalmente, se expusieron en capítulos anteriores varias técnicas y métodos realizados por algunos investigadores. Esta cuestión también puede ser atacada por la teoría de detección de anomalías en secuencias genómicas. Una anomalía se entiende como una irregularidad, anormalidad o falta de adecuación a lo que es habitual, lo cual cumple perfectamente con el objetivo de distinguir genes no habituales en una secuencia genómica. Entonces un problema de **detección de anomalías** puede ser declarado como un problema de clasificación de dos clases: dado un elemento de algún espacio, clasificarlo como normal o anormal. Existen terminologías diferentes para diferentes aplicaciones, tales como "detección de novedad o sorpresa", "detección de falla", "detección de valores atípicos" [14].

Inclusive, la detección de anomalías se considera como uno de los requisitos fundamentales para una buena clasificación o identificación del sistema, ya que a veces los datos de prueba contienen información acerca de los objetos que no se conocían en el momento de entrenar el modelo. Esta detección puede basarse en métodos estadísticos o redes neuronales, con aplicaciones a procesamiento de señales, reconocimiento de patrones, minería de datos y robótica. [15].

La detección de anomalías se puede realizar de acuerdo a la clase de metodología usada. El siguiente esquema muestra la clasificación de algunos tipos de aproximaciones[16]:

Si se utiliza una metodología de aprendizaje supervisado, específicamente para la distinción entre genes horizontales y genes nativos, el sistema se entrena con vectores que describen genes nativos y vectores identificados como HGT, con el objetivo de aprender una función del espacio de características con etiqueta (HGT o NG). Una vez que el sistema esta entrenado y alcanza un bajo error, nuevas secuencias cuyo origen se desconocen se presentan al sistema. El sistema emite una decisión que puede ser NG, HGT, o una probabilidad de ser HGT. Hay que tener en cuenta que en esta metodología, se requiere conocimiento externo, es decir, un supervisor el cual tiene que decidir si una secuencia dada es HGT o NG, al menos en la fase de entrenamiento [18].

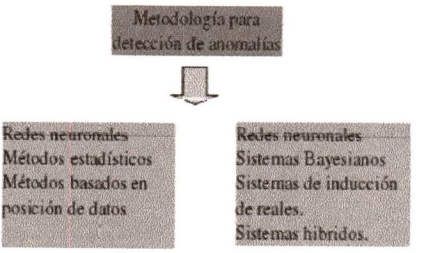

Figura 4.1: Metodologías para la detección de anomalías

Esta metodología se usa en las redes neuronales [15, 36]: Perceptron multicapa, Máquinas de soporte vértical, Teoría de resonancia adaptativa, Función de base radial, Hopfield, etc. Específicamente, un modelo de red neuronal se aplica para distinguir clases de genes, y para detectar genes extraños en [18]. En [19] una clase de soporte de máquinas de vectores es capáz de clasificar correctamente las secuencias nunca antes vistas como HGT cuando las etiquetas fueron asignadas previamente a los genes de formación.

Otras investigaciones sobre la detección de anomalías utilizan métodos estadísticos. Los métodos estadísticos se basan principalmente en el modelado de datos en función de sus propiedades estadísticas y el uso de esta información para estimar si una muestra de la prueba viene de la misma distribución o no. Las técnicas utilizadas varían en función de su complejidad.

El método más simple puede estar basado en la construcción de una función de densidad para los datos de una clase conocida. Estos enfoques a veces tienen limitaciones en cuanto a la elección de los parámetros y el ruido presente en los datos [15]. Algunos métodos paramétricos y no paramétricos, se muestran en la figura 4.2.

Podemos estudiar en que consiste cada uno de los métodos estadísticos, tanto paramétricos y no paramétricos mostrados en el esquema anterior, pero para nuestro caso, toma mayor interés el método con aproximación no paramétrica: **Estimación del vecino más cercano**. Existen varias aplicaciones y variaciones acerca de este método, pero la esencia es la misma; clasificar elementos de distintas clases en base a una métrica entre vecinos con respecto a las características de cada elemento.

Por ejemplo; Hellman utiliza el vecino más cercano como clasificador para rechazar los patrones con mayor riesgo de estar mal clasificados o al detectar fallos de rotor utilizando de vecindad una elipse como lo realizó Guttor. Algunos otros métodos proponen modificar el cálculo de la vecindad para determinar una anomalía de la siguiente manera: la distancia entre el patrón de prueba y su vecino más cercano en los datos de entrenamiento se encuentra junto con la distancia del vecino y su vecino más cercano. El cociente entre las dos distancias es una indicación de anomalía. La distancia euclídea se utiliza para este propósito.

También Jiang propuso un algoritmo de agrupamiento de dos fases para la detección de va-

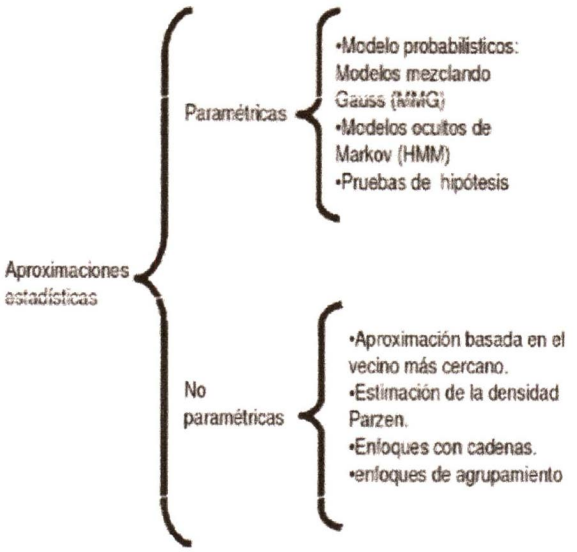

Figura 4.2: Aproximaciones estadísticas, métodos paramétricos y no paramétricos

lores atípicos basado en una modificación de k-medidas y un árbol de expansión mínima (MST). El algoritmo de k-medidas se modifica para calcular la distancia mínima entre cualquier par de centros de cúmulos. Si la distancia de un patrón y su cúmulo más cercano es mayor que esta distancia, entonces el patrón se asigna a un nuevo grupo. Incluso puede ser un modelo muy simple como el que realizó Yang aplicado a la clasificación de documentos.

El algoritmo es muy simple, cuando llega un nuevo documento, se compara con todos los documentos disponibles. Si su vecino más cercano en su pasado tiene una puntuación de similitud por debajo de un umbral, entonces el documento se etiqueta como novedoso y crea su primer historia de lo contrario se etiqueta como antigüo y se añade a la historia [15].

Hasta el momento hemos mostrado algunas de las metodologías para la detección de anomalías. Tomando algunas ideas se propone un método de identificación de genes por transferencia horizontal dentro del campo de algorítmos no supervisados, específicamente con la técnica del vecino más cercano, antes de comenzar con nuestros supuestos hay que destacar que la mayoría de estas investigaciones tienen que hacer especificaciones sobre los datos dependiendo de su naturaleza, es decir, determinar las características que se rescatarán de cada patrón o dato, para posteriormente realizar la identificación o clasificación.

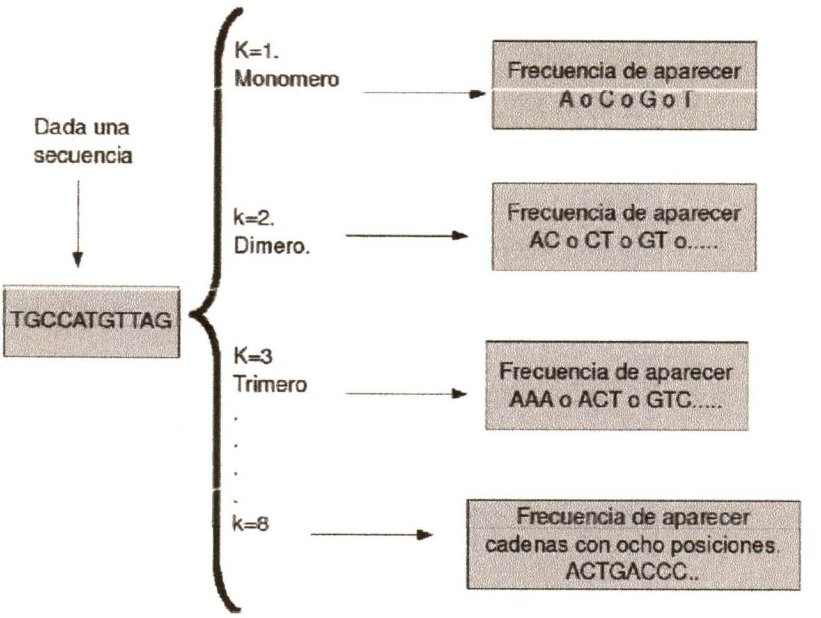

Figura 4.3: Visualización de los k-meros en una secuencia determinada.

4.1. Supuestos para la identificación del transporte horizontal de genes mediante detección de anomalías.

Para buscar los genes que son transferidos horizontalmente, debemos suponer que en una secuencia dada existen otros genes nativos, los cuales son transferidos verticalmente por sus ancestros denotados por NG en este caso los HGT son aquellas anomalías presentes en la secuencia según la definición que se dió en el capítulo 4.

Dadas las consideraciones anteriores debemos definir en que espacio de características tomaremos cada secuencia, y en base a ésto realizar el análisis correspondiente, de acuerdo a las propiedades que debe cumplir un espacio de características mostrado en el capítulo I y haciendo una semejanza con la firma genómica, podemos tomar dos tipos de espacios de características: el espacio de características composicionales (CFS) y el espacio de características organizacionales(OFS).

Para el **espacio de características composicionales** utilizaremos la idea de identificar la frecuencia relativa de un patrón en un genoma predeterminado, es decir, ubicar que tanto se repiten ciertas subsecuencias con tamaño 1, 2, 3 y 4 dentro de cada secuencia. Notemos que para $k = 3$ la subsecuencia son los llamados codones, la Figura 4.3 facilita la comprensión.

En el caso del **espacio de características organizacionales** debemos recordar que éstos

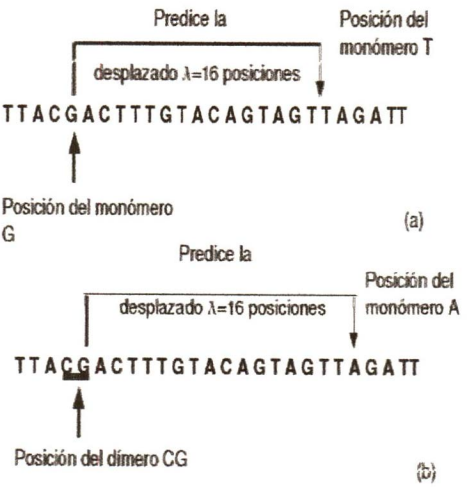

Figura 4.4: Correlación entre nucleótidos considerando características organizacionales

suponen que los nucleótidos no se presentan en forma aleatoria, sino que un nucleótido en una posición dada puede predecirse a partir de nucleótidos en otras posiciones, es decir, existe correlación entre ellos. Dado los buenos resultados nos centraremos en el uso de métodos de teoría de información, especificamente en el uso de la Función de Información Mutua (MIF), la cual es una herramienta muy útil en el calculo de correlaciones.

Para definir el espacio de características organizacionales se utilizará la MIF entre los k-meros (monómeros, dímeros, trímeros y tetrámeros) y las próximas bases a una distancia λ, para $\lambda = 1..,100$ formando una dimensión de espacio de características de 100 para cada k-mero, pues el desplazamiento de la función de información mutua va de 1 a 100.

En la Figura 4.4 puede observarse un caso específico para $\lambda = 16$ con casos particulares de monómero (a) y diméros (b), en este trabajo las subsecuencias consideradas son hasta tetrámeros con el corrimiento del λ desde 1 a 16, según la figura sólo tendríamos una componente del vector para el monómero y una para el dímero.

En terminos generales, cada elemento del espacio de características será un vector con dimensión correspondiente a cada atributo, ahora utilizando la metodología del vecino más cercano se calculará un centroide para cada una de las clases, aquí cada clase se supone esta formada por elementos que pertenezcan a OFS o CFS. Una vez que se identifique el centroide para la clase NG y HGT, se especifica una distancia umbral para construir un cúmulo, donde las secuencias que sobrepasen esa distancia (que esten fuera del cúmulo) no se consideran como parte de esa clase.

El algoritmo que se describe en esta contribución se basa en la distancia entre un gen y su

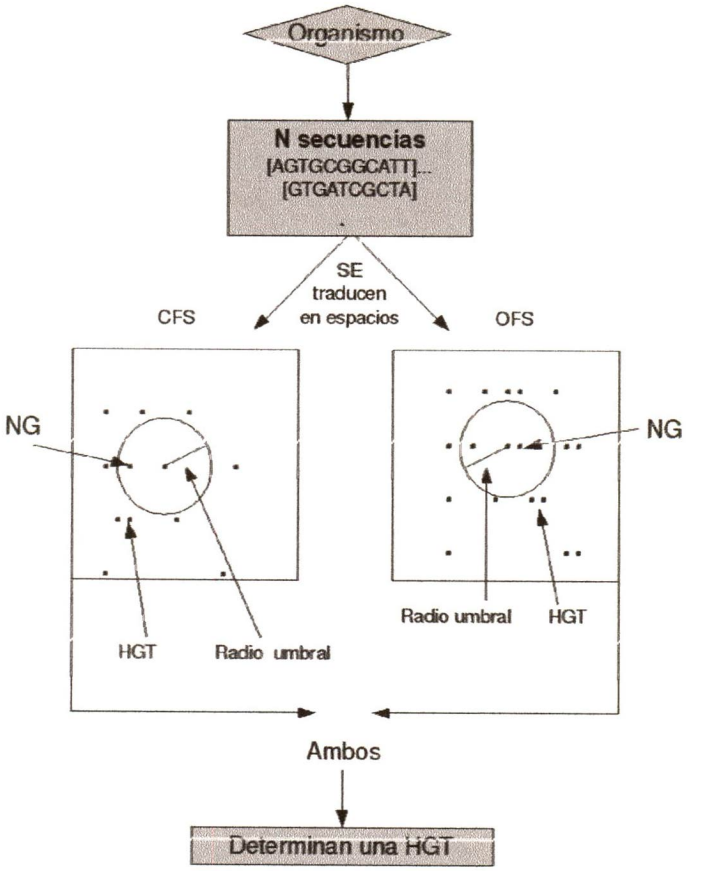

Figura 4.5: Metodología para la detección de HGT en ambos espacios de características

vecino más próximo de genes en un espacio de características. Si esa distancia no está dentro de un intervalo típico, entonces, el primer gen es un candidato a ser transferido horizontalmente.

La variante que se propone en este trabajo es no sólo detectar un HGT considerando la posición de los genes, sino que utilizar una métrica relativa entre un gen y otros genes, específicamente con sus vecinos, dando importancia a la posición que ocupan estos genes (puntos en el espacio de características). Si al determinar un umbral de distancia definido de acuerdo a cada genoma, un gen no se encuentra dentro de ese rango entonces dicho gen será un candidato a ser HGT. Recalcando que para que un gen sea HGT debe de satisfacer la condición de ser HGT en ambos espacios de características organizacionales y composicionales. La Figura 4.5 esboza nuestra intención.

En resumen, en este trabajo el algoritmo propuesto funciona de la siguiente manera: primero

cada secuencia es mapeada como un punto en cada espacio de características después se identifica su vecino más cercano mediante una distancia. Si esa distancia no está dentro de un cierto rango, entonces, el gen es un candidato a ser HGT. Si ese mismo gen aparece como un candidato para ser HGT en todos los espacios de características consideradas (organización y composición), entonces ese gen finalmente es marcado como HGT.

4.2. Exploración del problema

Una vez hecho los supuestos correspondientes, procedemos a la formulación matemática estableciendo la notación correspondiente para justificar el algorítmo, y después mostrar las consideraciones pertinentes en la fase experimental.

4.2.1. Formulación matemática.

Sea O_i la secuencia genómica del organismo i con una longitud l. Por ejemplo, la secuencia $O_i = [ATGGGTAC]$ tiene longitud $l = 8$ y si k denota el tamaño de ventana o subsecuencia a identificar en una secuencia, entonces tenemos que para $k = 1$ hay 4^1 posibles subsecuencias, para $k = 2$ hay 4^2 posibles subsecuencias, en general, si el tamaño de ventana es k, hay 4^k posibles subsecuencias.

Definamos a $A_k \subset N^{4^k}$ como el conjunto de las posibles combinaciones de secuencias con tamaño k. Es decir si $k = 1$; $A_1 = \{A, G, C, T\}$ y si $k = 4$; $A_4 = \{AGCT, ACTG, ACGT, ..., \}$.

Dado lo anterior podemos definir a CFS como sigue:

Para $k = 1$ (monómero), $CFS = \{X : X = [f_A, f_G, f_C, f_T]\}$ o $CFS = \{X : X = [f_a], a \in A_1\}$ donde f denota la frecuencia relativa de que aparezca cada base nitrogénada en la secuencia, así $X \in \Re^4$.

Para $k = 2$ (dímero), $CFS = \{X : X = [f_a]\}$, donde f representa la frecuencia relativa para la subsecuencias $a \in A_2$. Así a recorre todas las combinaciones de longitud 2, por tanto $X \in \Re^{16}$.

En general, sí $k = l$ entonces $CFS = \{X : X = [f_a]\}$, donde f representa la frecuencia relativa para la subsecuencias de tamaño l,..etc. Así $a \in A_l$ recorre todas las combinaciones de longitud l por tanto $X \in \Re^{4^l}$.

Por otra parte, para calcular el espacio de características OFS, utilizamos la función de información mutua como sigue:

Sea,

$$I_\lambda(X;Y) = \sum_a^{A_k} \sum_b^{A_1} P(a,b)_\lambda log \frac{P(a,b)}{P(a)P(b)},$$

donde $P(a,b)_\lambda$ representa la probabilidad de que en la presente posición un símbolo $a \epsilon A_k$ sea observado y que λ posiciones después se observe un símbolo b con $b \epsilon A_1$. $P(a)$ representa la probabilidad de que aparezca a con $a \in A_k$, análogamente $P(b)$ es la probabilidad de aparecer el símbolo b con $b \in A_1$.

Ahora, para $k = 1$ (monómero), tenemos que:
$OFS = \{X : X = [I_\lambda(x_j, y_{j+\lambda})], x_j, y_{j+\lambda} \in A_1$, con $X \in \Re^{100}\}$
Aquí, $y_{j+\lambda}$ representa la base en la posición j más y $I_\lambda(x,y)$ denota la función de información mutua, ente x (monómero) y y (la siguiente base) recorrida $\lambda = 1, ..., 100$.

Para $k = 2$ (dímero),
$OFS = \{X : X = [I_\lambda(x_j, y_{j+\lambda})], x_j \in A_2$ en la posición j-ésima, $y_{j+\lambda} \in A_1 X \in \Re^{100}\}$
donde $\lambda = 1, ..., 100$.

En general,

$OFS = \{X : X = [I_\lambda(x_j, y_{j+\lambda})], x_j \in A_k$ en la posición j-ésima, $y_{j+\lambda} \in A_1 X \in \Re^{100}\}$
donde $\lambda = 1, ..., 100$.

Concluyendo CFS esta formado por vectores de dimensión 4, 16, 64 y mientras que OFS estan formados por vectores de dimensión 100.

Finalmente para calcular la similitud entre secuencias definamos un métrica correspondiente:

Sea HTG el conjunto de genes que fueron horizontalmente transferidos y NG el conjunto de genes nativos, inicialmente estos conjuntos son vacíos. Sea N el número de genes.

Para OSF ,
$\mu_{CFS} = \frac{1}{N} \sum d(i,j)$ representa la distancia promedio entre cada gen y su gen cercano en OFS $d(i,j)$ es la distancia entre cada gen i y su gen cercano, y σ la desviación estandar.

Así que
$i \in NG$, si $\mu_{CFS} - \sigma_{CFS} \leqq d(i,j) \leqq \mu_{CFS} + \sigma_{CFS}$
$i \in HGT$, si $\mu_{CFS} - \sigma_{CFS} \nleqq d(i,j) \nleqq \mu_{CFS} + \sigma_{CFS}$

Análogamente para OFS,
$\mu_{OFS} = \frac{1}{N} \sum d(i,j)$ la distancia promedio entre cada gen y su gen cercano en OFS, donde $d(i,j)$ es la distancia entre cada gen i y su gen cercano. Con σ como desviación estandar.

Así que
$i \in NG$, si $\mu_{OFS} - \sigma_{OFS} \leqq d(i,j) \leqq \mu_{OFS} + \sigma_{OFS}$
$i \in HGT$, si $\mu_{OFS} - \sigma_{OFS} \nleqq d(i,j) \nleqq \mu_{OFS} + \sigma_{OFS}$

4.3. Algoritmo: Identificación de transferecia horizontal de genes por detección de anomalías

El algoritmo conocido como la identificación HGT por la detección de anomalías a través de la distancia al vecino más cercano, identifica genes en un espacio de características cuya distancia más cercana al gen vecino no es típica (de acuerdo a una medida promedio determinada para cada organismo) y los asocia a la lista de candidatos HGT identificándolos como anomalías, considerando métricas euclideanas.

Denotamos a **HGTiADnnd** como el algoritmo de identificación horizontal de transferencia de genes por detección de anomalías y presentemos su estructura como sigue:

1. Para un organismo O, tómese un gen O_i representado como secuencia de símbolos i sobre el alfabeto $S = A, G, C, T$.
2. Defina a F como un mapeo del espacio de secuencias (SS) hacia el espacio de características composicionales (CFS) y calcule:
   - CFS como la frecuencia relativa de no sobreponer subsecuencias de tamaño τ, de tal forma que la dimensión de CFS es $N = 4^{\tau}$.
3. Inicialice a la lista de genes que son identificados con HGT en CFS: $HGT_{CFS} = []$
4. Para cada gene representado como i en SS:
5. Obténgase esta representación en $CFS : C_i = F(i)$
6. Obtenga la matriz de distancias D para todos los puntos representados por genes en CFS
7. Obtenga la distancia promedio entre pares de vecinos cercanos $\mu(CFS)$ y la desviación estandar $\sigma(CFS)$
8. Para cada renglón de D representado por un gen O_i:
9. Identifica el gen más cercano a C_i y denotalo por $\eta(C_i)$, después calcúlese: La distancia entre C_i y $\eta(C_i), d(C_i, \eta(C_i))$
10. Si $d(C_i, \eta(C_i)) \ni [\mu(CFS) - \sigma'(CFS), \mu(CFS) + \sigma'(CFS)]$.
11. Agregar O_i a la lista HGT_{CFS}
12. Defina a G como un mapeo del espacio de secuencias (SS) hacia el espacio de características organizacionales (OFS) y calcule:
    - OFS como los primeros k desplazamientos de la función de información mutua de una secuencia dada i y llámele $m_i f(i)$, de tal forma que la dimensión sea k.
13. Inicialice a la lista de genes que son identificados con HGT en OFS: $HGT_{OFS} = []$
14. Para cada gene representado como i en SS:

15. Obténgase esta representación en $CFS : O_i = F(i)$
16. Obtenga la matriz de distancias D para todos los puntos representados por genes en OFS
17. Obtenga la distancia promedio entre pares de vecinos cercanos $(\mu(CFS))$ y la desviación estandar $\sigma(CFS)$
18. Para cada renglón de D representado por un gen O_i:
19. Identifica el gen más cercano a O_i y denótalo por $\eta(O_i)$, después calcúlese: La distancia entre O_i y $\eta(O_i), d(C_i, \eta(O_i))$
20. Si $d(O_i, \eta(O_i)) \ni [\mu(OFS) - \sigma'(OFS), \mu(OFS) + \sigma'(OFS)]$.
21. Agregar O_i a la lista HGT_{OFS}
22. Finalmente, $HGT = HGT_{OFS} \bigcap HGT_{CFS}$ representa a los genes adquiridos lateralmente

4.3.1. Observaciones sobre el algoritmo

Nuevamente, recordemos que el espacio de características de composición (CFS) se refiere a la frecuencia relativa de cada una de las subsecuencias 4^τ de longitud τ, logrando que no se sobre encimen las ventanas, mientras que el espacio de características de organización (OFS) se definen en términos de la función de información mutua para una determinada secuencia o gen. Notemos que se define la distancia promedio para cada organismo en cada espacio de características, por tanto cambiará de acuerdo a la estructura del organismo y además para definir los rangos de intervalo, se utiliza a $\sigma'(.)$ que indica el valor de la pendiente para $\sigma(.)$ es decir, cuando cambia la $P(HGT)$. De acuerdo a los supuestos que realizamos, si $\sigma'(.)$ es muy grande, el intervalo crece y entonces el algoritmo no puede identificar los HGT, si al contrario $\sigma'(.) \to 0$ el intervalo se reduce y todos los puntos estan fuera del rango provocando que todos sean catalogados como HGT. Esto refuerza la idea de dado que la distribución de genes varía se debe calcular los rangos independientemente para cada espacio de características, es decir, para CFS habrá cuatro valores distintos para un misma secuencia dependiendo del kamero y para OFS tendrán igual 4 valores distintos de σ',

Debemos aclarar que un punto en un espacio dado se declara una anomalía si se aísla. Es decir, una anomalía, no se puede definir en términos estadísticos de primer orden y de segundo, pero si en términos de relaciones espaciales que son más relevantes para capturar eventos anormales[23]. Actualmente en los algoritmos hay una desventaja, pues no es posible obtener un error de reconocimiento 0, si la función que describe perfectamente la secuencia de bases es desconocida. Así, algunos falsos positivos (FP) como algunos falsos negativos (FN) pueden estar presentes en los resultados.

Si en el algoritmo de HGTiADnnd se considerará que la distancia entre un gen y su gen más cercano fuera inferior a un umbral en lugar de incluirla en un intervalo, algunos FN se presentarían. Esto sucede cuando dos o más secuencias se transfieren desde el mismo organismo, lo que significa que las secuencias presentan características similares y, por tanto, se encuentra cerca

una de la otra en los espacios de características. Como son secuencias similares representadas en puntos cercanos en los espacios de características, ninguno de ellos se aislaría en el espacio de características. La solución obvia sería comprobar la distancia no sólo la del vecino más cercano, si no a los r mas cercanos. Sin embargo esto sólo desplazaría la tasa de FN.

El método a seguir para reducir el número de FN en el algoritmo propuesto también va en la dirección de encontrar la distancia correcta entre pares de genes más próximos a clasificar entre HGT y NG. Por tanto, en el algoritmo propuesto considera la distancia típica entre un gen y su vecino más próximo. Para cada gen i representado como un punto en el espacio F de características, se identifica su punto más cercano j (representando también un gen). Ahora, se obtiene la distancia media $\mu_G(F)$ para todos los genes i en el genoma G y su más cercanos j genes en F, junto con la desviación estándar $\sigma_G(F)$. La práctica de éstas hipotesis se presentan en las simulaciones del siguiente capítulo.

Capítulo 5

Simulaciones y resultados experimentales

En base a los supuestos y metodología realizamos el siguiente análisis y simulación.

5.1. Simulaciones

Se recolectaron 29 genomas del sitio FTP en la página del NCBI (Centro Nacional de Información sobre Biotecnología) para usarse en el algorítmo propuesto, dentro de estos genomas hubieron 22 bacterias, 6 arqueobacterias y 1 cromosoma humano, en los cuales se consideraron sólo genes con más de 300 bases, la siguiente tabla muestra su caracterización. En la primer columna aparece el nombre del organismo en la segunda el locus (la ubicación dentro del ADN), en la tercera columna el código o abreviación del organismo, en la cuarta la cantidad de genes en el genoma y en la ultima muestra la taxonomía a la cual pertenece cada organismo.

Al aplicar el algoritmo, se obtuvieron los dos grupos de espacios de características: de composición (CSF) y de organización (OFS). La dimensión de los espacios de composición son 4, 16, 64 y 256 de acuerdo a cada $k-mero$ $(4^i, i = 1, .., 4)$, mientras que la dimensión de los espacios de organización es 100 para todos los casos, debido al desplazamiento de la función de información mutua desde 1 a 100.

La heurística de discernir grupos de HGT del organismo donante mismo, es que los genomas tienen diferentes distancias promedio entre pares de puntos más cercanos, que es $\mu_{G1}(F) \neq \mu_{G2}(F)$, donde $G1$ y $G2$ son diferentes genomas, y F es una función espacio. La figura. 5.1 muestra la distancia media y la desviación estándar para los organismos analizados, así como los mapas bidimensionales obtenidos por un mapa de auto-organización (ver anexo) para las ocho dimensiones de OFS, CFS, y las 16-dimensiones de ambos espacios de características.

De la fig. 5.1, se observa que la μ distancia media entre pares de genes más cercanos en espacios de características es diferente en todos los organismos. Por lo tanto, elegimos μ como un filtro para decidir qué secuencias genómicas son NG y que son HGT. Ahora, el parámetro de umbral específico θ se indica en el algoritmo 6 y se calcula como un rango. Si la distancia entre gen i y su más cercana gen j tanto en el genoma de G y F en función de espacio está en el intervalo $[\mu_G(F) - \sigma'_G(F), \mu_G(F) + \sigma'_G(F)]$, entonces i es NG, de lo contrario es HGT en el espacio de características F.

Cuadro 5.1: Análisis de los genomas

Organismo	locus	código	genes	Tax.
Aeropyrum pernix K1	NC_000854	apx	991	A
Agrobacterium tumefaciens C58	NC_003062	atf	1986	B
Aquifex aeolicus VF5	NC_000918	aae	600	B
Archaeglobus fulgidus DSM 4304	NC_000917	afg	1332	A
Bacillus subtilis str. 168	NC_000964	bst	2123	B
Borrelia burgdorferi B31	NC_001318	bbg	421	B
Chlamydia trachomatis A/HAR-13	NC_007429	ctr	590	B
Deinococcus radiodurans R1	NC_001263	drd	1829	B
Escherichia coli O55:H7 str. CB9615	CP001846	eco	2882	B
Haemophilus influenzae 86-028NP	NC_007146	hiz	1072	B
Helicobacter pylori 26695	NC_000915	hpy	968	B
Homo sapiens chr 1	NW_921350	Hs1	1295	E
Listeria innocua Clip11262	NC_003212	lin	1754	B
Lactococcus lactis Il1403	NC_002662	llc	1294	B
Mycoplasma genitalium G37	NC_000908	mgt	325	B
Mycoplasma pneumoniae M129	NC_000912	mpn	448	B
Mycobacterium tuberculosis H37Rv	NC_000962	mtb	1823	B
Nanoarchaeum equitans Kin4-M	NC_005213	nne	280	A
Pseudomonas aeuruginosa PA7	NC_009656	pae	3811	B
Pyrococcus abyssi GE5	NC_000868	pab	1098	A
Salmonella enterica Typhi Ty2	AE014613	sen	2292	B
Staphylococcus aureus MRSA252	NC_002952	sau	1545	B
Thermoplasma acidophilum DSM 1728	NC_002578	tac	888	A
Thermoplasma volcanium GSS1	NC_002689	tvc	879	A
Thermotoga maritima MSB8	NC_000853	tmt	1212	B
Treponema pallidum str. Nichols	NC_000919	tpd	689	B
Vibrio cholerae O1 str. N16961	NC_002506	vce	1164	B
Xylella fastidiosa Temecula1	NC_004556	xyf	1186	B
Ureaplasma urealyticum 10 ATCC 33699	NC_011374	uuy	419	B

0.8

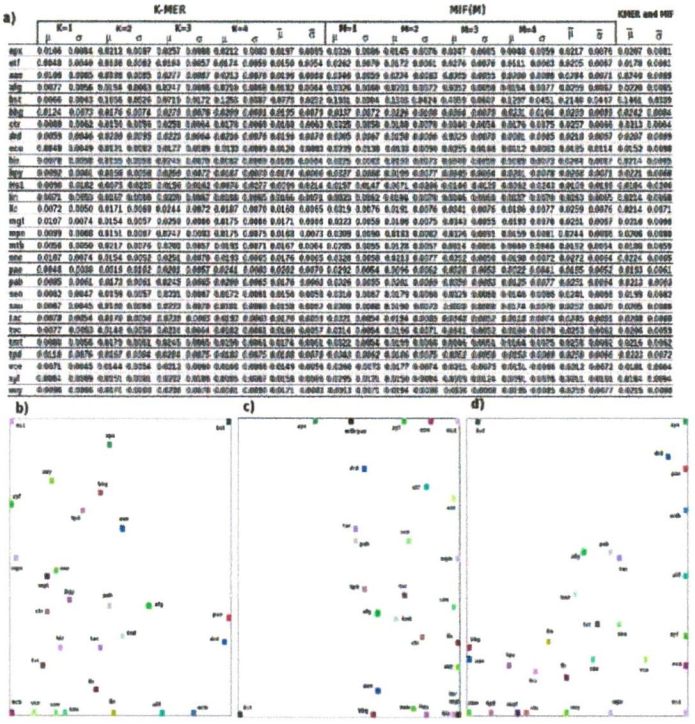

Figura 5.1: La distancia media (μ) y desviación estándar (σ) de los 29 genomas analizados. a) distancia promedio entre pares de puntos más próximos en los espacios de características de composición (kamero-1 a kamero-4) y el promedio de la distancia entre los pares de genes en los espacios de características organizacionales (MIF-1 a MIF-4). b) la organización de (SOM) para el SFC. c) SOM para OFS. d) SOM para OFS y CFS

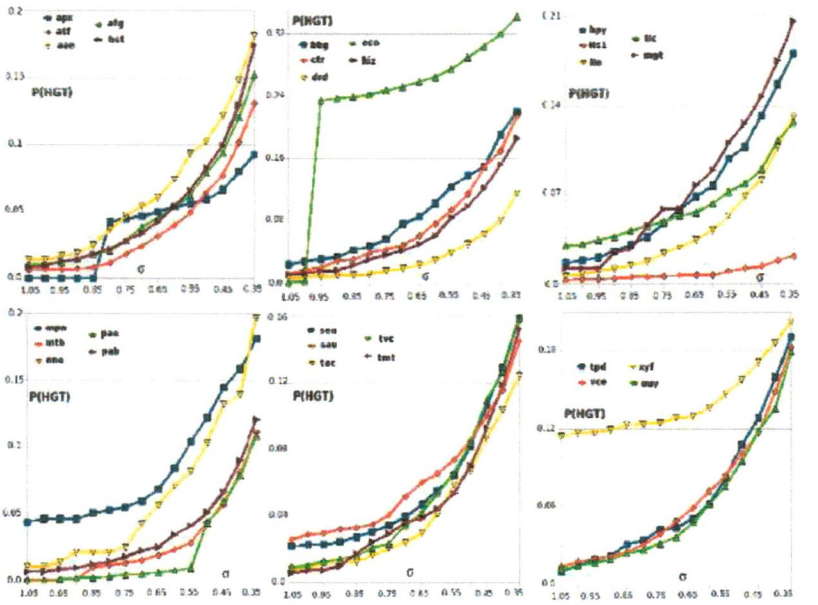

Figura 5.2: Frecuencia de la HTG

A medida que consideramos varios espacios de características (cuatro composicionales y cuatro organizacionales), un gen i se declara HGT si y sólo si:

$$\sum_{f\in[CFS,OFS]} \delta_f(i) = 0, \text{donde } \delta_f(i) = \begin{cases} 1 & \text{si } d(i,j) \in [\mu_G(f) - \sigma'_G(f), \mu_G(f) + \sigma'_G(f)] \\ 0 & \text{En otro caso} \end{cases} \quad (5.1)$$

Donde $\sigma'_G(F)$ es el valor para $\sigma_G(F)$ cuando la pendiente de la derivada $P(HGT)$ cambia.

Así que en ambos espacios, en el CFS y el OFS, se identificaron los genes más cercanos cuya secuencia se encuentra a una distancia no incluida en un intervalo típico, que se define en términos de la distancia media entre cada gen y su vecino más cercano en el espacio de características, dichos genes se consideraron candidatos HGT. En la Figura 5.2 podemos observar la frecuencia relativa de una HGT para diferentes valores de la desviación estandar $\sigma \in [0,35, 1,05]$ de acuerdo a el valor de μ_G la distancia media entre un gen y su vecino más cercano para todos los genes en el genoma de G. Hay que recalcar que esta distribución es en base al promedio de probabilidad de HGT en cada espacio de características y que la visualización que se presenta se agrupó de acuerdo al orden alfabético para facilitar el análisis.

5.1.1. Observaciones

Visualmente, notamos una variación en la P(HGT) para algunos organismos, aunque P(HGT) es definida positiva en σ, el incremento no tiene la misma proporción para todos los organismos estudiados, incluso para un mismo organismo cambia drásticamente su comportamiento. Por ejemplo en:

Aeropyrum pernix P(HGT) es casi nula para $\sigma \leqslant 0{,}85$ y de inmediato alcanza una altura de 0,042.

E. coli P(HGT) vale casi cero en en un intervalo muy corto de σ y se dispará para valores de $\sigma = 0{,}95$ hasta 0,238

La interpretación que podemos dar a esta conducta es la siguiente: Esto se ocasiona por que el número de genes que quedan fuera del intervalo típico es mayor que lo que se espera durante unos puntos distribuidos al azar en el espacio de características, así que el salto de probabilidad determinado por σ define al intervalo llamado σ'. El intervalo de la distancia típica (NG) entre un gen y su vecino más cercano en un FS se determina por $[g - \sigma'_G, g + \sigma'_G]$.

Ahora, qué significa que en la grafica algunos organimos tengan valores elevados de de P(HGT), incluso para los grandes valores de σ, En el caso de :

Mycoplasma pneumoniae Aquí $P(HGT) = 0{,}048$ para $s = 1,05$. El incremento en P(HGT) para ese organismo permanece casi constante hasta que σ alcanza un valor de 0,6. Se ha documentado que HGT está presente en Eukarya, pero es en general infrecuente [24].

Cromosoma humano 1 Se observa que la P(HGT) es baja y aumenta muy lentamente, evidenciando que el algoritmo esta realizando una detección correcta de acuerdo a la teoría existente.

5.2. Resultados experimentales

En la tabla 7.1 mostramos los resultados sobre la detección de HGT en los genomas analizados. Se muestra σ' para cada genoma, así como el número de genes detectados como HGT y su P(HGT) (frecuencia relativa de HGT dentro de un genoma). Se observa que el genoma con mayor porcentaje de HGT entre los organismos analizados, es E. coli con un 24 % , después con un 15 % se encuentra Chlamydia trachomatis y Vibrio cholerae. Por otra parte los organismos con menor P(HGT) son Agrobacterium tumefaciens, Archaeglobus fulgidus, cromosoma humano 1, y Thermoplasma volcanium, todos con 2 %. Estos resultados difieren con otros algoritmos de detección de HGT, especificamente con los resultado que expone Garcia [6] en la tabla 2 de ese mismo artículo.

El alto contenido de HGT en E. coli contrasta con algunos de los algoritmos existentes, ya que estiman que el contenido de HGT en este genoma es alrededor del 12 % [6]. Esta diferencia no significa necesariamente que que nuestro modelo esta determinando a un dato anómalo sin serlo (existencia de un falso positivo) o bien que que otros métodos no están detectando a los

Cuadro 5.2: Frecuencia relativa de HGT de los 29 organismos y su clasificación de acuerdo a la categoria funcional

Code	σ'	No.HGT P(HGT)	%Info	%cell	%Meta	%Poorly
apx	0.80	42 (0.0424)	13	13	13	63
atf	0.75	35 (0.0181)	17	42	8	33
aae	0.80	22 (0.0367)	6	9	15	69
afg	0.80	27 (0.0203)	9	14	14	63
bst	0.40	277 (0.1305)	12	48	32	7
bbg	0.70	32 (0.0760)	14	8	15	61
ctr	0.45	88 (0.1492)	21	17	33	30
drd	0.40	147 (0.0804)	27	11	22	40
eco	0.95	678 (0.2353)	13	27	13	47
hiz	0.55	90 (0.0840)	28	13	33	27
hpy	0.55	95 (0.0981)	14	14	28	44
Hsl	0.30	32 (0.0247)	41	32	14	14
lin	0.65	61 (0.0348)	2	1	6	90
llc	0.45	116 (0.0896)	26	6	41	27
mgt	0.80	15 (0.0462)	22	15	44	19
mpn	0.55	47 (0.1049)	22	19	26	34
mtb	0.45	224 (0.1228)	0	33	17	50
nne	0.75	7 (0.0250)	10	0	0	90
pae	0.45	229 (0.0601)	25	0	25	50
pab	0.50	48 (0.0436)	12	0	15	74
sen	0.50	189 (0.0825)	32	21	41	6
sau	0.75	63 (0.0408)	18	13	47	22
tac	0.65	26 (0.0293)	14	7	18	61
tvc	0.75	20 (0.0228)	32	37	18	13
tmt	0.45	113 (0.0932)	10	6	43	41
tpd	0.60	43 (0.0624)	15	15	25	46
vce	0.40	173 (0.1486)	10	8	44	38
xyf	0.65	154 (0.1298)	18	22	40	21
uuy	0.70	15 (0.0358)	47	3	45	5

datos anómalos(falsos negativos). En la subsección de validación del algorítmo discutiremos esta observación.

También es un tema de controversia que los genomas archaebacterial tiendan a presentar un mayor contenido de HGT, ya que estan inmersos en un ambiente donde las condiciones de estrés favorecen la aceptación de genes extraños [24]. Sin embargo, cuando se aplica el HGTiADnnd para detectar la HGT, las arqueas presentan un contenido más bajo que el observado en las bacterias. En el archea taxón, el contenido promedio de los genes de origen extranjero es del 3.1 %, mientras que en las bacterias es del 8.9 %.

5.2.1. Análisis

Recordemos una de la hipótesis sobre la transferencia horizontal de genes donde se establece que algunos genes tienen mayor participación en este proceso, de acuerdo a su funcionalidad dentro del organismo, en base a esto se presenta el porcentaje de participación de algunos genes con ciertas características, es decir dependiendo la funcionalidad que realizen, de manera semejante a como se estable en el trabajo de (Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes)[6] podemos establecer cuatro grupos funcionales: Info (El cual consiste en genes informacionales), cell (aquellos que participan en procesos celulares), poor (los que tiene muy poca participación en los procesos), y meta (los que participan en proceso del metabolismo).

La distribución funcional del HGT detectada no presenta una clara preferencia sobre una de las tres categorías. Para los genes informativos, Xylella fastidiosa es el genoma de mayor porcentaje con un 47 %, seguido por el cromosoma humano 1 con un 41 %, Salmonella enterica, y Thermoplasma volcanium ambos con un 32 %. Para los genes relacionados con los procesos celulares, el porcentaje más alto se observa en Bacillus subtilis con un 48 %. Los genes relacionados con el metabolismo están más presentes en Staphylococcus aureus con un 47 %, y Ureaplasma urealyticum con el 45 %. Para los porcentajes más bajos, Mycobacterium tuberculosis no presenta información relacionada con HGT. Ésto denota uno de dos procesos. Un proceso de mejora se ha producido en los genes, mediante el cual los genes han adquirido modificaciones logrando parecerse más a las pautas generales que describen los genes nativos, o que nuestra resolución del método no es suficiente como para detectarlas. Sin embargo, otros métodos también reportan bajos porcentajes de genes relacionados con la información [24]. Equitans Nanoarchaeum, Pseudomonas aeruginosa, y abyssi Pyrococcus no presentan HGT relacionada con los procesos celulares.

Entre los genomas analizados, la función más común en la detección de HGT, es la función del metabolismo y con menor frecuencia de transmisión son los genes de la categoria de información. Esto es coherente con la teoría existente, que establece que se debe esperar una transferencia relativamente menos frecuente entre genes relacionados con la información. Según autores esta baja probabilidad es causada por el hecho de que los genes informativos son generalmente miembros de sistema grande y complejo, mientras que los genes operativos no tienen estas funciones[10].

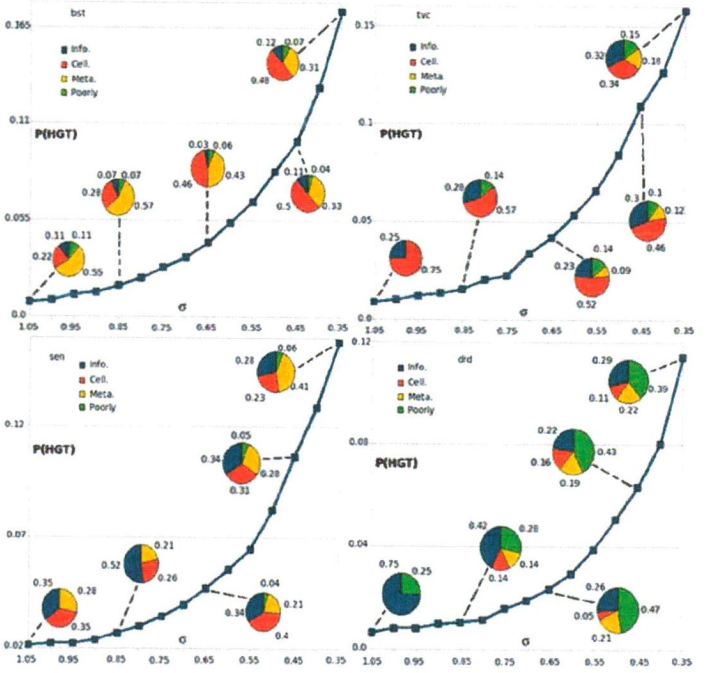

Figura 5.3: Distribución funcional, variando el valor σ

En la fig 5.3 se muestra la distribución funcional de las secuencias detectados como HGT para cuatro de los genomas analizados en función de σ.

De acuerdo a la construcción del algoritmo, los datos de entrada estan dados en una gran dimensión, pues determinan descripciones especificas de los genes y genomas, lo cual complica realizar una visualización de los genes que son HGT y los NG, pues se representan como puntos en espacios de dimensión 4,16, 64, 256 y 100.

Para tener una buena idea de los genes presentes en la distribución de esos espacios de características, es necesario realizar una una proyección en un espacio bidimensional. Para lo cual utilizaremos una proyección lineal denominada Mapeos Auto-organizados (SOM), en el anexo correspondiente se expone la metodología que utiliza ésta herramienta. Según nuestro problema, en [20] se presenta un análisis semejante a nuestro objetivo el cual consiste en visualizar la distribución de genes en los espacios de características de alta dimensión que hemos considerado.

5.3. Visualización de HGT y NG mediante SOM

La dimensión de los espacios de composición son 4, 16, 64 y 256 para $\tau = 1..,4$, mientras que para los espacios informativos, la dimensión es de 100, debido al número de desplazamientos

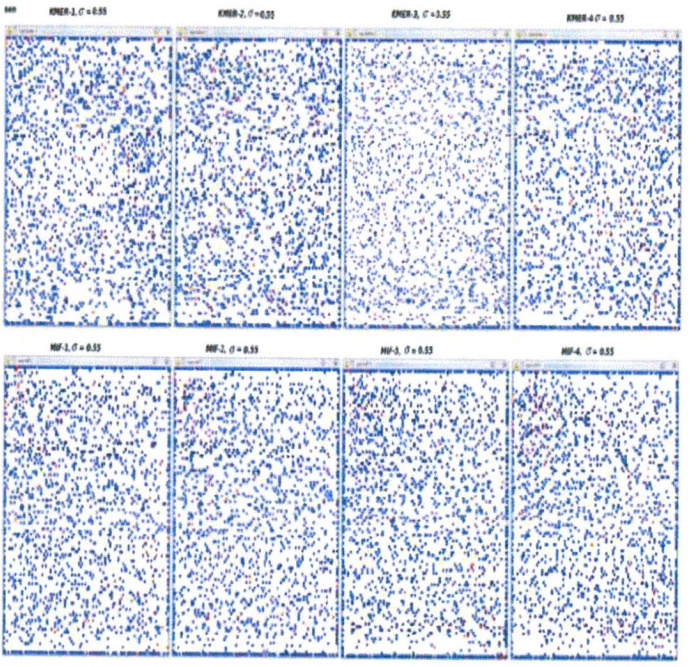

Figura 5.4: SOM para Salmonella enterica. En azul se muestran los NG y en rojo los HGT, de acuerdo a HGTiADnnd con $\times\sigma = 0,055$. Considerando los cuatro espacios de características composicionales y organizacionales

para el análisis de la función de información mutua. Hemos construido SOMs bidimensionales para la distribución de datos en los espacios de alta dimensión para visualizar cómo los genes están distribuidos.

Los mapas autoorganizados o SOM (Self-Organizing Map), también llamados redes de Kohonen son un tipo de red neuronal no supervisada, competitiva, distribuida de forma regular en una rejilla de normalmente dos dimensiones, cuyo fin es descubrir cumúlos de los datos introducidos en ella, vease anexo B.

Los genes de cada organismo se representan como puntos en un espacio de alta dimensión. Las coordenadas de cada gen se determinan por la frecuencia relativa de cada uno de los K-meros 4^{τ} en el gen, para el caso de CFS, mientras que para OFS, las coordenadas son dadas por la información mutua entre K-meros en la posición i y las bases en la posición $i+1$ hasta $i+100$.

Para cada uno de los organismos en la tabla 1, se obtuvieron ocho SOMs. Los cuatro primeros fueron para cada uno de los CFS: monómero dímero, trímero y tetrámero de frecuencias relativas. Los cuatro restantes se obtuvieron del OFS, es decir, mediante la MIF entre los k-meros ($k = 1, 2, 3$ y 4) y las siguientes bases 100. En el SOM obtenido, los genes identificados como HGT por HGTiADnnd están representados por cuadros rojos, y los NG se representan como cuadros

de color azul. En la 5.4 se muestran los SOMs de Salmonella enterica. Tenga en cuenta que no hay grupos de HGT y el NG en las proyecciones, ya que el método propuesto no se basa en la identificación de cluster. HGTiADnnd se basa en la identificación de genes cuya distancia a su vecino más cercano no es típica para el genoma que forman parte.

5.4. Validación del algoritmo

No hay una precisión absoluta acerca de cuáles genes son transferidos horizontalmente y cuáles son nativos, ya que no existe un registro puntual de los eventos de transferencia horizontal [24]. En términos computacionales, esto significa que no hay a priori una etiqueta asociada a los genes, por lo que es difícil de validar cualquier modelo de clasificación de una manera directa.

Con el fin de dar más pruebas de que HGTiADnnd es capaz de discernir entre las secuencias de nativos y extranjeros, se procede a comparar los resultados de nuestro modelo con los obtenidos por otros modelos, y también para comparar la HGT detectado por HGTiADnamnd que está presente en la la literatura biomédica. En primer lugar, se realizó un experimento en el que se crearon varios genomas artificiales para demostrar que HGTiADnnd es capaz de detectar secuencias extrañas cuando el postulado de la firma genómica es válido.

Se construyeron 100 genomas artificiales $A_1, ..., A_{100}$, cada uno con un número de genes entre 450 y 1200, y los genes con una serie de bases en el rango de [300..,5000]. Los genes fueron ge$\neq$ *radosporunamatrizdeMarkovdeorden4(veaseanexoA), ycadagendentrodecadagenomatuvounadiferente* se aplica una diferente matriz de Markov M_i. Con probabilidad $p = 0{,}01$, cada base de cada gen sufrió una mutación.

El siguiente paso consistió en simular la HGT entre genomas. El genoma B_i, $0 < i \leq 100$, fue construido con un porcentaje q de genes nativos (los genes del genoma de A_i). y el restante $(1 - q)$ porcentaje de genes seleccionados al azar de otros genomas $j \neq i$. En todos los casos, el $80\% < q \leq 95\%$, es decir, entre 5 y 20 % de los genes de los genomas de B_i fueron adquiridos por HGT. Esos genes transferidos a B_i de los genomas distintos de A_i constituyen anomalías en el contexto de A_i.

Pusimos a prueba la capacidad de detección HGT en un conjuto de genomas artificiales de HGTiADnnd y otros dos algoritmos de la literatura.

- El primero es el que se utiliza en [6], en donde un gen se declara HGT si su contenido de G + C (T) se desvía por $> 1{,}5\sigma$ del valor medio del genoma o si las desviaciones de G+C(1) y G+C(3) eran del mismo signo y al menos una fue $> 1{,}5\sigma$, donde σ es la desviación estándar.

- El segundo método es el que se propone en [19], en el que las plantillas de tamaño 8 se analizan en términos de relación a la frecuecia de las bases. La idea es calcular la frecuencia relativa de todos los tipos de plantilla en un gen, y luego comparar dicha frecuencia a la

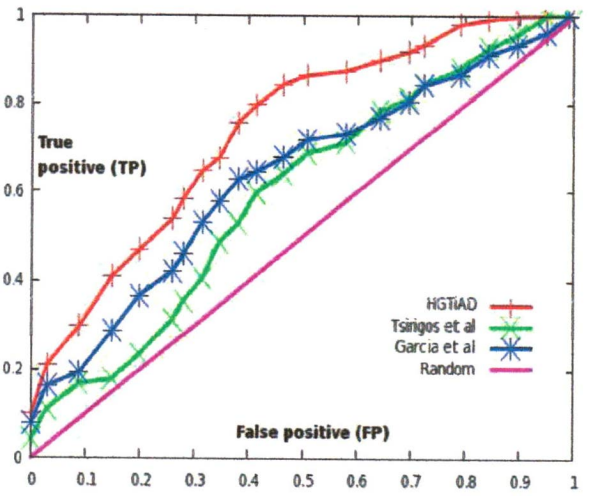

Figura 5.5: ROC para HGTiADnnd, comparación de algoritmos, un clasificador ideal sería presentar una tasa de 0 para un verdadero positivo y una tasa de 1 para un falso positivo.

observada en el genoma entero. La comparación se hace por cinco diferentes medidas, y un perfil de tipicidad dado.

Fue así como se comparó la capacidad de detección de nuestro algoritmo y el de los algoritmos mencionados.

Dado que se conoce con certeza el origen de las secuencias, hemos sido capaces de evaluar sobre la misma base de genes que están etiquetados (GN o HGT). La figura 5.5 muestra el promedio de ROC para los 100 casos. Se observa que HGTiADnnd logra un mejor rendimiento, ya que está más cerca de la curva de reconocimiento ideal. Nótese que, como las mutaciones estan presentes y las matrices de Markov puede ser similares, el error de reconocimiento será siempre mayor que cero.

Como resumen del reconocimiento, se obtuvo el ROC que se muestra en la 5.5. Tiene que hacerse notar de nuevo que se conoce con precisión absoluta si un gen es HGT o no, y por lo tanto, en muchos algoritmos se supone que los reconoce, como el presentado en esta contribución. A raíz de este hecho, podemos decir que nuestro algoritmo funciona mejor que otros, bajo el supuesto de que las bases en los genes son descritas por una matriz de Markov. Sin embargo, creemos que la capacidad de reconocimiento mostrado por HGTiADnnd pueden ser de interés para la detección de genes extraños.

Así, podemos decir que dependiendo de la característica del espacio y la función discriminante, los genes detectados como HGT pueden variar considerablemente. Por ejemplo, en el caso de Mycobacterium tuberculosis, la intersección de la HGT detectado por el método propuesto en [7] y los obtenidos por el método usado en [6] es el conjunto vacío. La razón de que difer-

entes algoritmos de detección de HGT identifiquen genes diferentes como verdaderos positivos es que cada algoritmo hace diferentes supuestos sobre el espacio de características relevantes y la función discriminante.

Hay muchos casos estudiados en la literatura que muestran una fuerte evidencia de HGT en varios organismos. Se reporta que los genes adquiridos por HGT son los responsable de algunas condiciones patogénicas identificadas en Salmonella según [8] y en [7] para Mycobacterium tuberculosis. Se presenta en la tabla 3 el número de HGT detectado bajo consideraciones biomédicas y algoritmos de alineación (Veyrier para el primer organismo y Lémire para el segundo), y su comparación con otros métodos de detección de HGT (CAI y HGTiADnnd).

Cuadro 5.3: La intersección entre HGT dectado por diferentes algoritmos tendiendo a ser muy pequeños

Mycobacterium tuberculosis(NC_000962)						
Algorithm	Veyrier	CAI	HGTiADnnd ($\sigma' = 0{,}45$)	veyrier ∩ HGTiADnndp	CAI ∩ HGTiADnndp	Veyrier ∩ CAI
No HGT	143	50	105	0	0	0
Salmonella enterica (AE014613)						
Algorithm	Lémire	CAI	HGTiADnnd ($\sigma' = 0{,}5$)	Lémire ∩ HGTiADnndp	CAI ∩ HGTiADnndp	Lémire ∩ CAI
No HGT	9	192	189	2	8	1

Capítulo 6

Conclusiones

Dado que cada genoma G se define como una colección de genes, podemos decir que algunos genes son nativos (N) y algunos son transferidos horizontalmente. Según las investigaciones encontrar un método que determine la clasificación no es tarea fácil, pues intervienen varios aspectos. Dentro de toda esa gama de investigaciones la propuesta de considerar a los genes horizontales como anomalías es motivante. La variación depende de cada atributo a considerar, como se mencionó en éste trabajo hay básicamente dos espacios que incluyen algoritmos de detección de HGT.

En el primero se consideran características de composición (contenido G-C, CAI, etc) y en el segundo basado en hipótesis de la existencia de una estructura de base en la secuencia genómica, es decir, las bases en una posición dada puede ser correlacionada con las bases anteriores. Recalcando la importancia de herramientas de teoría de información, como lo es la Función de Información Mutua de gran utilidad para calcular esta correlación.

De acuerdo a las hipotesis establecidas y comparadas con otros autores se presenta en este trabajo un algorítmo que primero identifica los dos grupos de espacio de características(organizacionales y composicionales), y después determinar el rango adecuado entre los genes más cercanos con el fin de discernir los genes nativos de los genes transferidos horizontalmente, utilizando herramientas computacionales que facilitan la interpretación de los resultados obtenidos de las simulaciones para comprender mejor los procesos biológicos en los genes ligados a esta investigación.

Podemos rescatar varios aspectos de este trabajo:

- Nuevo enfoque, La identificación de genes horizontales como un problema de detección de anomalías.
- Considerar no sólo atributos de composición, sino también introducir formalmente la importancia de los atributos de organización dentro del gen.
- La utilización de un algoritmo basado en análisis de distancias entre vecinos más cercanos, modificando el tradicional análisis de agrupación.

- Corroboración y discusión de algunas hipótesis y resultados presentados por otros autores, como por ejemplo en la participación de genes de específica categoria (operacionales, informativos, etc.)

Es importante notas que ningún método puede decirse que es mejor en la detección de HGT que otros, a menos que exista una certeza absoluta sobre el origen de las secuencias, sin embargo en términos generales y de acuerdo a la validación realizada se considera que este algorítmo es bueno para la identificación de genes.

Bibliografía

[1] Soberon Mainero, *"La ingeniería genética. La nueva biotecnología y la era de la genética"*, Ciencia para todos, Edit Fondo de Cultura Económica, 2005.

[2] G. Audesirk, y B. E. Byers, *"Biología: La Vida en la Tierra"*, 6a Ed. México: Prentice Hall, 2003

[3] Stryler, Jeremy N, Berg. John L. Tymoczco, *"Bioquímica"*, Reverte, 2008.

[4] J. Colin Sanderson, *"Understanding Genes and GMOS"*, World Scientific.

[5] Aristotelis Tsirigos, Isidore Rigoutsos, *"A new computational method for the detection of horizontal gene transfer events"*, Nucl. Acids Res. 2005. 33 No. 3 (922-933).

[6] García-Valle ´ S, Romeu A., Palau J. ' *"Horizontal Gene Transfer in Bacterial and Archaeal Comeplete Genomes"*, Genome Res, 2000. 10, 1719-1725.

[7] Veyrier F, Pletzer D, Turenne C, Behr MA. *"Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis"*, BMC Evolutionary Biology, 2009.9 196.

[8] Michael Hensel, Herbert Schmidt. *"Horizontal gene transference in the evolution of pathogenesis"*, Cambridge University Press, 2008.

[9] Philippe H, Douady CJ. *"Horizontal gene transfer and phylogenetics"*. Curr. Opin. Microbiol. 6, 2003, 498–505.

[10] Ravi Jain, Maria C. Rivera, and Jamees A. Lake, *"Horizontal gene transfer among genomes: The complexity hypothesis"* Molecular biological, Proc Natl Acad Sci USA 1999 Vol: 96:3801-3806. DOI: 10.1073/pnas.96.7.3801.

[11] Ragan M. *"Detection of lateral gene transfer among microbial genomes"*. Current Opinion in Genetics Development. 2001. 11:620–626.

[12] Koski LB, Morton RA, Golding GB: *"Codon bias and base composition are poor indicators of horizontally transferred genes"*. Mol Biol Evol 2001, 18:404-412.

[13] Glansdorff N: *"About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal"*. Mol Microbiol 2000, 38:177-185

[14] Gonzales Fabio, Dipankar Dasgupta, *"Neuro-Immune and Self-Organizing Map Approaches to Anomaly Detection: A Comparison"*, Division of Computer Science, Canterbury, UK, sept 9-11, 2002.

[15] Markou M, Singh S. *"Novelty detection: a review part 1: statistical approaches"*. Signal Processing 2003. 83: 2481-2497.

[16] A. Cervecera, A. Neme, L. Hernandez, *"Identification of horizontal gene transference by nearest neighbor based anomaly detection techniques"*,Helsinki University, Finland, 2011.

[17] Markou M, Singh S. 2003. *"Novelty detection: a review part 2: neural network based approaches"*. Signal Processing 2003. 83: 2499-2521.

[18] Wang HC, Badger J, Kearney P, Li M. *"Analysis of Codon Usage Patterns of Bacterial Genomes Using the Self-Organizing Map"*. Mol. Biol. Evol. 2001. 18(5):792–800.

[19] Tsirigos A, Rigoutsos I."A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes" Nucl. Ac. Res. 2005. 33, No. 12 3699–3707. doi:10.1093/nar/gki660

[20] Abe T, Kanaya S, Kinouchi M, Ichiba Y. *"Informatics for Unveiling Hidden Genome Signatures"* 13. 693-702. doi:10.1101/gr.634603

[21] F. Sánchez, J.L.Gutierrez. *"clasicos de biología matemáticas"*. Edit. siglo XXI editores, 2002, 32-34.

[22] Oskan Ufuk Nalbantoglu, *"Computational Genomic Signatures and Metagenomics"*, University of Nebraska, 2011.

[23] Pierce A. Benjamín, *"Genética, un enfoque conceptual"*. Edit Paramericana, España, 2009, 726-727

[24] Koonin E, Makarova K, Aravind L. *"Horizontal gene transfer in prokaryotes: Quantification and Classification"*. Ann. rev. Microbiology. 2001. 55, 709-42.

[25] Dalevi D, Dubhashi D,Hermansson M. *"Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. Bioinformatics"*. 2006. 22: 5, 517–522.doi:10.1093/bioinformatics/btk029.

[26] Azad R, Lawrence J. *"Detecting laterally transferred genes: use of entropic clustering methods and genome position"*. Nucl. Ac. Res. 2007. 35, 14 4629–4639. doi:10.1093/nar/gkm204.

[27] Bauer M, Schuster S, Sayood K. *"The Average Mutual Information Profile as a Genomic Signature. BMC Bioinformatics"*. 2008. 9:48 doi:10.1186/1471-2105-9-48.

[28] Deschavanne PJ, Giron A, Vilain V, Fagot G, Fertil B. *"Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences"*. Mol. Biol.Evol. 1999. 16:1391–1399.

[29] Jeffrey HJ. *"Chaos game representation of gene structure"*. Nucleic Acids Res. 1990. 18:2163–2170.

[30] Boto L. *"Horizontal gene transfer in evolution: facts and challenges"* Proc. R. Soc. B. 2010. 277:819-8. doi: 10.1098/rspb.2009.1679

[31] Steuer. *"The mutual information: Detecting and evaluating dependencies between variables"*, Korean J Parasitol. 2012 Jun; 50 (2):119-26 22711922.

[32] U. Liane Rosewich, H. Corby Kistler, *"Role of horizontal gene transfer in the evolution of fungi"*.Annu. Rev. Phytopathol. 2000. 38:325–63

[33] J. N. Davidson, R. L. P. Adams *"Bioquímica de los ácidos nucleicos"*,España, Edit. Reverte S.A. 1980

[34] Alon, Uri, *"An introduction to systems biology, design principles of biological circuit"*, Chapman and Hall, (2006), 5-6

[35] Kurland, C.G., Canback, B. And O. Berg. (2003). *"Horizontal gene transfer: A critical view"*. Proc Natl Acad Sci USA, 100, 9658-9662.

[36] C. Calderon, L. Delaye, V. Mireles , P. Miramontes (2012). *"Detecting Lateral genetic material transfer"*. DRAFT, México,2-4.

[37] Sydney Brenner, *"Mi vida en la ciencia, las aportaciones de un biólogo experimental"*, Catedrá de la divulgación de la ciencia, (2006), 37-39.

[38] D. Mark, McDonnell, Shiro Ikeda,Manton, *"An introductory review of information theory in the contextof computational neuroscience"*, Biol Cybern (2011) 105:55–70

[39] Zhenxiang Xi1, Robert K Bradley, Kenneth J Wurdack, KM Wong, M Sugumaran, Kirsten Bomblies, *"Horizontal transfer of expressed genes in a parasitic flowering plant"*, Xi et al. BMC Genomics 2012, 13:227.

[40] Aaron A Vogan, Paul G Higgs, *"The advantages and disadvantages of horizontal gene transfer and the emergence of the first species"*, Vogan and Higgs Biology Direct 2011, 6:1

Apéndice A

A.1. A. Cadena y matriz de Markov

Una cadena de Markov, en nombre de Andrei Markov, es un sistema matemático que se somete a las transiciones de un estado a otro, entre un número finito o contable de los estados posibles. Se trata de un proceso aleatorio caracterizado por ser sin memoria: el siguiente estado sólo depende de la situación actual y no en la secuencia de acontecimientos que la precedieron. Este tipo específico de "memorylessness" se llama la propiedad de Markov. Las cadenas de Markov tiene muchas aplicaciones, como modelos estadísticos de procesos del mundo real.

Matemáticamente la cadena de Markov es una secuencia de variables aleatorias X_1, X_2, X_3, ... con la propiedad de Markov, es decir que, dada la situación actual, los estados pasados y futuros son independientes.
Formalmente,

$$Pr(X_{n+1} = x | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = Pr(X_{n+1} = x | X_n = x_n).$$

Los valores posibles de X_i forman un conjunto numerable S llamado el espacio de estados de la cadena. Las cadenas de Markov se describe a menudo por un grafo dirigido, donde los bordes están marcados por las probabilidades de pasar de un estado a los otros estados.

A.2. B. Mapeos Autoorganizados (SOM)

Hay varias formas para comparar las secuencias genómicas en términos de las características obtenidas por este método. En el método de comparación primero cada secuencia genómica puede ser representado en un gráfico de dispersión, en la que cada posición en el eje x se refiere a una de las características que definen el vector característica. El eje y indica el valor de la función correspondiente. Así varias secuencias pueden ser visualizados al mismo tiempo, pero después de sólo unas pocas secuencias presentes en la trama, se hace difícil tratar de hacer comparaciones y extraer las relaciones pertinentes. Una herramienta común para visualizar espacios de alta dimensión es los mapas autoorganizados (SOM), cuyo objetivo es realizar una proyección lineal de espacios de alta dimensión hacia espacios de baja dimensión, permitiendo una visualización aproximada en éstos espacios.

En el trabajo, se aplicó una herramienta de mapeo para visualizar los puntos que representan los genes en alta dimensión en los espacios de composición y organización. El SOM es una aproximación de baja dimensión de puntos en alta dimensión espacios. El SOM es capaz de tener en cuenta las estadísticas de orden superior, no sólo de segundo orden, como PCA y otras herramientas relacionadas. El SOM también es un modelo de red neuronal no supervisada en el que un entrenamiento de unidades o neuronas modifica sus vectores de pesos, originando que los datos de alta dimensión sean analizados al presentarse a través de la SOM mediante un un proceso iterativo, los pesos se modifican en consecuencia a la ecuación:

$$w_i(t+1) = w_i(t) + (t)h_i(G,T)(x_j - w_i(t)),$$

donde x_j es el vector de entrada, w_i es el vector de pesos para la unidad i, (t) es un parámetro de aprendizaje y $h_i(G,T)$ es la función de distancia. Cuando los datos de entrada presentados al SOM, la unidad cuyo vector de pesos sea más similar a la de entrada se activará. La unidad activada, también conocida como la mejor unidad de acoplamiento, modifica su vector de pesos y permite que las neuronas dentro de su entorno se adapten. Una unidad activa intenta afectar todas las otras unidades, pero sólo aquellas para los cuales $h_i(g,t) > 0$ será capaz de modificar sus pesos. Este proceso se realiza para todos los vectores de entrada y, a continuación la función de vecindad disminuye. Una vez que la condición de parada se cumple, el algoritmo termina.

El SOM ha sido ampliamente aplicado al estudio de las secuencias genómicas[9].

- al hacer inferencias acerca de los cambios de dirección de secuencias H1N1/09, y proponer mecanismos para la vigilancia para detectar los precursores potenciales para situaciones de pandemia.
- Al clasificar las secuencias de ADN dentro de una y entre varias especies en subgrupos que corresponden a las categorías biológicas.
- En los fragmento de pequeños racimos de ADN y agruparlos en grupos filogenéticos

El software desarrollado para la ejecución y análisis de los resultados de HGTiADnnd, se puede consultar en la dirección electrónica www.modelos-predictivos.org.mx/grudeca/software/HGTiADnnd

A.3. C. Genomas de las bacterias

Agrobacterium tumefaciens - sinónimo de Rhizobium radiobacter - es una bacteria que causa en las plantas dicotiledóneas unos tumores conocidos como .ªgallas.º "tumores del cuello", que crecen en la zona donde se unen la raíz y el tallo (cuello).

Agrobacterium es una proteobacteria alpha de la familia Rhizobiaceae, la cual también incluye a las fijadoras de nitrógeno que viven en simbiosis con las legumbres. A diferencia de éstas, Agrobacterium es un parásito y causa grave daño a la planta afectada. Es de notar que Agrobacterium no es ni el único ni el más común causante de tumores en las plantas. Muchos pueden ser causados por insectos o larvas que segregan ciertas sustancias que producen el mismo

efecto aparente.

Aquifex es un género de bacterias, uno de los pocos en el filo Aquificae. Las dos especies generalmente clasificadas en Aquifex son A. pyrophilus y A. acolicus. Ambas son altamente termófilas, creciendo mejor en temperaturas acuáticas de 85 a 95 °C. Crecen a menudo cerca de los volcanes subacuáticos o fuentes hidrotermales.1 A. aeolicus fue descubierto en el norte de Sicilia, mientras que A. pyrophilus fue primero encontrado en el norte de Islandia. Son bacterias verdaderas (del dominio Bacteria) al contrario que otros habitantes de ambientes extremos, las Archaea.2

Archaeoglobus es un género de microorganismos hipertermófilos del dominio Archaea. Se compone de dos especies, A. fulgidus y A. profundus que fueron aisladas de fuentes hidrotermales. Archaeoglobus se puede también encontrar en los yacimientos de petróleo a alta temperatura donde puede contribuir a la descomposición del petróleo. El crecimiento óptimo de estos organismos se produce a aproximadamente 83 °C. **Bacillus subtilis** es una bacteria Gram positiva, Catalasa-positiva, aerobio1 comúnmente encontrada en el suelo. Miembro del Género Bacillus, B. subtilis tiene la habilidad para formar una resistente endospora protectora, permitendo al organismo tolerar condiciones ambientalmente extremas.

Borrelia burgdorferi es una especie de bacteria de la clase Spirochaetes y del género Borrelia. B. burgdorferi , es el agente de la enfermedad de Lyme. Esta es una enfermedad zoonótica transmitida por garrapatas. Borrelia burgdorferi lleva su nombre en honor al investigador Willy Burgdorfer que fue el primero que la aisló en 1982. Es una de las pocas bacterias patógenas que puede sobrevivir sin hierro, sustituyendo todas las enzimas de la familia hierro-sulfuro con enzimas que contienen manganeso, evitando el problema que tienen muchas bacterias patógenas para conseguir el hierro.

Chlamydia trachomatis (clamidia), es una bacteria que pertenece al género Chlamydia, familia Chlamydiaceae, orden Chlamydiales. Es una bacteria intracelular obligado que infecta sólo a humanos; causa tracoma y ceguera, infecciones óculogenitales y neumonías. Algunos individuos desarrollarán el artritis reactiva, que no tiene cura.

El Deinococcus radiodurans (antes Micrococcus radiodurans) es una bacteria extremófila, y el segundo organismo conocido más resistente a la radiación siendo el primero el Thermococcus gammatolerans. Mientras que una dosis de 10 Gy es suficiente para matar a un ser humano, y una dosis de 60 Gy es capaz de matar todas células en una colonia de E. coli, la D. radiodurans puede resistir una dosis instantánea de hasta 5000 Gy sin pérdida de viabilidad, y dosis de hasta 15000 Gy con un 37 % de pérdida de viabilidad. Además, puede sobrevivir en condiciones de calor, frío, deshidratación, vacío y ácido. Debido a estas características, se ha sugerido que estas bacterias podrían ser capaces de sobrevivir en el espacio exterior.