UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

Biología Experimental.

“Análisis evolutivo y estructural de genes traslapados en procariontes.”

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS
Biología Experimental

PRESENTA:

Biol. Hiram Massa Anaya.

TUTOR PRINCIPAL DE TESIS: Dr. Carlos II Arturo Becerra Bracho
Facultad de ciencias, UNAM.

COMITÉ TUTOR: Dr. Pablo Padilla Longoria
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. (IMAS)

Dr. Alexander de Luna fors
Laboratorio Nacional de Genómica para la Biodiversidad (langebio)

Dr. Luis Delaye Arredondo
Centro de Investigación y de Estudios Avanzados del Instituto (CINVESTAV-IPN)

MÉXICO, D.F. JULIO, 2013.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

Biología Experimental.

“Análisis evolutivo y estructural de genes traslapados en procariontes.”

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS
Biología Experimental

PRESENTA:

Biol. Hiram Massa Anaya.

TUTOR PRINCIPAL DE TESIS: Dr. Carlos II Arturo Becerra Bracho
Facultad de ciencias, UNAM.

COMITÉ TUTOR: Dr. Pablo Padilla Longoria
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. (IMAS)

Dr. Alexander de Luna fors
Laboratorio Nacional de Genómica para la Biodiversidad (langebio)

Dr. Luis Delaye Arredondo
Centro de Investigación y de Estudios Avanzados del Instituto (CINVESTAV-IPN)

MÉXICO, D.F. JULIO, 2013.

VNIVERSIDAD NACIONAL AVTONOMA DE MEXICO

POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
DIVISIÓN DE ESTUDIOS DE POSGRADO

OFICIO FCIE/DEP/106/13

ASUNTO: Oficio de Jurado

Dr. Isidro Ávila Martínez
Director General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **10** de **diciembre** de **2012**, se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS (BIOLOGÍA EXPERIMENTAL)** del (la) alumno (a) **MASSA ANAYA HIRAM** con número de cuenta **96352032** con la tesis titulada **"Análisis Evolutivo y Estructural de Genes Traslapados en Procariontes"**, realizada bajo la dirección del (la**) DR. ARTURO CARLOS II BECERRA BRACHO:**

Presidente:	DR. PEDRO EDUARDO MIRAMONTES VIDAL
Vocal:	DR. LUIS JOSÉ DELAYE ARREDONDO
Secretario:	DR. ALEXANDER DE LUNA FORS
Suplente:	DRA. LUCIANA RAGGI HOYOS
Suplente:	DR. PABLO PADILLA LONGORIA

Sin otro particular, me es grato enviarle un cordial saludo.

A t e n t a m e n t e
"POR MI RAZA HABLARA EL ESPIRITU"
Cd. Universitaria, D.F., a 6 de marzo de 2013

DRA. MARÍA DEL CORO ARIZMENDI ARRIAGA
Coordinadora del Programa

MCAA/MJFM/ASR/

Agradecimientos.

El término de este trabajo no abría sido posible sin los apoyos fundamentales del Posgrado de Ciencias Biológicas, UNAM, el cual admitió el desarrollo del proyecto conacyt 100199, además agradezco al CONACYT que me otorgó apoyo económico durante el desarrollo del mismo, permitiendo así que me dedicara tiempo completo a este trabajo.

Agradezco también al Laboratorio Nacional de Genómica Para la Biodiversidad (LANGEBIO) y al Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional, unidad Irapuato. que apoyaron con las herramientas necesarias para la realización del proyecto, en especial al laboratorio de sistemas genéticos, liderado por el doctor Alexander de Luna Fors, así como al laboratorio de genómica evolutiva y al doctor Luis Delaye Arredondo.

Además de agradecer a todos y cada uno de los miembros del comité que ayudaron a mi formación académica con sus consejos y criticas.

A los tres locos por la ciencia (H.E.H.)... Arturo Becerra, Alexander de luna y Luis Delaye, por que de ellos aprendí aparte de lo académico, Humildad, Esfuerzo y trato Humano, respectivamente, que de eso hay poco en el mundo en estos dias.

Al doctor Pablo Padilla por permitirme seguir en esta locura llamada ciencia.

Gracias.

Dedicatoria.

Este trabajo esta dedicado a todas las historias conocidas y platicadas, a los que se fueron sin llevarse el abrigo de la puerta, a los que durmieron bajo las fantasías surrealistas de Remedios varo, a las ideas mórbidas que modificaron la visión de esto que tan poco conocemos la vida, a la simbiosis orgánica resultado de amenas charlas con los que hoy ya nos se encuentran entre nosotros, a los que con sus criticas crearon la fortificación de lo que hoy se conserva firme cerca de las estrellas.

Samuel, gran amigo que acompañaste aquella juventud y nos dejaste para dar un viaje profundo a otros planos.

A esa familia que iba y venía cual oleaje de mar en mayo que no se quedaba por temor a no entender la fragilidad de la playa, a los que llegaron de una dura travesía y lograron permanecer entre nosotros "Samantha", a los que crecieron sin permitirme darme cuenta de ello "Casandra e ivan", a mi pequeña criatura que me salvo de la triste oscuridad regalándome sus risas neonatas "Jacqueline" y a los que siempre estuvieron ahí.

A usted Doctor "Arturo Becerra" que brindo su tiempo, su paciencia y 500 pesos los cuales nadie apostaría en un juego de pokar con una mano de un par de 2's, para darme una oportunidad que ahora devuelvo en unas cuantas paginas y en una mano extendida que invita a una larga amistad.

A ti hermosa mujer que has compartido esta parte de mi vida, te agradezco por estar y darme fuerzas para seguir en esta locura, por lo que has creado para ambos, por todo lo que has ofrecido, te agradezco con todo mi corazón por invitarme a participar en tu vida y acompañarte en tu camino.

"Te amo preciosa, podríamos morir en un instante, pero eso, es lo de menos cuando se ha escrito en las paginas del destino un amor como el nuestro que durará hasta la eternidad."

Gracias por existir.

Índice

Resumen.

El origen de los genes es una de las preguntas interesantes que la biología evolutiva y molecular intentan resolver, ya sea por el fenómeno mismo de la aparición de nuevas regiones genéticas funcionales, como por los efectos de dichos genes ejercidos sobre el genoma y su resultado fenotípico (las novedades genéticas). Hasta el momento a este proceso se puede suponer una sucesión de diferentes pasos que dan como resultado la permanencia de los nuevos genes, a) la aparición de genes resultado de distintos tipos de mutación; b) la divergencia y evolución de dichas secuencias y por último; c) la fijación en la población por efectos de selección natural.

La mayoría de los mecanismos requieren una preexistencia de genes bien establecidos, por lo que el surgimiento de nuevos genes de manera espontánea ha sido poco estudiado. Tal es el caso del gen *htgA,* el cual se ha sugerido que su origen se dio a partir de mutaciones puntuales que permitieron un nuevo marco de lectura que superó las barreras evolutivas impuestas por la selección natural dando como resultado su permanencia en *E.coli k-12.*

Rastrear los nuevos genes que presentan una historia evolutiva distinta a la visión de duplicación y divergencia, es un enorme reto que la biología evolutiva ha intentado abordar, por lo que identificar este tipo de genes es crucial.

Por el momento no se conoce un mecanismo que promueva la compactación genética en las secuencias de los organismos, lo que nos lleva a aquellas secuencias traduccionalmente activas que se encuentran compartiendo varios genes, los cuales se les conoce como genes traslapados. Estos logran ser un excelente ejemplo de genes de aparición reciente por sus características (estan limitadas a un solo grupo, se encuentran compuestos por un distinto uso de codon al promedio, etc..) y por lo tanto una herramienta utilizada en este trabajo para observar algunos fenómenos evolutivos relacionados con la aparición de un nuevo gen.

Se han analizado una serie de genes que presentan características que nos permitieron obtener información que nos deja conocer mas sobre las interacciones entre los genes recientes (traslapados) y sus portadores genes funcionales bien determinados.

En este trabajo se buscaron efectos selectivos de las secuencias traslapadas sobre sus portadores, utilizando métodos bioinformaticos. Se analizaron 9 genes traslapados correspondientes a *Pseudomonas flourescences y de Escherichia coli k-12. S*e encontraron solo 3 genes con evidencia de selección negativa, así como descubrir la estructura de 3 productos, que nos permitirían poner en evidencia si los genes recientes presentan diferentes formas o se encuentran dentro del universo de proteínas que se conoce hasta la fecha, lo cual apuntaría que los nuevos genes que han aparecido por mutaciones puntuales siguen una ruta independiente a los productos ya conocidos.

El análisis revela que el tipo de traslape, el costo informacional por región del gen (terminal-inicial), por posición y la longitud son fundamentales para rastrear los efectos ejercidos sobre las secuencias portadoras, dependiendo la relación que existe entre tamaño, tipo de traslape y regiones altamente conservadas.

El método falla al rastrear los efectos ejercidos del traslape sobre el portador cuando la longitud del traslape es menor con respecto a la del portador.

Se pudo observar en el caso nov 6 selección negativa en la región no traslapada así como en el análisis del gen completo, lo cual nos indica que el traslape aumenta los efectos selectivos sobre el gen portador. El caso nov 13 es distinto, con el método se puede observar selección negativa en la región traslapada así como en el análisis del gen completo, lo cual sugiere que los efectos por tamaño y tipo de traslape afectan de mayor medida al traslape y no al portador. El caso de htgA revela que la selección negativa impuesta por un traslape pudiera ser debida a la localización del traslape, ya que se encuentra en el centro del gen, además de la cohesión del grupo de secuencias que se encuentran muy cercanas y altamente conservadas.

Se realizaron pruebas de purificación de 3 genes, htgA en sus dos versiones y cosA. Se encontraron las condiciones para la purificación de solo uno de ellos (cosA) lo cual permite que el proyecto siga en trabajos posteriores.

La selección negativa evidente en los casos nov 6, nov 13 y htgA, son resultado de un traslape y sus diferentes efectos que producen en los portadores que a su vez permiten la estabilidad, permanencia de las secuencias traslapadas y ¿por qué no? del propio portador.

Abstract.

The origin of genes is one of the interesting questions that evolutionary and molecular biology has been trying to solve, whether by the phenomenon per se involving the emergence of new functional gene regions, or as the effects of these genes on the genome and the phenotypic outcome (genetic novelties). This process can be assumed as a succession of different steps which results in the retention of new genes: a) the appearance of new genes resulting from different types of mutation, b) the evolution and divergence of these sequences, and c) the fixation of these sequences in the population by natural selection.

Most mechanisms require pre-existing well established genes, so the spontaneous emergence of new genes has been little studied. Such is the case of *htgA* gene, that has been suggested that comes from punctual mutations that caused a new reading frame that overcame the evolutionary barriers imposed by natural selection resulting in its permanence in *E.coli k-12*.

Tracking new genes that have an evolutionary history different from duplication or divergence is a major challenge that evolutionary biology has attempted to address, that's why the identification of such genes is crucial.

Nowadays the mechanism that promotes genetic compaction of the sequences in organisms is unknown, this leads us to the sequences shared by several genes and that are translationally active, better known as overlapping genes, which are an excellent example of recent onset genes and therefore a tool used in this work to observe some evolutionary phenomena related to the emergence of a new gene.

Has been analyzed a series of genes, wich presented characteristics that allowed us to obtain information that lets us know more about interactions between recent genes (overlapping genes) and its well-defined and functional carrier genes.

This study sought selective effects of overlapping sequences on their carriers, using bioinformatic methods. Nine overlapping genes corresponding to *Pseudomonas flourescences* and *Escherichia coli k-12* were analyzed. Three of the genes we found show evidence of a negative selection and the structure of 3 products was discovered, this allow us to highlight if there are different forms on recent genes or if they are included in the proteins known to date, this would suggest that the new genes that appeared as a result of punctual mutations do follow a independent path of that of the products already known.

The analysis reveals that the overlapping type, and the informational cost from each gene region (end, start), by position and length are critical to track the effects exerted on the carrier sequences, depending on the relation between size, overlapping type and highly conserved regions.

The method fails in tracking the effects exerted by the overlap on the carrier when the length of the overlap is lower with respect to that of the carrier.

Negative selection was observed in the case of nov 6 in the non-overlapping region and in the analysis of the entire gene, this indicates that the overlap increases the selective effects on the carrier gene. In nov 13 the case is different, the method indicates negative selection in the overlapped region and in the complete gene analysis, this suggests that the effects of size and type of the overlapping affects the overlap and not the carrier. In the case of htgA, the negative selection imposed by an overlap may be due to the location of the overlap that is localized in the center of the gene, besides the cohesion of the sequences that are closely and highly conserved.

Has been performed purification tests of 3 genes, htgA in its two versions and cosA. Only the conditions for the purification for cosA were found, which allow to do further studies to continue this project.

The negative selection present in nov 6, nov 13 and htgA, are the result of an overlap and the different effects that produce in its carriers which in turn allow stability, permanence of overlapping sequences and, why not, of the carrier itself.

1.- Antecedentes.

1.1 Orígenes

Entender el origen de los genes es para la biología evolutiva uno de los retos más importantes. Se ha intentado explicar el origen desde diferentes puntos de vista: la transferencia horizontal de genes, los rearreglos génicos (transposición, retro transposición, fusión y fisión de genes), la aparición *de novo* y las duplicaciones con divergencia de módulos ancestrales, propuesta por *Haldane (1933) y Müller (1935)* que ha sido el más estudiado hasta el momento.

La discusión se torna interesante cuando se contraponen las ideas de un surgimiento continuo de genes (Keese & Gibbs, 1992) contra lo que denominan un "*big bang"* genético (Kaessmann, 2010). Se han encontrado independientemente diferentes ejemplos que pudieran apoyar ambas hipótesis, desde la aparición *de novo* en virus y procariontes (surgimiento continuo), así como de mecanismos de duplicación, transposición y fusión de secuencias (*big-bang* genético).

De los trabajos que abordan el tema la mayoría se enfoca en duplicación y/o transposición y divergencia, dejando a un lado los demás mecanismos que se describirán en seguida.

1.2 Transferencia horizontal.

La transferencia horizontal es definida como - "el intercambio de material genético entre dos o más especies y/o diferentes tipos celulares que provengan de distintos ancestros" - (Ochman, Lawrence, & Groisman, 2000). Este mecanismo puede proveer de novedades genéticas a los organismos, asimilando fragmentos genéticos que confieran un nuevo grupo de genes al receptor, esto ocurre principalmente en procariontes aunque en eucariontes se ha demostrado que se puede dar principalmente entre parásitos y sus hospederos, la importancia de la transferencia horizontal ha sido muy discutida y ampliamente documentada (Keeling & Palmer, 2008; Koonin, Makarova, & Aravind, 2001).

El fenómeno de transferencia horizontal se puede llevar a cabo por tres mecanismos: la transducción, transformación y conjugación (figura.-1).

La transducción es el mecanismo el cual esta mediado por un vector parásito, (virus principalmente), en donde la maquinaria del hospedero es utilizada para la replicación del virus facilitando el intercambio del material genético entre el virus y su hospedero promoviendo en muchas ocasiones el secuestro de porciones del genoma, las cuales suelen intercambiar cuando vuelven a infectar a la misma u otra población de células que incorporan las regiones secuestradas a su genoma , en ocasiones permite la aparición de un nuevo gen o complementan una región transcripcionalmente activa lo que resulta en el surgimiento de nuevos productos proteicos que pueden afectar la adecuación organismo, dando una oportunidad de fijarlo en la población (Ochman et al., 2000).

La transformación implica la obtención de DNA liberado al medio y hacerlo propio, las células receptoras previamente debieron pasar por un proceso llamado competencia, esto quiere decir susceptibles a la incorporación de materiales liberados por células lisadas, que se da por efecto de estrés ambiental. Los fragmentos de DNA obtenidos se encuentran en el medio principalmente en forma de plásmidos (aunque no necesariamente), asimilarlos proporciona a los organismos ventajas que les permiten atravesar las barreras impuestas por la selección como es el caso de las resistencias a antibióticos en procariontes, los fragmentos pueden insertarse en regiones del genoma volviéndolas activas o pueden mezclarse con genes bien definidos modificandolos permitiendo el surgimiento de un nuevo gen (Ochman et al., 2000).

La conjugación se basa en la obtención de material genético por contacto físico mediado por una estructura proteica conocida como *Pili* el intercambio de DNA es llevado acabo entre una célula receptora y una donadora. El proceso es más complejo de lo que parece, aun no se acaba de entender las bases moleculares y genéticas que desencadenan la respuesta de la conjugación ya que se ha visto que el proceso no es dependiente de los efectos de estrés en el medio (Ochman et al., 2000).

La obtención de genes a través de transferencia horizontal aunado a una divergencia de los genes recién obtenidos presentan una oportunidad para su remodelaje, alterando el contexto de las secuencias a partir de diferentes presiones de selección a lo largo del gen (ya sea en regiones grandes como dominios o simplemente en pequeñas como en un nucleótido particular), permitiendo que el parecido entre el ancestro y el descendiente génico sea menor, confiriéndole así una identidad propia, lo que se podría considerar como el nacimiento de uno, el cual independiente de su origen pudiera fijarse en la población y comenzar una historia evolutiva propia.

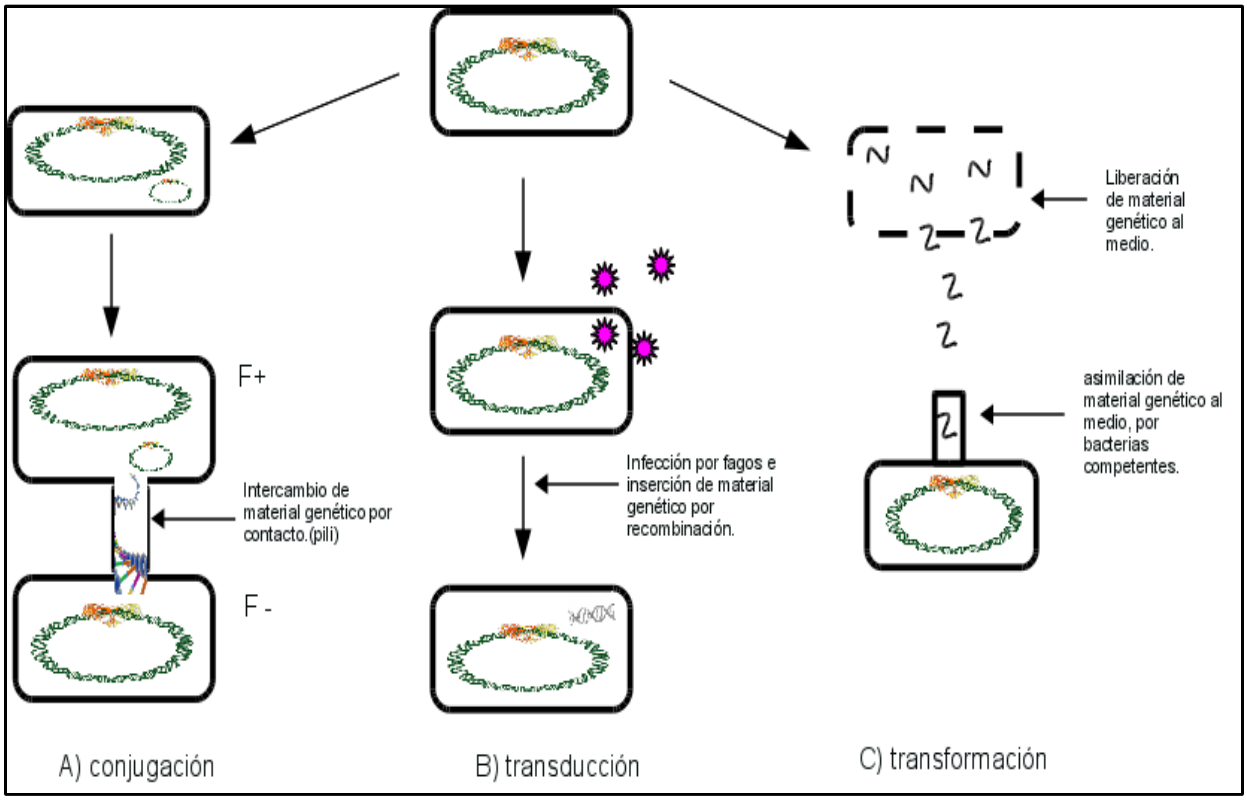

Figura 1.- diagrama que representa los tres tipos de transferencia horizontal de genes, A) conjugación .- intercambio de material genético por contacto físico, B) transducción obtención de material por infección viral, C) transformación asimilación de material genético directamente del medio.

1.3 Rearreglos genéticos.

1.3.1 Transposición

La recombinación genética permite a los organismos tener una gran variedad de fuentes de expresión de las diferentes copias de un gen, intercambiar unos por otros dentro de un genoma y regularlos dándole una fidelidad óptima al sistema, pero a su vez dejan la puerta abierta a un gran numero de "parásitos" génicos, como es el caso de los virus y los transposones (elementos móviles), además se vuelven susceptibles a errores durante el proceso provocando recombinaciones que lleven a la perdida de información, estas aparentes desventajas en las secuencias pueden ser utilizada por la selección natural redefiniéndolas de vez en vez y a diferentes tiempos permitiendo la aparición de secuencias codificantes.

Los elementos móviles (EM) son definidos como entidades genéticas o pequeñas secuencias de DNA de entre 15pb a 20pb en tándem que permiten que se lleven a cabo recombinaciones de fragmentos flanqueados por ellas además se dividen en dos clases: los dependientes de un intermediario de RNA clase I o retrotransposones y los independientes de intermediarios de RNA clase II transposones (Finnegan, 1990; M G Kidwell & Lisch, 2001; M G Kidwell, 2002; Kleckner, 1981). Estas entidades tienen la característica de poder "saltar" a diferentes partes del genoma llevando consigo fragmentos secuestrados provocando recombinaciones inespecificas, estas inserciones pueden llevarse a cabo en regiones que ya eran codificantes o reguladoras, modificándolas a ellas y a sus productos pero a su vez pueden insertar los fragmentos en otras regiones que pudieran complementar o permitir la aparición de nuevos marcos de lectura y como resultado nuevos genes cuando se utiliza un EM de tipo II a esto se le conoce como transposición genética.

1.3.2 Retrotransposición.

El proceso conocido como retrotransposición se determina cuando un RNAm atraviesa por un mecanismo de transcripción reversa, obteniendo un producto que es insertado de nuevo al genoma por recombinación, a esto se le conoce también como una retroduplicación siempre y cuando se vea implicado un gen completo (Kaessmann, 2010), usualmente el resultado tiende a la pseudogenización, ya que la retrotranscripción no tiene incluida una región regulatoria (Brosius, 1991), aunque en algunas ocasiones la inserción del producto puede darse en zonas cercanas a módulos de regulación permitiendo su transcripción y posible traducción. En la literatura hay varios ejemplos de este proceso como en *Drosophila melanoganster (Long & Langley, 1993)*, diferentes plantas y algunos mamíferos como en *Homo sapiens (Mathias, Scott, Kazazian, Boeke, & Gabriel, 1991)*.

1.3.3 Fusión y Fisión de genes.

La fusión de genes se define como la unión de dos o más regiones codificantes en una simple unidad de transcripción (Kaessmann, 2010) y puede darse por yuxtaposición de genes que han sufrido una sucesión de duplicaciones (figura.-2). Estas inserciones pueden darse en las cercanías de otros genes uniéndolos, provocando estructuras quiméricas resultado de las mezcla (Moran, DeBerardinis, & Kazazian, 1999). Una secuencia bien definida que produce este tipo de fenómeno es el retrotransposon

L1 el cual puede secuestrar porciones del genoma posicionándolas río abajo de diferentes genes modificando el producto. Un ejemplo clásico de la fusión de genes es el encontrado gen *jingwei (Long & Langley, 1993),* la quimera surge a partir de la fusión de distintas partes de un duplicado de *ynd* con una retro copia de una alcohol deshidrogenasa, se ha descrito además que se encuentra relacionada con funciones del metabolismo de hormonas y a su vez se encuentra influenciada por efectos de selección positiva (Jones & Begun, 2005). Así, como existe la fusión también hay pruebas de genes formados por fisión en donde una secuencia se disocia del gen, el producto no pierde su función aunque se vea alterada su estructura o en caso contrario llegar a obtener una función y forma diferente a la original, estos genes "incompletos" pueden llegar a alterarse por presiones selectivas obteniendo características propias. Las implicaciones de la fusión y la fisión aun no se han descrito por completo, una gran cantidad de trabajos aun intentan descifrar la importancia de dicho mecanismo en la evolución de los genes y sus productos.

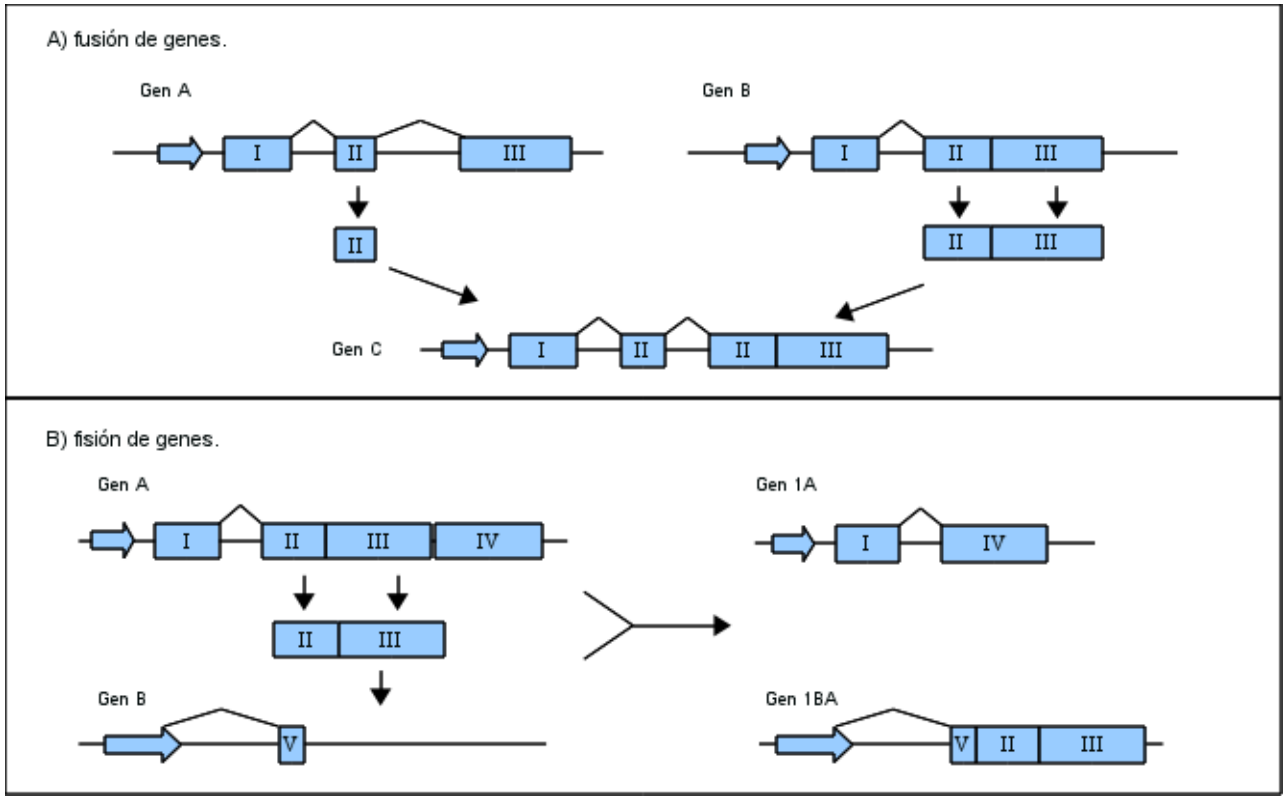

Figura. 2.- diagrama que representa los probables eventos de A) fusión de genes: dos genes (gen A y gen B) atraviesan por un proceso de duplicación parcial de alguna de sus regiones, con una subsecuente inserción en un marco de lectura abierto generando un nuevo gen (gen c). y B) fisión de genes donde un ge (gen A) pierde por recombinación inespecifica una secuencias que puede insertarse en otro gen (gen B) dando como resultado dos nuevos productos uno con una deleción que no afecta su funcionalidad (gen 1A) y otro que por inserción obtenga una nueva función (gen 1BA).

1.4 Duplicación.

Las ideas de Haldane (1933) y Müller (1935) proponen que los eventos de duplicación son causantes principales de la aparición de nuevos genes, esto fue retomado por Ohno (1970) lo que permitió una serie de descubrimientos sobre su importancia en acontecimientos evolutivos a gran escala, Ohno enfatiza que las copias juegan un papel central en la aparición de novedades genéticas, donde una de las copia permanece con la función ancestral y la otra obtiene nuevas propiedadesa esto se le conoce como como neofuncionalización. Además puede presentarse otros destinos para las nuevas secuencias las cuales pueden ser 3 probables, a) perderse en el genoma derivando a pseudogenes (pseudogenización); b) pueden compartir una función el gen original con alguna de las copias y llegar a la subfuncionalización o por el contrario c) como lo predicen las ideas de Ohno obtener una nueva función (neofuncionalización) (figura.-3).

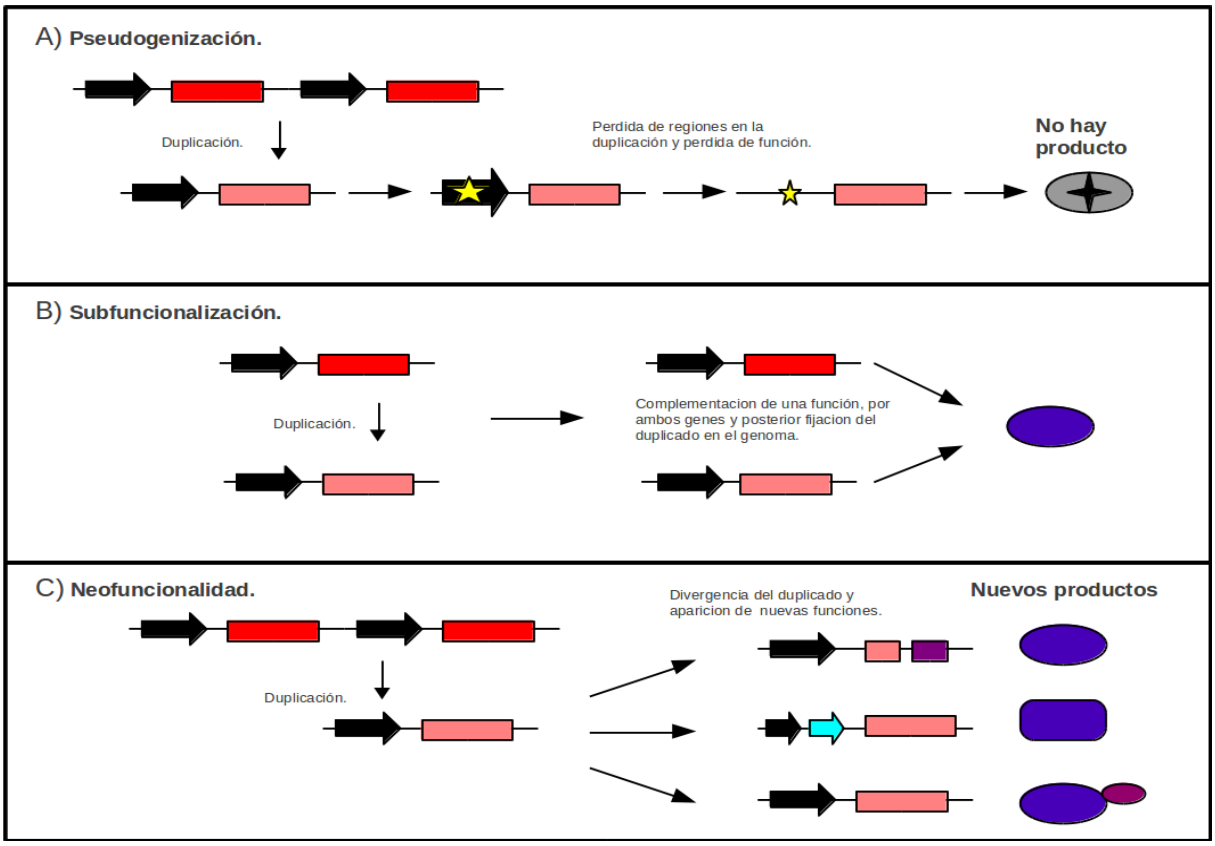

Figura 3.- Los destinos de las nuevas secuencias generadas a partir de una duplicación pueden ser tres a) la pseudogenización en donde una de las copias pierde sus cualidades codificantes por mutaciones azarosas convirtiéndose en un pseudogene, b) La subfuncionalización que permite al producto de la copia compartir o complementar una función que por lo general puede ser el resultado de una duplicación de una proteína bifuncional y c) la neofuncionalización que permite la aparición de nuevas funciones con ayuda de la selección natural por divergencia de los genes resultados de una duplicación.

Existen por el momento descritos 2 tipos de duplicaciones (figura.- 4):

Las duplicaciones de genoma completo que se encuentra principalmente en eucariontes cuyo efecto en los fenotipos es mucho más evidente que las duplicaciones de fracciones o de genes completos del cual existen a su vez dos subtipos:

La duplicación en productos de DNA provocada por recombinación inespecifica, ya sea por fallas en la replicación o recombinaciones genéticas y la duplicación basada en productos de RNA o retro-duplicación resultado de eventos de recombinación provocados por elementos móviles de tipo II o retrotransposones, los cuales interaccionan con moléculas de RNAm provocando una transcripción reversa como se describió previamente (Zhou & Wang, 2008).

Las retro duplicaciones son mediadas por productos de RNAm, como se explicó previamente, la cual depende de la transcriptasa reversa que es codificada por diferentes elementos móviles de tipo II, ejemplos claros de retro duplicaciones existen varios, como el caso de una ribonucleasa pancreática de monos, la cual es codificada por una de las copias del gen ancestral *RNASE1* (J. Zhang, Zhang, & Rosenberg, 2002), se observó que el duplicado *RNASE1B* cuenta con una cantidad considerable de sustituciones en su secuencia, además de encontrarse bajo efecto de selección positiva.

Los productos de las retroduplicaciones podrían perderse debido a que las secuencias codificantes resultado de este proceso difícilmente cuentan con regiones regulatorias, diversos trabajos proponen la idea de que algunos retrogenes pueden obtener secuencias regulatorias (principalmente promotores, potenciadores y/o silenciadores) de diferentes formas: a) utilizando las secuencias de genes adyacentes o b) reclutando secuencias conocidas como protopromotores con altos contenidos de CpG de las proximidades que no se encuentren previamente asociados a genes.

También se ha demostrado que algunos retrogenes insertados cerca o adyacentes a sus copias pueden ser controlados por la misma secuencia reguladora, en otros casos pueden haber heredado promotores alternativos integrados y por último, la aparición *de novo* a través de sustituciones de regiones adyacentes (Kaessmann, 2010; Kaessmann, Vinckenbosch, & Long, 2009), ya obtenidas las

regiones reguladoras pueden ser traducidos y eventualmente remodelados por la selección natural, un ejemplo de ello es el retrogen de ratón *Rps23* que proveniente de una proteína ribosomal del cual se han encontrado cientos de copias en distintos genomas de mamíferos los cuales han sido pseudogenizados, *Rps23* obtuvo la región regulatoria (protopromotor) a partir de que se insertó el duplicado en una región adyacente, se mantiene funcional y evita la formación de laminas amieloides un factor causante de Alzheimer (Y. wu Zhang et al., 2009).

No solo las regiones codificantes pueden pasar por este proceso, se han encontrado algunos casos de secuencias no codificantes funcionales como es el caso de la gran cantidad de los microRNAs y siRNAs que son reguladores postranscripcionales que surgen por este mecanismo (Carthew & Sontheimer, 2009). Varios trabajos apoyan la idea de que las duplicaciones han provocado cambios fenotípicos a gran escala en los organismos, por lo que en los varios trabajos se les ha dado una mayor relevancia evolutiva sobre otros mecanismos que permiten la aparición de nuevos genes.

1.5 Aparición *de novo.*

La aparición *de novo* es conocida como el proceso evolutivo en el cual los genes surgen a partir de mutaciones puntuales que dan la posibilidad de la apertura de un nuevo marco de lectura que se vuelve codificante en el genoma. En los últimos años se han encontrado evidencias de genes que han aparecido en regiones no codificantes, por mencionar algunos ejemplos relacionados con este proceso están varios genes en *Drosophila melanogaster* los cuales han sido bien identificados con un origen *de novo* como el gen *CG33235* que se identificó en patrones de expresión de *testis* y unos cuantos reportados en levadura como el caso de el gen *BSC4* el cual está implicado en la reparación de DNA durante la fase estacionaria y aumenta la robustez en cuando crece en un medio mínimo, por el momento se conocen pocos ejemplos en bacterias debido a lo difícil que es encontrar este tipo de genes (Cai, Zhao, Jiang, & Wang, 2008; Long & Langley, 1993) (figura.-4).

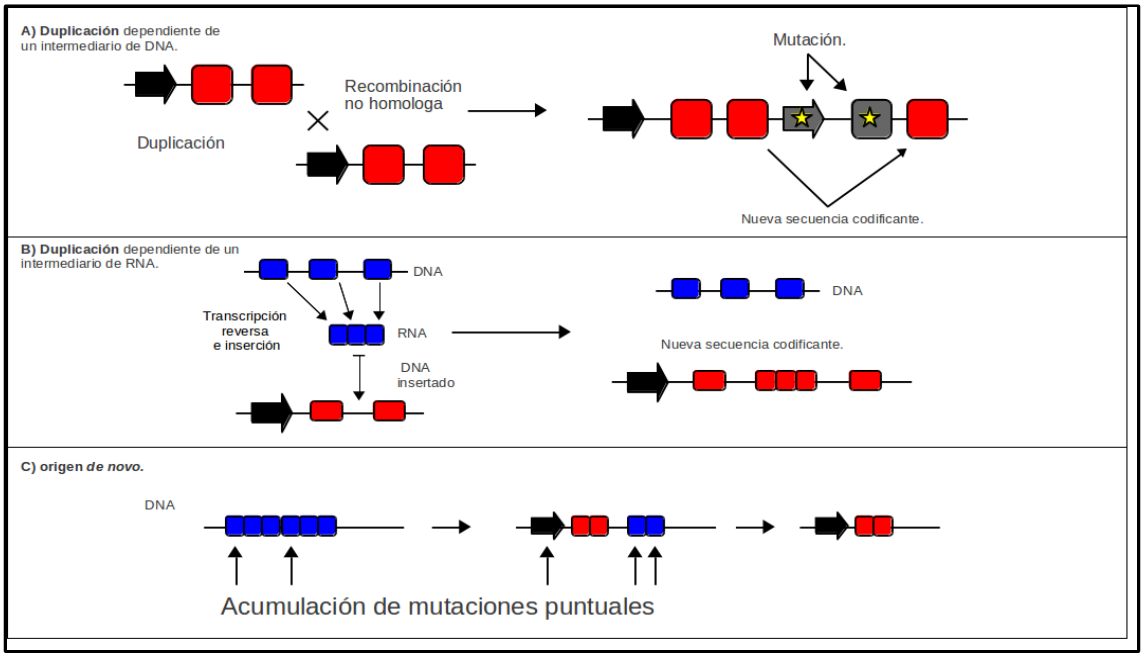

Figura 4.- Representación de los tipos de duplicaciones A) Dependientes de DNA, la cual es el resultado principalmente de recombinaciones no Homologas entre secuencias de DNA, B) Dependientes de RNA que son el producto de la proteína reverso transcriptasa, además el origen de nuevos genes también puede ser *de novo* en donde la acumulación de mutaciones puntuales permiten la aparición de nuevos marcos de lectura, los cuales eventualmente se vuelven transcripcionalmente activos.

Se puede inferir que los nuevos marcos de lectura que surgen podrían presentarse en cualquier parte del genoma y deben de presentar dos características que les permitirán fijarse en la población, a) que se vuelva una región transcripcionalmente activa y b) sea codificante de productos que afecten directamente al individuo portador en el espacio-tiempo adecuado (*momentum*), para lograr estar bajo presiones selectivas que son indispensables para tener la posibilidad de ser elegido dentro del *pool* de genes del organismo que sera heredado a las generaciones siguientes.

Por las complicaciones metodológicas los nuevos genes que surgen en las regiones intergénicas se han analizado poco, identificar y diferenciar a los genes que surgen *de novo* de los que han aparecido por otro mecanismos es muy complicado, las filogenias pueden ser una buena herramienta para localizarlos pero en la mayoría de los casos, las historias evolutivas rebuscadas pueden dar información errónea, localizar genes fuera de fase puede ser otro parámetro útil aunado a la cantidad de homólogos encontrados del gen en cuestión.

Debido a la dificultad de su búsqueda, otra estrategia mucho más confiable es encontrando genes traslapados (Rancurel, Khosravi, Dunker, Romero, & Karlin, 2009). Por el momento no se

conoce de algún mecanismos que promuevan la compactación del genoma lo que provocaría a su vez que la información contenida se sobreponga y por lo tanto origine traslapes génicos lo que podría sugerir que los genes traslapados son resultado y una buena referencia de genes *de novo*. Estos presentan características que podrían ser una buena guía para localizarlos, como lo es el hecho de tener una mínima o nula cantidad de homólogos; la probabilidad de que los genes traslapados sean productos de transferencia horizontal es baja por que la inserción de un fragmento en una región transcripcional activa altera al producto y podría perderse. Los trabajos que se han desarrollado hasta la fecha acerca de genes traslapados pocos hasta ahora los colocan como resultado de origen *de novo*, la mayoría de ellos analizan características, su trayectoria y/o relevancia evolutiva, pero no su origen.

Las tecnologías bioinformáticas se han visto limitadas por lo complejo que es su identificación; por el momento, los virus se han considerados como modelos óptimos para el estudio del fenómeno ya que su alta taza de mutación, su velocidad de reproducción y la gran cantidad de genes traslapados presentes en sus genomas permiten monitorear su dinámica génica. Recientemente se han localizado una gran cantidad de genes traslapados en los tres dominios de los cuales se han intentado describir características propias y su relevancia evolutiva (Scherbakov & Garber, 2000).

1.6 Los genes traslapados.

La definición de los códigos de información genética es un tema complicado de abordar, ya que delimitar en un solo concepto; producto, regulación y los mecanismos de expresión, es difícil ya que existen muchos ejemplos que serían la excepción. La definición reciente de "gen" hecha por Pearson: "Es una región en una secuencia genómica la cual corresponde a una unidad de herencia asociada a regiones regulatorias, de transcripción y funcionales" (Helen Pearson, 2006), permite la inclusión de fenómenos más complejos, como los productos de RNA reguladores o las quimeras genéticas, entre otros.

La interacción entre cada una de las piezas de lego que permite al organismo estar adecuado en su *momentum* refleja la enorme capacidad de cambio del DNA y RNA. Por lo tanto, maximizar el flujo y la fidelidad del mensaje debe de ser crucial, de tal manera que alterarlo podría llevar a la pérdida de información lo cual se traduciría en algún problema subsecuente para el organismo. La aparición de un nuevo gen modifica la interacción entre los genes ya establecidos ¿cómo supera el genoma este

problema? Varios trabajos proponen que se intentan minimizar las modificaciones en las secuencias, por lo tanto, el problema es superado con la aparición de genes traslapados, definidos como aquellas secuencias transcripcionalmente activas que se encuentran compartiendo una región, tal es el caso de los genes traslapados *gag* y *pol* en el genoma del virus de HIV.

Los genes traslapados permiten la continuidad de los portadores sin alterar la interacción que hay entre ellos debido a que se ven influenciados por un doble efecto de selección impuesto por la secuencia compartida (Johnson & Chisholm, 2004), estas restricciones evolutivas contrapesan los efectos de las mutaciones y por lo tanto, disminuyen el efecto de selección positiva (Chirico, Vianelli, & Belshaw, 2010; D C Krakauer, 2000), además proveen una ventaja adaptativa (en virus) cuando existe una alta tasa de mutación (Chirico et al., 2010) (A. Pavesi, 2000) y forman parte de los mecanismos de antiredundancia genética que afecta la longitud del genoma (David C. Krakauer, 2002).

Las diferentes funciones que pueden presentar los genes traslapados son variadas, entre las que se encuentran la compresión genómica (tema de mucha polémica) y la regulación de la expresión genética entre las más importantes (Keese & Gibbs, 1992), además de presentar en la mayoría de los casos no virales, efectos de regulación de otros productos. La mayoría de los genes traslapados se han descrito como reguladores transcripcionales y traduccionales (Scherbakov & Garber, 2000) (Boi, Solda, & Tenchini, 2004) (Makałowska, Lin, & Hernandez, 2007).

1.6.1 Clasificación y características.

Los genes traslapados se pueden clasificar dependiendo su localización, la dirección y su marco de lectura: pueden ser internos o terminales por su localización, ya sea en la misma cadena o en la cadena complementaria por posición, divergente, convergentes y unidireccionales por su orientación, a su vez tener alguno de los 5 marcos de lectura diferentes con respecto a su portador (Makalowska, Lin, & Makalowski, 2005) (Sabath, 2009)(figura.- 5)

En algunos trabajos se ha encontrado una correlación entre la cantidad de genes de un genoma y el número de traslapes que es consistente en la mayoría de los genomas analizados en procariontes *(Eubacterias y Archeobacterias)*, así como en la de los elementos extracromosomales de diferentes

linajes como es el caso de los virus, lo cual puede sugerir que los traslapes son una característica importante en su evolución (Johnson & Chisholm, 2004; Scherbakov & Garber, 2000).

Se han encontrado en una gran cantidad de genomas características que definen los traslapes genéticos: a) por lo menos un tercio de su genoma se encuentra dividido en genes traslapados; b) hay una correlación entre la longitud de genomas, el número de genes y la cantidad de genes traslapados; c) la distribución a lo largo de los genomas es aleatoria, ya sea en la misma cadena formando traslapes unidireccionales o, en cadenas opuestas, traslapes antiparalelos (en su mayoría) ya sean divergentes o convergentes; d) los marcos de lectura más comunes son +1 y +2 para aquellos que son unidireccionales, aquellos que han sido encontrados en fase complementaria con su portador son inestables y tienen efectos deletereos por la aparición de codones de paro que se vuelve frecuente (Johnson & Chisholm, 2004; D C Krakauer, 2000).

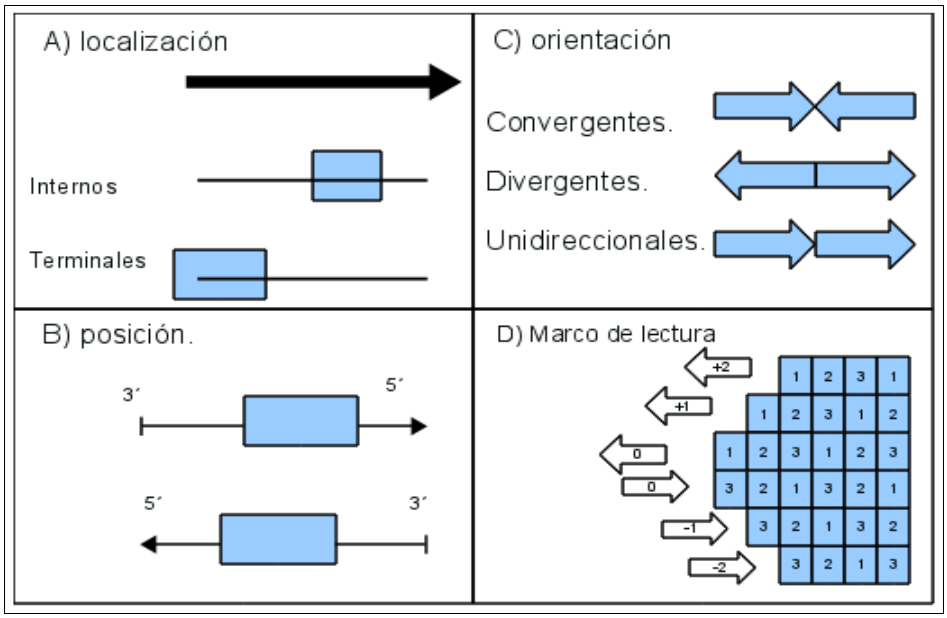

Figura 5.- diagrama que representa la clasificación de los genes traslapados A) por su localización dentro de un gen (portador), B) por su posición en donde el traslape se encuentra en la hebra complementaria la cual correspondería al gen portador, C) por orientación que pueden ser de 3 tipos convergentes donde el portador y el traslape se encuentran, divergentes en donde ambos genes se alejan en sentido contrario o unidireccionales que presentan el mismo sentido transcripcional y el último, por el tipo de marco de lectura con respecto a su portador los cuales pueden ser 5 posibles.

Además de las características antes mencionadas se encontró en un estudio donde se analizaron 50 genomas de bacterias, que el 71.4 % de los genes traslapados son unidireccionales y la proporción entre los genes convergentes es mayor que los divergentes. Se piensa que la baja proporción de genes divergentes es debida a las limitaciones selectivas que afectan a la región 5´ río arriba la cual es

determinante para la expresión de los genes ya que está incluida la región promotora, los traslapes unidireccionales son más fáciles de formar que los divergentes y los convergentes debido a la pérdida del codón de paro. Se piensa que existe una tasa de aparición y desaparición de genes traslapados muy conservada en los genomas bacterianos por lo que se han comenzado a utilizar como marcadores en la construcción de filogenias (Fukuda, Nakayama, & Tomita, 2003).

Debido a que hay una interdependencia en los genes traslapados se ha encontrado que la tasa de cambios sinónimos es mayor que la de no sinónimos, la de fijación de mutaciones se ve reducida en las regiones compartidas, se puede considerar que la necesidad de mantener dos genes funcionales traslapados inevitablemente limita la habilidad de que ambos se encuentren adecuados en su *momentum*, lo cual puede ser contrarrestado por efectos de una o varias duplicaciones de alguno de los genes dando la oportunidad de tener una historia evolutiva independiente (Rogozin et al., 2002), para ello como se ha mencionado antes el traslape y el portador deben de proveer de una ventaja adaptativa al organismo, lo que permite que los productos del traslape estén bajo tres posibles escenarios evolutivos:

1.- Las nuevas secuencias proteicas en particular las que se encuentren traslapadas en la posición C-terminal puedan estar bajo efectos mínimos de selección funcional estando así bajo neutralidad.

2.- Que las nuevas regiones codificantes estén bajo el influjo de la selección positiva favoreciendo las sustituciones mejorando las propiedades físico-químicas del producto.

3.- Las regiones terminales traslapadas pueden estar bajo dos diferentes efectos de selección uno positiva y la otra región bajo selección negativa.

De los probables tres escenarios se ha encontrado que las secuencias mantienen un patrón bien definido debido a que varían su comportamiento dependiendo diferentes factores como serían el tamaño del traslape, la dirección y la fase en que se encuentran uno con respecto a otro, por lo tanto dos de estos escenarios han sido encontrados en la mayoría de los genomas bacterianos a) cuando se encuentra un efecto de selección positiva y b) cuando ambas partes se están bajo diferentes efectos selectivos (Rogozin et al., 2002).

Los productos de estos genes se encuentran codificados por un uso de codón inusual, este presenta un alto grado de degeneración conteniendo grandes cantidades ya sea leucina y/o arginina que da a los productos propiedades físico-químicas parciales los cuales pueden volverse adaptaciones ventajosas como lo son en el caso en los virus lo que les permite mantener pocos genes bien conservados, así como aceptar por medio de traslapes y/o complementación la aparición de nuevos productos proteicos sin alterar a los genes bien establecidos (Pavesi, De Iaco, Granero, & Porati, 1997; Rancurel et al., 2009).

1.7 El universo de Proteínas.

El universo de las proteínas que contiene todos y cada uno de los productos de los genes conocidos se encuentra delimitado por solo unos cuantos plegamientos proteicos existentes en comparación a la infinidad de posibles formas que pudieran encontrarse. Las leyes evolutivas que afectan a los organismos nos permiten asumir que también dicho universo sigue el mismo patrón lineal ancestro-descendiente donde las forman se mantienen por efectos selectivos y que tienen historias evolutivas que se entrelazan proviniendo todas las formas conocidas hasta ahora de un gran big bang proteico, en este contexto surgen muchas preguntas, por mencionar algunas: ¿qué sucede con los productos de genes que han aparecido *de novo*? ¿tendrían patrones de plegamientos similares? ¿cuáles son las fuerzas que rigen dichas conformaciones? intentando resolverlas se ha tratado de definir al universo de proteínas, cosa que hasta el momento no ha sido fácil.

Por el momento se tienen representadas en PDB aproximadamente 79,851 proteínas (martes-6-2012) las cuales son clasificadas desde diferentes puntos de vista pero si hablamos de estructuras se dividen en un número reducido, lo que nos permite asumir el hecho que solo unos pocos plegamientos son posibles ¿por qué razón se da este fenómeno? ¿es acaso el resultado de un big bang biológico, lo que encontramos? Hasta la fecha se sabe que el número de familias de proteínas es delimitado debido a que se han encontrado muy pocas nuevas familias en los últimos años, además cada que se encuentra una estructura nueva se incluye en los grupos ya preexistentes. Hasta el momento el crecimiento de las familias ha sido lento con respecto a la descripción de nuevas secuencias encontradas que va a un ritmo muy acelerado cada año; por otro lado y en el mismo sentido el descubrimiento de familias multidominio es mayor al de las familias de un solo dominio, cuyas cuales solo son rearreglos o

mezclas de las familias ya existentes que presentan un solo dominio, además entre más secuencias funcionales se descubren se adicionan a las que ya se conocen (Denton & Marshall, 2001; Levitt, 2009; cyrus chothia, 1992). Ahora bien, si las estructuras nuevas forman parte de las familias preexistentes ¿qué sucede con los productos de los nuevos genes generados *de novo*? ¿las leyes físico-químicas podrían ser el factor determinante? delimitando así los posibles plegamientos en vez de los efectos generados por la selección natural, "Si es el caso", las restricciones físico-químicas afectan más a las restricciones históricas en los productos orgánicos lo cual nos daría una respuesta viable a la pregunta pero aun así, no podríamos descartar el hecho de que la selección natural es de importancia para la permanencia de las secuencias generadas *de novo,* ambas propuestas no son excluyentes una de otra, ambos efectos pudieran ser determinantes de igual medida a la hora de la aparición de un nuevo producto el cual asumimos que es funcional y a su vez está fijado en una población.

1.8 Genes *Nov´s, cosA y htgA.*

Ahora bien, retomando la idea principal de analizar genes de aparición reciente, nos permite acercarnos un poco más a entender el efecto de la selección en las secuencias, tanto en las regiones portadoras como en las traslapadas y a su vez descubrir cómo son las estructuras de estos nuevos productos. Para ello se buscó un grupo de genes que tienen las siguientes características: ser traslapados, que sean funcionales y un indicio de que sean resultado de una aparición *de novo*.

1.8.1 Los genes *Nov´s.*

Los genes Nov´s fueron identificados en *Pseudomonas fluorescens (P. flourescens)* por medio de espectrometría de masas, además de análisis de péptidos por medio de HPLC también se extrajo RNA para pruebas de RT-PCR, complementado con análisis bioinformáticos, se realizaron la búsqueda de nuevas secuencias codificantes por los programas GenemarkS (http://opal.biology.gatech.edu/GeneMark/genemarks.cgi) y *Glimmer* (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi), encontrando secuencias codificantes de las que se realizaron TBlastN en búsqueda de homólogos en GenBank de NCBI así, con ayuda de *clustalw* se alinearon con las obtenidas por los análisis de proteómica, obteniendo 16 secuencias que no habían sido anotada ni identificadas previamente.

Para cumplir los objetivos de este trabajo se tomaron 7 de las 16 identificadas (Nov´6, Nov´7, Nov´8, Nov´11, Nov´13, Nov´14, Nov´15) (Kim et al., 2009).

1.8.2 El gen *cosA*

De los genes traslapados con funciones bien definidas hay pocos ejemplos en la literatura debido a las complicaciones metodológicas de caracterizar genes putativos. Para este trabajo se necesitaron genes con funciones bien caracterizada, ejemplo de ello son los genes *nov´s* que a pesar de haber sido detectados por distintas técnicas moleculares, aún no se han realizado análisis suficientes que nos den pruebas de su función y/o si afecta la presencia de estos la adecuación al organismo que lo porta y ya que suponemos que han aparecido *de novo* sería de gran ayuda tomar en cuenta algunos ejemplos de genes traslapados que se haya comprobado su función, por tal motivo el gen *cosA* fue elegido para complementar este estudio, porque existen pruebas sobre su funcionalidad.

El gen *cosA* fue descubierto mientras eran localizados genes traslapados funcionales por métodos bioinformáticos en el genoma de *P. fluorescens,* posteriormente se encontró en trabajos experimentales que era codificante para una proteína que es importante para la colonización del suelo bajo diferentes tipos de estrés (aunque no se ha descrito adecuadamente como es que afecta dicho proceso), además se encontró que la sobre expresión por medio de múltiples copias del gen aumentaba la velocidad del proceso de colonización, para su detección se utilizó un método de PCR reverso; las deleciones del codón de paro de *cosA* y su relación con la baja o tardía colonización del suelo de *P. fluorescens* comprueban que es transcripcionalmente activo y funcional (Silby & Levy, 2008).

1.8.3 El gen *htgA.*

Además de ser funcionales debemos tener algún gen traslapado que haya sido analizado en un contexto evolutivo para poder tener una idea sobre los efectos que puede ejercer la selección en una secuencia codificante, el gen *htgA* fue elegido porque ya se tiene evidencia experimental y bioinformática; la proteína fue identificada por mutagénesis de transposición mientras era estudiado un operon que contiene el gen dna*K* y su producto una proteína *hsp70 (heat shock protein 70kD),* el producto de *htgA* es una proteína que funciona como regulador positivo del factor de transcripción

sigma 32 en respuesta al crecimiento a altas temperaturas de *Escherichia coli K-12.* (James. & Deyrick, 1991; Missiakas, Georgopoulos, & Raina, 1993); también fue utilizado como modelo para explicar la posibilidad de la aparición de nuevos genes por el mecanismo de sobreimpresión o calca génica (Overprinting) y la evolución de secuencias transcripcionalmente activas a partir de regiones no codificantes, en el análisis se observó su distribución filogenética, se identificaron los dominios proteicos y se observó la disminución en la tasa evolutiva de *yaaw* debida al traslape de *htgA* el cual impone un efecto doble sobre la secuencia que codifica para ambos productos (Delaye, Deluna, Lazcano, & Becerra, 2008).

La elección de estos genes nos permite tener varios grupos entre los que encontramos unos con evidencias claras y otros que suponemos son funcionales basados en resultados experimentales, así como un trabajo evolutivo previo, lo que nos permitirá abordar las incógnitas que rodean a los genes traslapados en diferentes contextos.

2. Objetivos.-

En este trabajo se trata de analizar uno de los mecanismos que generan novedades en los genomas "los traslapes genéticos", encontrando las huellas que la selección natural deja en las secuencias portadoras, utilizando programas bioinformáticos que nos permitan rastrear estos efectos en los genes; además de obtener cristales de los productos de los genes *cosA y htgA* que nos den pauta para discutir *a groso* modo la aparición de nuevas estructuras en el universo proteico.

En este análisis se hacen preguntas sobre la evolución de este tipo de genes ¿cual es el efecto de la selección en las regiones traslapadas con respecto a la portadoras? ¿La funcionalidad de los nuevos productos puede ser determinada viendo los efectos de la selección natural de las secuencias? ¿como son las estructuras tridimensionales de estos nuevos productos? ¿los nuevos productos son regidos principalmente por su historia evolutiva o presentan características que los autodefinen? estas preguntas nos dan pauta para delimitar el análisis, dividiéndolo en una serie de objetivos metodológicos:

1.-Analizar la distribución filogenética de los portadores.
2.-Predecir posibles traslapes en los genes homólogos de cada uno de los portadores.
3.-Medir los efectos de la selección natural sobre los genes *nov´s, cosA y htgA*.
4.-Sobre expresar, purificar y cristalizar los productos de los genes *cosA y htgA*.

Que a su vez, nos llevaran a resolver 2 objetivos principales:

1.- Estudiar los efectos y los mecanismos en que la selección natural actúa sobre las nuevas secuencias codificantes.
2.- Cristalizar las proteínas *cos-A y htg-A* en sus dos versiones (short y long).

3. Métodos bioinformáticos.

Se obtuvieron las secuencias de los portadores de los genes *nov's, cosA y htgA* reportados en los trabajos anteriores a este (Delaye et al., 2008; Kim et al., 2009; Silby & Levy, 2008) de la base de datos de NCBI (Gen bank):

cosA.

Acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=1091232&to=1093424&RID=RYDMUSPR01N
GenBank: CP000094.2 región: 1091232..1093424
nombre: *Pseudomonas flourescens pf0-1.*

htgA.

Acceso: http://www.ncbi.nlm.nih.gov/nucleotide/359330873?report=gbwithparts&from=10830&to=11315&RID=RYEB77ZX016
GenBank: AP012306.1 región: 10830..11315
nombre: *Escherichia coli str. K-12* substr. MDS42 DNA

Nov6.

Acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=3059193&to=3060938&RID=RYEW9ZUX013
GenBank: CP000094.2 región: 3059193..3060938
Pseudomonas fluorescens Pf0-1, complete genome.

Nov7.

acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=3524719&to=3526674&RID=RYFHKCKD016
GenBank: CP000094.2 región: 3285317..3286696
Pseudomonas fluorescens Pf0-1, complete genome.

nov8.

acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=3524719&to=3526674&RID=RYGHMC0D01S
GenBank: CP000094.2 región: 3524719..3526674
Pseudomonas fluorescens Pf0-1, complete genome.

Nov11

acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=4134938&to=4135837&RID=RYGUTYNJ01S
GenBank: CP000094.2 región: 4134938..4135837
Pseudomonas fluorescens Pf0-1, complete genome.

Nov13

acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=4476661&to=4477458&RID=RYHUKDH0016

GenBank: CP000094.2 región: 4476661..4477458

Pseudomonas fluorescens Pf0-1, complete genome.

Nov14

acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=4755213&to=4756433&RID=RYK0KX8E013

GenBank: CP000094.2 región: 4755213..4756433

Pseudomonas fluorescens Pf0-1, complete genome.

Nov15

acceso: http://www.ncbi.nlm.nih.gov/nucleotide/253992019?report=gbwithparts&from=3407543&to=3409681&RID=RYKARR9V016

GenBank: CP000094.2 región: 3407543..3409681

Pseudomonas fluorescens Pf0-1, complete genome.

La búsqueda de homólogos de los portadores fue hecha con todas las secuencias que se encontraban en la base de datos por medio de Blast-nr, de las secuencias resultantes se excluyeron aquellos genes que pertenecían a fagos y eucariontes, además se tomaron valores por encima del 75% de identidad como parámetro de elección, cada uno de los resultados fue analizado por separado buscando inconsistencias propias de la anotación en la base de datos u de otro tipo que nos provocara malas interpretaciones como lo son: longitud y cobertura.

Las secuencias fueron alineadas con sus respectivos homólogos con el programa *muscle* de la paquetería PHYLIP versión - 3.69 en sistema iso (ubuntu 10.4) las cuales fueron revisadas y repetidas varias veces descartando secuencias que fueran inconsistentes y mal alineadas.

Basándonos en las alineaciones previas se hicieron filogenias con el método de *Máxima verosimilitud* de la paquetería de MEGA 5, con boostrap de 100 repeticiones y sin boostrap. Las filogenias se revisaron cuidadosamente, se buscaron los clados en donde se encontraban los genes bajo análisis y sus grupos cercanos basándonos la cercanía evolutiva.

Teniendo las secuencias de las filogenias internas se buscaron en la base de datos de NCBI-genebank las secuencias en nucleótidos de los genes en cuestión, se alinearon con el mismo programa *(muscle)* y se repitieron las filogenias revisándolas cuidadosamente en busca de incongruencias del método o de la alineación que suelen aparecen con facilidad.

Con el programas *bioedit* se editaron las secuencias manteniendo los parámetros permitidos por la paquetería del PAML (Phylogenetics Analysis Maximum Likelihood version 4.5). PAML es una plataforma que permite análisis filogenéticos de alineaciones a nivel de secuencia de aminoácidos y de nucleótidos utiliza valores de *máxima verosimilitud.* Permite la estimación de parámetros con sofisticados método de sustitución de codones, estima tiempos de divergencia bajo modelos locales o relojes moleculares globales, hace reconstrucción de secuencias tanto de aminoácidos como de nucleótidos entre otros, utiliza modelos complejos para la estimación de tasas de sustitución y detección de selección positiva, para el análisis evolutivo fue utilizado *codeml* un programa que se basa en el uso de modelos de sustitución de aminoácidos (cambios sinónimos y no sinónimos), utiliza un archivo con las secuencias alineadas en formato .sphy, y/o .phy nombradas adecuadamente y sin repetición, un archivo con la filogenia en formato newick (.nwk) y las opciones deseadas para cada uno de los modelos utilizados ya sea el caso (figura.- 6)**.**

Los resultados obtenidos dependiendo el tipo de modelo utilizado por *codeml* son analizados bajo varios supuestos que uno determina, para cada uno de los modelos se obtienen los valores de *máxima verosimilitud* (lnL) y los valores relativos de cada modelo, los cuales se contraponen con la ayuda de una chi^2 obteniendo el valor "P" que es >,< o = al .05 para rechazar o aceptar un supuesto dado por los dos tipos de modelos.

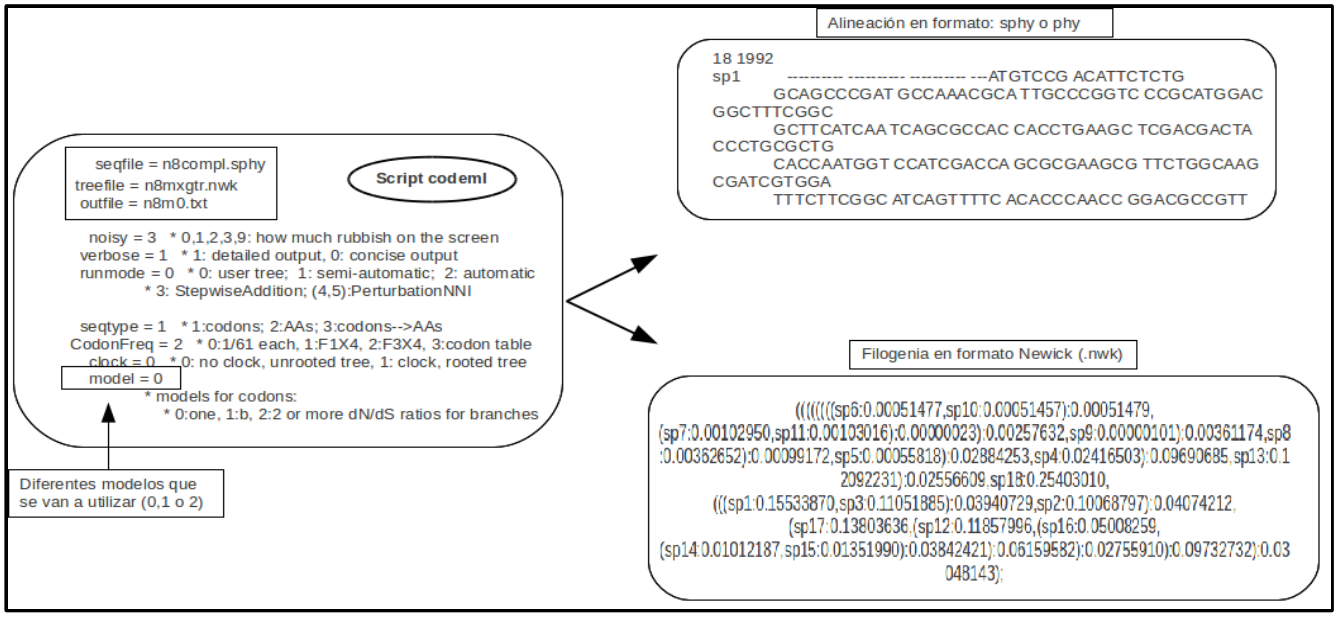

Figura 6.- Representación de los 3 diferentes parámetros que son modificados principalmente en el *script* de *codeml* para obtener resultados del análisis de tasas de sustitución a partir de distintos modelos o supuestos.

El programa *codeml* toma en consideración que cada uno de los genes se encuentran bajo dos modelos, a) el modelos 0 hipótesis nula (M0) dice que todas las secuencias tienen la misma proporción de cambios sinónimos y no sinónimos a lo largo de ellas y b) el modelos 2 (M2) que propone que las proporciones para el gen traslapado y su portador difieren con respecto a sus homólogos (figura.- 7).

Para el análisis de los resultados obtenidos de cada uno de los modelos se utiliza una sencilla formula que es una prueba de ajuste entre los modelos ($LRT = 2[(\ln L_2 - \ln L_0)]$), donde los valores de *máxima verosimilitud* (lnL) nos dan un producto relativo que es el LRT (likelihood ration test). los LRT se contrapone con los esperados de una chi^2, la cual nos da un valor de P ya sea >,< o = a .05 permitiendo así la aceptación o el rechazo de uno de los dos modelos, los grados de libertad utilizados en la prueba de chi^2 son iguales a la resta de los parámetros del modelo más complejo menos los parámetros del modelo más simple, en otras palabras, $df = nP_2 - nP_0$, donde los parámetros son las diferentes omegas (ω) de cada modelo, basándonos en esto podemos proponer el tipo de efecto que pudiera ejercer la selección sobre las secuencias.

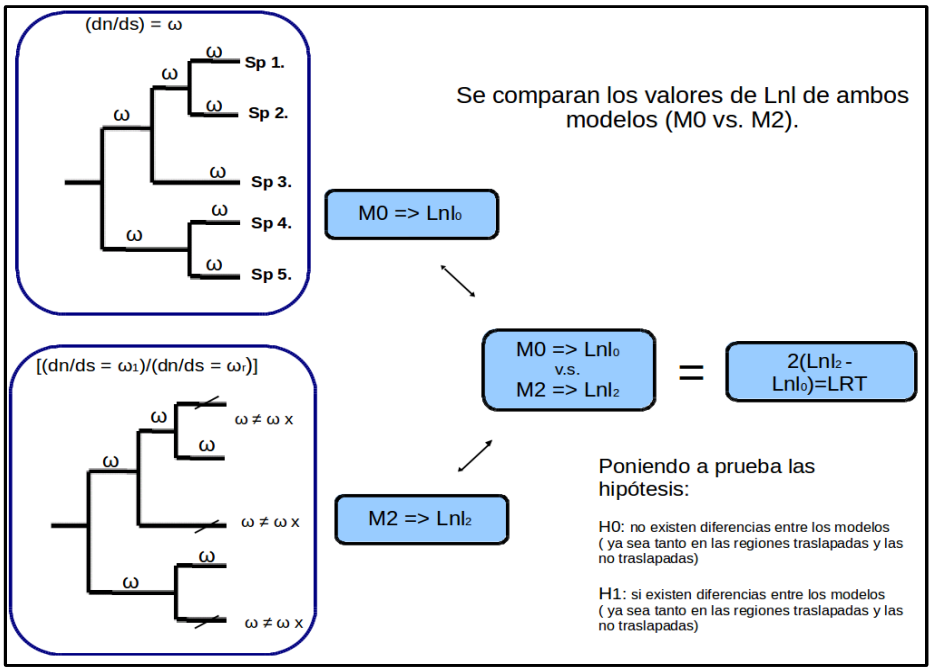

Figura 7.- Diagrama que representa de manera gráfica los dos modelos; M1 las tasas de sustitución son constantes en el tiempo y entre las ramas de la filogenias confiriéndole una omega (ω) para cada una de las ramas el modelo 2 que calcula dos distintas omegas una para aquellas secuencias que tienen un traslape y otra para las que no.

Con el supuesto de que los portadores no serían los únicos que presentaban un traslape se infirieron alguno posibles marcos de lectura que podrían estar transcripcionalmente activos. Basándonos en las secuencias de codones homólogas a los portadores se buscaron con ORF-Finder-NCBI (http://www.nih.go.jp/~jun/cgi-bin/frameplot.p) y manualmente todos los probables codones de inicio que correspondieran al de los genes traslapados reportados complementados con codones de paro que se encontraran en los alrededores del traslape, a partir del sitio de inicio predicho se tomaron 4 codones río arriba y río abajo, por lo que cada uno coincidía con su correspondiente, las predicciones se utilizaron para hacer el mismo análisis cambiando los parámetros del modelo, lo que nos permitiría encontrar un traslape no reportado y así descartar algún efecto que nos evitara ver lo que ocurre en las secuencias.

Las secuencias alineadas se cortaron en las regiones donde existe un probable traslape debido a la idea de que el comportamiento de las secuencias es distinto tanto en el gen completo como en cada una de las regiones en que se dividió, en otras palabras si fuera el caso de que existiese un efecto de selección debido a la presencia de un traslape se vería reflejado tanto en el análisis del gen completo como el de las regiones separadas, debido a que la medición de las tasas de sustitución varían por la presencia de un gen traslapado con respecto a las regiones que no la presentan, por lo tanto el numero de análisis de cada uno de los genes depende de las regiones de las que se compone un gen (región no traslapada y traslapada) además se analizaron los genes completos independientemente con los mismos parámetros y modelos.

Los comandos del archivo de entrada para *codeml* tenían que ser modificados en el archivo de entrada donde se encontraba la filogenia debido a que en cada organismo que presentara un posible traslape se marcaba, para que el programa *codeml* lo pudiera reconocer y analizar desde otra perspectiva con el método antes mencionado, repitiendo así el análisis y comparando resultados de ambos métodos.

La discusión de los resultados del análisis estuvieron bajo los parámetros de longitud del traslape, del portador y marco de lectura, pudiendo observar el tipo de efecto de la selección natural a lo largo de las secuencias y sus consecuencias evolutivas.

4. Discusión y resultados bioinformáticos.

Las ideas de que la aparición de nuevos genes provienen de un solo camino evolutivo de ancestro-descendiente se ha ido atenuando poco a poco en los últimos años, se conocen muchos casos en que la aparición de nuevos genes no siguen esta ruta, como es el caso de los genes traslapados que por el momento han sido poco estudiados.

La explicación de la dinámica evolutiva a lo largo del genoma es complicada ya que no es homogénea y se ve influenciada por las presiones de selección que evitan o permiten el cambio en las distintas regiones que lo conforman. Un ejemplo, son las regiones no codificantes en donde hay un aumento en la tasa de sustitución con respecto a las codificantes, por el hecho de que si un nucleótido cambia en esa región no se ve alterada la dinámica entre los genes ni los productos de los mismos. Por lo tanto si se miden los efectos de la selección en las distintas regiones del genoma se tendría una gama bien definida en donde se presentan cambios mínimos y donde se permiten una mayor cantidad. Pero ¿qué sucede si aparece un gen traslapado sobre una región ya codificante? la proporción de cambios en el gen portador se vería alterada sin duda, pero la pregunta más profunda es ¿de qué manera? ¿los cambios que aparecen pueden ser tenues para el portador y radicales para la nueva región transcripcionalmente activa?, en cualquier caso un análisis que nos permita ver qué sucede en las secuencias nos daría pauta a responder dichas preguntas.

Krakauer propone que entre los más frecuentes se encuentran aquellos que se presentan en las regiones 3' y en un marco de lectura en fase -2, seguido de las fases +1 y +2 en orden descendiente, la fase 0 y por último las fase -1, el estudio realizado observa como el costo informacional del uso de codón interviene o limita la aparición de traslapes en un genoma, por medio de algoritmos matemáticos se infiere la estabilidad de los traslapes y así su permanencia en el genoma, y de manera teórica la evolución de los traslape en los genomas. Además, se puede observar que la dependencia del tamaño, dirección y marco de lectura con respecto a su portador, es fundamental para la evolución de los genes traslapados. Por lo tanto y tomando en cuenta estas ideas, este estudio revela como la selección natural afecta estas regiones, en los diferentes casos analizados no se encontró un efecto tan marcado como el que el que propone Krakauer, los costos que implica un traslape en un gen solo se revelan en tres de los nueve casos analizados, se encontraron 3 tipos de traslapes diferentes 7 son traslapes en fase 0, uno fase +1 y uno en fase -2, el tipo de traslape se encontró 4 convergentes, 3 divergentes, 1 unidireccional y 1

se localiza en medio del gen por lo que se puede clasificar en divergente o convergente. De los genes analizados se discute los efectos de la selección natural sobre las secuencias y sus regiones, basándonos en el costo informacional tanto de longitud y el tipo de traslape que presentan.

4.1 El costo informacional del código genético.

La información contenida en el código genético es crucial en los seres vivos, de ello depende la armoniosa interacción, síntesis y funcionamiento de las proteínas. Se puede ver a cada uno de los genes como módulos que contiene un fragmento de información que permite que todos los demás trabajen adecuadamente, ahora bien cada uno de los genes está compuesto por secuencias divididas en tripletes que también guardan información la cual es necesaria para la construcción de los productos ya sean proteicos o no. Cada uno de los niveles que estructuran un gen pueden considerarse como objetos en donde la selección natural actúa, tanto a nivel de nucleótido, como de codones, de dominios, etc... Cada uno tiene un costo informacional que es conferido por la selección natural, un ejemplo es la probabilidad de sustitución o intercambio de dominios que claramente modifica el funcionamiento y la estructura del producto. Los codones que permiten el ensamblaje de las proteínas, también tiene un costo informacional dependiendo la posición en que son decodificados, concediendo así que los cambios en las regiones sean a mayor o menor escala dependiendo si los aminoácidos determinan la función como es el caso de la interacción con los grupos funcionales. Si nos fijamos a nivel de codones podemos ver que las posiciones de cada uno de los nucleótidos juega un papel importante para decidir el aminoácido que se va a anexar a la cadena polipetídica. Retomando estas ideas podemos definir el costo de cada una de las posiciones de los nucleótidos de un codón. Por el momento, se conoce bien el tipo de producto que puede generar una mutación en un sitio especifico, basado en el hecho de que los codones que codifican un mismo aminoácido muchas veces tienen los dos primeros nucleótidos iguales, cambiando sólo el tercero. Así, mutaciones que se localicen en la tercera posición no suponen cambios en el aminoácido (mutaciones silenciosas). De este modo se minimiza el impacto de mutaciones puntuales cuando ocurren en la posición 3 del codón. En cambio las mutaciones en la primera y segunda posición del codón suelen suponer un cambio de aminoácido (mutaciones sin sentido). Por lo general, los aminoácidos con las mismas características físico-químicas presentan el mismo nucleótido en la posición 2 del codón. Las mutaciones que alteren la posición 1 dan lugar a aminoácidos similares mientras que cambios en la posición 2 del codón, dan lugar a los aminoácidos

con propiedades muy diferentes. Mutaciones en cualquiera de las tres posiciones del codón pueden dar lugar a la aparición de codones de paro provocando una terminación de la traducción. En el caso de tener un traslape supondría una restricción adicional que delimitaría el cambio que pudiera presentarse debido a un restricción evolutiva sobre la secuencia compartida dada por la permanencia de uno de los dos genes por efecto de la selección natural.

4.2 Búsqueda de homólogos a las secuencias *nov y htgA* .

Para cada uno de los genes reportados se buscaron secuencias homologas usando BLASTP en la base de datos GenBank (NCBI), en su versión de aminoácidos. Las secuencias obtenidas reportaron diferentes grados de similitud y cobertura, se seleccionaron aquellos con mejores valores los cuales están descritos en la tabla 1. El total de las secuencias seleccionadas se encuentran disponibles en el archivo anexo 1A.

Con el objetivo de seleccionar las secuencias homologas más relacionadas entre si, se realizó un análisis filogenético de cada uno de los genes portadores reportado en el anexo 2A**,** alineando las secuencias con *muscle* e infiriendo su relación filogenética por *máxima verosimilitud* en el programa MEGA 5 (figura.- 8).

Tabla 1.- Parámetros de selección de homólogos basados en similitud y cobertura, numero final total de genes utilizados en el análisis de tasas de sustitución.

Gen	Homólogos totales.	Homólogos elegidos.	Similitud.	Cobertura.	Genes seleccionados.
Nov' 6	100	33	21% - 49%	> 75%	20 Genes
Nov' 7	100	100	59% - 88%	> 90%	17 Genes
Nov' 8	100	100	51% - 80%	> 95%	18 Genes
Nov' 11	100	100	62% - 93%	> 97%	17 Genes
Nov' 13	100	100	61% - 97%	> 98%	10 Genes
Nov' 14	100	100	68% - 97%	> 99%	15 Genes
Nov' 15	100	100	67% - 97%	>99%	23 Genes
cosA	99	99	49% - 92%	>95%	11 Genes
htgA	100	86	31% - 99%	> 77%	43 Genes

Una vez realizadas dichas filogenias, se eligieron aquellas secuencias de menor distancia a los genes reportados, y con clados soportados con un valor de bootstrap entre el 60% y el 70% . En caso de htgA se tomaron todas las secuencias sin considerar el valor de bootstrap ya que el grupo tenía una cercanía muy estrecha y los valores de boostrap eran bajos a lo largo la filogenia menos en la rama que contenía a htgA. El numero de secuencias seleccionadas por cada gen portador, se presenta en la

tabla 1.

De los genes seleccionadas, se obtuvieron las secuencias de nucleótidos, las cuales fueron alineadas con *muscle* y se hicieron filogenias de *máxima verosimilitud* en el programa MEGA 5 ver anexo 2A.

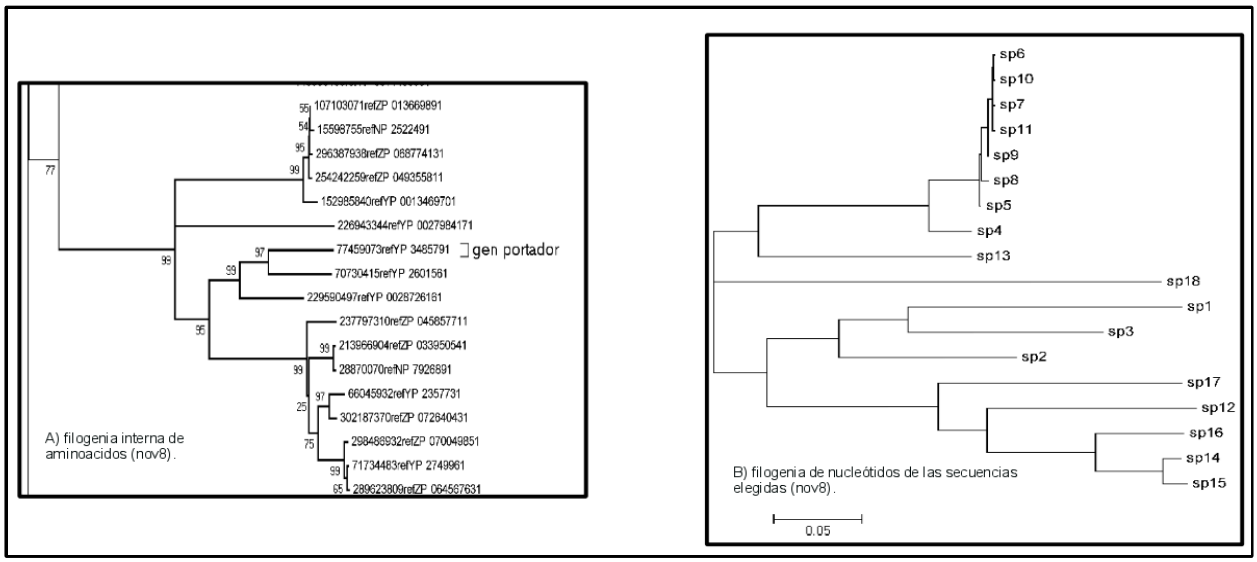

Figura 8 .- Ejemplo de la filogenia de nov 8. A) selección de una rama interna realizada con las secuencias de aminoácidos por máxima *verosimilitud* con valores de boostrap mayores de 70% utilizado para su elección. B) Filogenia de máxima *verosimilitud* los nombres originales fueron sustituidos por :>sp-1-|255961261:3524719-3526674; >sp-2-|70728250:4125427-4127382; >sp-3-|229587578:2896478-2894520 ; >sp-4-|152983466:3411032-3412987; >sp-5-|296279672:5485-7440; >sp-6-|107098994:2742574-2740586; >sp-7-|254234166:1054169-1052214 ; >sp-8-|116048575:3449781-3451736; >sp-9-|254239388:1173670-1171715; >sp-10-|110645304:2184407-2182452; >sp-11-|218888746:3681414-3683369 ; >sp-12-|170719187:3013103-3015064; >sp-13-|146305042:2241704-2239749; >sp-14-|26986745:3455887-3453935; >sp-15-|148545259:3031795-3033747; >sp-16-|167031021:3141287-3143245; >sp-17-|104779316:3378506-3380458; >sp-18-|223019529:5020-6978; que corresponden a los genes elegidos en sus versiones de nucleótidos.

4.3 Edición y predicción de traslapes en las secuencias de los portadores.

Las alineaciones fueron editadas con el programa *Bioedit,* dividiendo las secuencias en regiones traslapadas y no traslapadas, basándonos en las predicciones del traslape previamente reportadas (figura.- 9).

La predicción de las regiones traslapadas para cada uno de los genes se calculó con el programa FramePlot 2.3.2 directamente en su sitio (http://www.nih.go.jp/~jun/cgi-bin/frameplot.pl). que permite localizar todos los marcos de lectura de la secuencia (ver Anexo 3A). En las figuras del anexo 3A se pueden ver los posibles traslapes que presentan un codón de inicio y un codón de paro en el mismo marco de lectura lo que supondríamos pudiera ser una región transcripcionalmente activa, entre las

coordenadas del traslape del gen portador, se eligieron aquellos que estuvieran no más de 12 pares de bases (pb) río arriba y río abajo del reportado, además debía de tener una longitud aproximada del original y presentar el mismo marco de lectura, para los genes Nov15, cosA y htgA se encontró que las regiones predichas no cumplían con los parámetros establecidos tenían longitudes muy pequeñas y no se encontraban en el mismo marco de lectura, por lo tanto no se eligieron otras secuencias con posibles traslapes.

Las predicciones realizadas de los genes se utilizaron para el análisis de tasa de sustitución proponiéndolas con la misma omega que el gen reportado, esto descartaría que los resultados estuviesen ocultando evidencia de selección negativa por el hecho de analizar un gen contra todo un grupo. Los análisis realizados reflejan que en todos los casos las variaciones en los resultados no arrojan diferencias significativas entre las tasas de sustitución de ambos análisis. En los dos casos (de cada gen) se obtuvieron resultados similares (reportados mas adelante), los análisis de los genes con más de un traslape no se reportaron en este trabajo, aunque se tienen los resultados en el Anexo 4A.

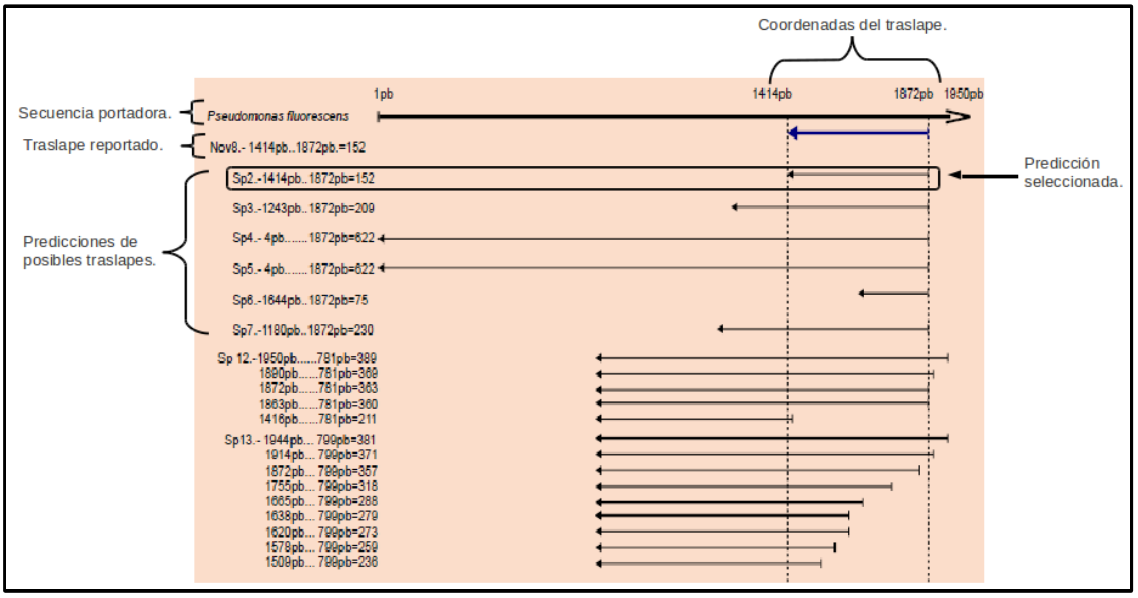

Figura 9 .- Diagrama que representa las predicción hechas por Orfinder. Se encuentran ordenadas en orden descendiente, los tamaños representados en la figura son proporcionales a los tamaños predichos.

4.4 Estimación de tasas de sustitución PAML.

Para realizar el análisis de tasas de sustitución se utilizaron los tres archivos creados que contenían, a) la alineación en formato phylip (.sphy), b) las filogenias del gen portador correspondiente y c) un script con los parámetros del análisis como se observa en las imágenes anteriormente mostradas (figura.- 7 y 8), como se ha mencionado de las secuencias se obtuvieron las tasas de sustitución con el programa *codeml* bajo dos tipos de supuestos que se contraponen.

El supuesto del **M0** (ver métodos) toma en cuenta que la tasa de sustitución es similar para cada una de las ramas y cada uno de sus clados, dichos valores denominados omega (ω) no varían entre las ramas y se toma como hipótesis nula. En otras palabras, bajo este modelo suponemos que no existe un efecto de selección negativa, debido a que las tasas de sustituciones a lo largo de las secuencias son similares. Por lo tanto todas las secuencias cambian a la misma velocidad.

Bajo el supuesto **M2,** suponemos que la tasa de sustitución varía entre las ramas de la filogenia, en donde se toman dos tipos de omega, a) una para aquellas secuencias que son seleccionadas por tener un traslape verdadero o uno predicho y b) en donde no existe un gen traslapado dentro de las secuencias, los resultados o valores relativos (LRT) de ambos modelos se evaluaron mediante una prueba estadística de chi^2 permitiéndonos rechazar alguna de las hipótesis. (Anexo 5A).

Los resultados se resumen en las siguientes figuras 10A y 10B además de encontrarse en el anexo 5A los resultados completos.

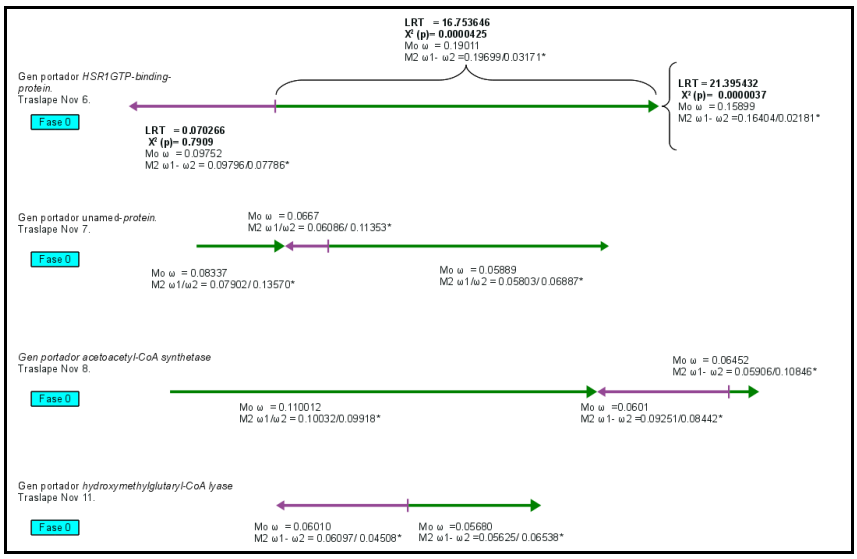

Figura 10A.- Resultado de los análisis de tasas de sustitución de los genes nov's del 6 – 11, las flechas magentas representan las regiones que corresponden al traslape, en verde las regiones no traslapadas. Las flechas corresponden a la división del gen portador, se resumen el nombre que corresponde al portador, la fase en la que se encuentran, valores de LRT, X^2 para los casos donde se encontró evidencia de selección negativa, valores omega para todos lo genes y para las porciones completas no se representan los valores omegas solo en los casos en donde se encontró selección negativa. Se encontró solo un gen bajo selección negativa nov'6.

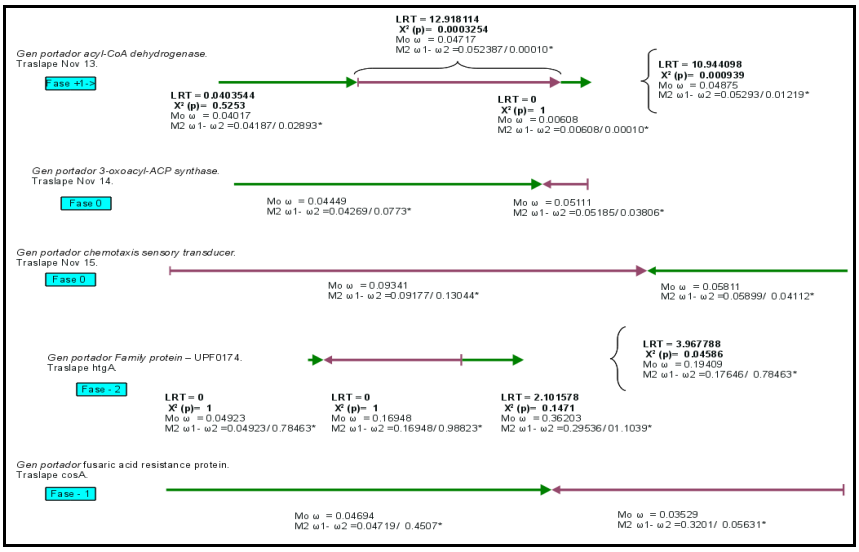

Figura 10B.- Resultado de los análisis de tasa de sustitución, representados por flechas verdes que representan las regiones no traslapadas y en magenta las regiones traslapadas, se encontró efectos de selección negativa en dos genes mas nov's 13 que se encuentra con la misma dirección además en fase + 1 y el gen htgA que se encuentra en fase -2.

En todos los casos se calcularon los valores de *máxima verosimilitud* de las proporciones de sustituciones sinónimas y no sinónimas, las cuales se utilizaron para obtener los valores relativos

[LRT= 2 ($Lnl_2 - Lnl_1$)] como se ha mencionado anteriormente, los productos se compararon bajo una prueba de chi^2 con 1 grado de libertad, los resultados con un error estándar mayor a .05 nos permite rechazar los supuestos de la hipótesis alternativa.

Los resultados de cada uno de los análisis se resumen en las tablas siguientes.

Tabla 2.- valores de *máxima verosimilitud* de M0 y M2, valores relativos, chi^2 y omegas utilizadas en el análisis, en cada uno de las regiones en las que se dividió el gen portador Nov6, Nov7.

Nov6.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	1785pb .	1302pb	483 pb	
fase.	0 compl.			
lnL-M0	24446.667264	18378.382139	5928.733776	
lnL-M2	24435.969548	18370.005316	5928.698643	
LRT [(2)(M2-M0)]	21.395432	16.753646	0.070266	
Chi2 - 1g.l. (P)	0.000003737	0.0000425	0.7909	
omega M0	0.15899	0.19011	0.09752	
omega M2	0.16404/ 0.02181*	0.19699/ 0.03171*	0.09796/ 0.07786*	
Nov7.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	1392pb.	291pb.	129pb	972pb.
fase.	0 compl.			
lnL-M0	8179.678927	1923.563955	808.307875	5378.371573
lnL-M2	8178.920751	1923.563955	807.953431	5378.236446
LRT [(2)(M2-M0)]	1.516352	1.211104	0.708888	0.270254
Chi2 - 1g.l. (P)	0.2182	0.2711	0.3998	0.6032
omega M0		0.08337	.0667	0.05889
omega M2		0.7902/ 0.13570*	0.06086/ 0.11353*	0.05803/ 0.6887*

4.4.1.- Nov 6.

El caso de Nov 6 es un traslape del gen *HSR1GTP-binding-protein* con una longitud de 1785 pb. que se divide en dos regiones una no traslapada de 1302 pb. y una traslapada que consta de 483 pb. se localiza en la región que correspondería al carboxilo terminal del producto, es divergente y se encuentra en fase 0 con respecto a su portador (tabla.-2).

Los resultados obtenidos en el análisis del gen completo nos permiten rechazar el supuesto del modelo 0, la obtención de un valor P menor al 0.05, por lo tanto se infiere que existe un restricción

evolutiva que evita que el gen tenga una tasa de sustitución menor al de sus homólogos.

Con respecto a sus regiones en las que se dividió se encontró, como se puede observar en la tabla 1 que en la región no traslapada que representa el 73% del gen no se encuentra algún efecto de selección negativa, el valor de P no es menor de 0.05 del esperado por el análisis de chi^2, debido a que la tasa de sustitución es igual tanto para el gen como para los demás homólogos que conforman el grupo, con estos resultados podemos decir que no hay indicio de alguna restricción selectiva que se esté ejerciendo en ninguno de los genes tanto en el traslape como en el portador en esta región.

La región traslapada que se localiza en el extremo 3´ y representa el 27% del gen nos muestra que los cambios en la tasa de sustitución es menor, teniendo un valor de P por debajo al 0.05 lo cual pone en evidencia un fenómeno de restricción selectiva del tipo negativo para esta porción de secuencia, la distribución del traslape con respecto a su portador es complementaria en fase anti paralela lo que quiere decir que los cambios que afecten a la posición 1 alteran a la posición 3 del otro gen y por lo tanto mientras en un gen el producto pudiera cambiar el otro no se vería afectado debido a que la posición 3 no es fundamental para el producto resultante.

Al analizar los datos en conjunto podemos darnos cuenta que el tamaño del traslape es pequeño en comparación a la región no traslapada y suponemos que es un factor que limita la observación de las restricciones selectivas en la secuencia del portador.

Como se ha mencionado anteriormente, en las distintas regiones que conforman un gen se encuentran implicados varios factores que provocan que esté bajo distintas presiones de selección a lo largo de su secuencia, a) el grupo funcional en donde se encuentra contenido los dominios que son funcionales, b) la región 5` que dirige la síntesis del producto y en donde se encuentran las secuencias que controlan la transcripción del producto (en la mayoría de los casos) y c) la región terminal en donde se detiene la síntesis tanto de los polímeros de DNA, RNA y/o proteicos, estos últimos delimitan el tamaño y por lo tanto los cambios en esta región modifican tanto la longitud y como consecuencia los posibles plegamientos del producto. EL gen nov 6 se localiza en la región 3' la cual podría estar influyendo en los efectos observables en las tasas de sustitución ya sea por la presencia del traslape o por lo antes mencionado, basados en los resultados suponemos que en el análisis del portador existe una relajación en el efecto de la selección debido al presencia del traslape.

4.4.2.- <u>Nov 7.</u>

El portador de Nov 7 es una proteína sin caracterizar con una longitud de 1392 pb. que se divide en tres regiones, dos no traslapadas de 291 pb que corresponde al 21%., otra de 972 pb. que representa el 70% respectivamente y una traslapada que consta de 129 pb. y tiene el 9% del tamaño total, se localiza en una región próxima al carboxilo terminal, es divergente y se encuentra en fase 0 con respecto a su portador (tabla.-2).

Se aceptó el modelo 0 en los 4 casos, las secuencias tienen una tasa de sustitución constante en el gen completo como en sus regiones en las que se divide. Si observamos la longitud relativa de las secuencias veremos que el tamaño del traslape es pequeño por lo que, si existe un efecto de selección sobre ella no se puede observar por este método debido a que las sustituciones en las secuencias que podemos registrar son proporcionales a la longitud, en el análisis el efecto que pudiera imponer un traslape pequeño en el gen completo se ve anulado por la distancia de la secuencia, por otro lado se podría esperar que al no encontrarse en una región catalítica o estructural importante para el producto, el efecto sobre el portador disminuya y termine con un costo informacional nulo.

Tabla 3.- valores de *máxima verosimilitud* de M0 y M2, valores relativos, chi^2 y omegas utilizadas en el análisis, en cada uno de las regiones en las que se dividió el gen portador Nov8 y Nov11.

Nov8.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	1992pb.	1449pb.	456pb.	78pb.
fase.	0 compl.			
lnL-M0	13698.49335	10118.078875	2968.655382	553.566034
lnL-M2	13698.490671	10118.0044	2968.416693	553.549562
LRT [(2)(M2-M0)]	0.005358	-0.14895	1.077378	0.032944
Chi2 - 1g.l. (P)	0.9416	0.6995	0.2993	856.0000
omega M0	0.09861	0.10012	0.09120	0.06452
omega M2	0.09869/ 0.09822*	0.10032/ 0.09918*	0.09251/ 0.08442*	0.05906/ 0.10846*
Nov11.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	993pb.		489pb.	504pb
fase.	0 compl.			
lnL-M0	5996.785369		2625.505805	3353.913565
lnL-M2	5996.785228		2625.409658	3353.867683
LRT [(2)(M2-M0)]	0.000282		0.192294	0.091764
Chi2 - 1g.l. (P)	0.9866		661.000	0.7619
omega M0	0.06064		0.06010	0.05680
omega M2	0.06062/ 0.06102*		0.06097/ 0.04508*	0.05625/ 0.06538*

4.4.3 .- <u>Nov 8.</u>

El portador de Nov 8 es una proteína *acetoacetyl-CoA synthetase* con una longitud de 1992 pb. que se divide en tres regiones, dos no traslapadas de 1449 pb. que corresponderían al 73%, otra de 78 pb. que representa el 4 % y una traslapada que consta de 456 pb. o el 23% total de la secuencia del gen portadores localiza en una región circundante al amino terminal, es convergente y se encuentra en fase 0 con respecto a su portador (tabla.-3)

En este caso ocurrió lo mismo que los análisis anteriores, donde no encontramos algún efecto de la selección negativa en las regiones en las que se divide el portador como en el gen completo. En el análisis de predicción de posibles traslapes no se encontró alguno que fuera viable de tener un producto en los homólogos, por lo que se puede suponer que el traslape sea falso y que aparente no ser traduccionalmente activo, por lo tanto suponemos que la relación entre las variaciones en los cambios de las secuencias del grupo es constante y que no es evidente algún indicio de selección negativa o simplemente no existe dicho efecto.

4.4.4.- <u>Nov 11.</u>

El portador de Nov 11 es una proteína *hydroxymethylglutaryl-CoA lyase* con una longitud de 993 pb. que se divide en dos regiones, la primera traslapadas de una longitud de 489 pb., y una no traslapada que consta de 504 pb. se localiza en una región circundante a lo que correspondería al carboxilo terminal, es divergente y se encuentra en fase 0 con respecto a su portador (tabla.-3)

El gen portador se dividió en 2 regiones a) una no traslapada que representa el 49%, y b) una traslapada que correspondía al 51%. Las proporciones de cambios sinónimos y no sinónimos en las secuencias se mantienen continuas a en todo el grupo. No encontramos rastro de que exista selección negativa en el gen portador o en alguna de las regiones en que se divide, no se encontró una correlación entre la longitud, la fase y las restricciones evolutivas, no se encontraron probables traslapes en los homólogos, al parecer el traslape “si no es hipotético” no da un producto con relevancia evolutiva evidente, el costo informacional por estar en fase 0 al parecer es bajo tanto para el traslape como para el portador y por consiguiente no se puede ver alguna selección negativa a lo largo de la secuencia.

Tabla 4.- valores de *máxima verosimilitud* de M0 y M2, valores relativos, chi^2 y omegas utilizadas en el análisis, en cada uno de las regiones en las que se dividió el gen portador Nov13 y Nov14.

Nov13.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	1173pb.	462 pb.	633 pb.	78pb.
fase.	1+ →			
lnL-M0	5488.427647	2205.97895	2941.885038	294.899413
lnL-M2	5482.955598	2205.777178	2935.425981	294.899413
LRT [(2)(M2-M0)]	10.944098	0.403544	12.918114	0
Chi2 - 1g.l. (P)	0.000939	0.5253	0.0003254	1
omega M0	0.04875	0.04017	0.04717	0.00608
omega M2	0.05293/ 0.01219*	0.04187/ 0.02893*	0.052387/ 0.00010*	0.00608/ 0.00010*
Nov14.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	1221	1029	192	
fase.	0			
lnL-M0	5571.933414	4558.148493	976.103385	
lnL-M2	5571.116084	4556.758741	976.039565	
LRT [(2)(M2-M0)]	1.63466	2.779504	0.12764	
Chi2 - 1g.l. (P)	0.2011	0.09548	0.7209	
omega M0	0.04858	0.04449	0.05111	
omega M2	0.04729/ 0.07707*	0.04269/ 0.08773*	0.05185/ 0.03806*	

4.4.5.- Nov 13.

El portador de Nov 13 es una proteína *acyl-CoA dehydrogenase* con una longitud de 1173 pb. que se divide en tres regiones, la primera no traslapada con una longitud de 462 pb., la segunda es traslapada que consta de 633 pb. y la tercera de 78 pb. se localiza en una región circundante a lo que correspondería al carboxilo terminal, es unidireccional (->) y se encuentra en fase +1 con respecto a su portador (tabla.- 4).

El gen portador se dividió en 3 porciones que corresponden a) la región no traslapada al 39%, b) la región traslapada al 54% y por último la región no traslapada 2 al 7%, como se hace evidente el tamaño de la región traslapada es mayor lo que podría ayudar en la capacidad de detección del método, asi como en el efecto que hay sobre la secuencia; como se ha mencionado antes el caso de la región no traslapada 2, es muy pequeña para que el método pueda detectar alguna restricción evolutiva.

De los traslapes predichos por Krakauer se considera que la fase en que esta Nov 13 es la segunda más común en las bacterias, este caso es peculiar en comparación a los demás genes analizados aquí, se puede observar que la longitud del traslape, el marco de lectura que presenta es +1 y

la dirección del traslape pueden ser importantes en su conjunto para que la selección negativa este ejerciendo presión en la secuencia afectando directamente al portador.

Los cambios que pudieran aparece en la posición 1 y 2 corresponden a la posición 2 y 1 del gen complementario, por lo que podemos ver que los cambios sinónimos para un gen se traducen en no sinónimos para el otro, por lo que las restricciones al cambio en estos sitios son muy fuertes la carga informacional presente podría aumentar las restricciones evolutivas en la secuencia. La posición 3 altera de diferente manera a su correspondiente por lo que no habría razón de elevar el costo en este sitio, debido a que en ambos genes se presenta la misma posibilidad de cambio, por lo tanto la fase en la que se encuentra el traslape es fundamental para mantener con una tasa de sustitución disminuida en la secuencia compartida por ambos genes y en el gen portador.

4.4.6.- Nov 14.

El portador de Nov 14 es una proteína *3-oxoacyl-ACP synthase* con una longitud de 1221 pb. que se divide en dos regiones, la primera no traslapada con una longitud de 1029 pb. y la segunda que consta de 192 pb. se localiza en la secuencia que correspondería al amino terminal del producto, es un traslape convergente y se encuentra en fase 0 con respecto a su portador (tabla.-4).

El portador del gen se dividió en 2 regiones que constaban una no traslapada que correspondería al 84% del total de la secuencia y la fracción traslapada de un tamaño del 16% respectivamente,

El caso de Nov 14 sigue el patrón en el que los valores de *máxima verosimilitud* son muy similares en ambos modelos para el gen completo y los fragmentos en que se divide, si observamos el tamaño de la secuencia traslapada podemos ver que es mucho menor a la región no traslapada por lo que suponemos que el tamaño influye en la búsqueda de alguna señal de restricción selectiva en la secuencia. Se esperaría que la tasa de sustitución se vería disminuida en donde correspondan ambas secuencias pero en este caso no es así por lo que nos permitimos asumir que la diferencia de longitudes en la secuencia compartida el método no percibe las restricciones selectivas en regiones traslapadas con longitudes pequeñas.

Tabla 5.- valores de *máxima verosimilitud* de M0 y M2, valores relativos, chi^2 y omegas utilizadas en el análisis, en cada uno de las regiones en las que se dividió el gen portador Nov15 y HtgA.

Nov15.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	2229		1545	684
fase.	0			
lnL-M0	25187.714991		6890.724438	18170.224166
lnL-M2	25187.20576		6890.318744	18169.256935
LRT [(2)(M2-M0)]	1.018462		0.811388	1.934462
Chi2 - 1g.l. (P)	0.3129		0.3677	0.1643
omega M0	0.08304		0.09341	0.05811
omega M2	0.08216/ 0.10145*		0.09177/ 0.13044*	0.05899/ 0.04112*
htgA.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	711	42	486	183
fase.	2- compl.			
lnL-M0	1950.283254	74.066172	1274.428079	515.251669
lnL-M2	1948.289936	74.066172	1274.428079	514.20088
LRT [(2)(M2-M0)]		0	0	2.101578
Chi2 - 1g.l. (P)	0.04586	1	1	0.1471
omega M0	0.19409	0.04923	0.16948	0.36203
omega M2	0.17646/ 0.78463*			0.05899/ 0.04112*

4.4.7.- Nov 15.

El gen portador de Nov 15 es una proteína *chemotaxis sensory transducer* con una longitud de 2229 pb. que se divide en dos regiones, la secuencia traslapada con una longitud de 1545 pb. y la no traslapada consta de 684 pb., Nov15 se localiza en la secuencia que correspondería al carboxilo terminal del producto, es un traslape convergente y se encuentra en fase 0 con respecto a su portador (tabla.-5).

El caso de Nov15 sigue los mismo patrones se dividió en dos fragmentos que representaban a) una porción traslapada el 69% y una no traslapada de 31% del total del gen portador, el traslape tiene un tamaño considerable, basándonos en su tamaño y en su posición esperaríamos encontrar selección negativa lo que no fue así, los datos obtenidos nos permiten suponer que el producto del traslape se localiza en una zona donde los efecto de la selección son mínimos o simplemente el traslape no forma una secuencia transcripcionalmente activa y lo cual evita en ambos casos observar las huellas de alguna restricción evolutiva.

4.4.8.- *htgA.*

El portador de htgA es una proteína de la familia *UPF0174 - yaaW*, el gen consta de una longitud de 711pb. se encuentra dividido en tres secciones dos no traslapadas de 42pb. que correspondería al 6%, y de 183pb que representaría el 26%. Respectivamente y una región traslapada que consta de un tamaño de 486 pb. que consta del 68% del total de la secuencia, htgA esta en fase de -2 se localiza en el centro del gen por lo tanto se puede clasificar en convergente o divergente (tabla.-5).

El gen htgA es un caso muy diferente a lo encontrado en los demás genes tanto por su clasificación como por su fase en la que se encuentra. Este análisis fue el único caso en donde la omega del portador era compartida por 9 secuencias homologas y por lo tanto tienen una distancia evolutiva entre ellas de 0 por el hecho de que el gen está altamente conservado en este pequeño subgrupo.

Ahora bien, al medir la tasa de sustitución de la secuencia completa del portador encontramos que hay diferencias significativas entre los modelos que nos permite desechar la hipótesis nula, a lo contrario encontrado al revisar los fragmentos por separado donde no observamos selección negativa en ningún caso.

Si tomamos en cuenta el costo informacional que tiene cada uno de los nucleótidos de un codón podemos observar más a detalle lo que sucede en este caso. Por el tipo de traslape los nucleótidos de la posición 3 comparten el sitio 1 y 2 de ambos genes, por lo que las sustituciones en este lugar no alteran el producto del gen complementario por lo que las sustituciones no se ven comprometidos a un solo tipo, sin cambio la posición 2 que corresponde al sitio 1 si afecta el producto debido a que una sustitución sinónima se convierte en una no sinónima para el otro alterando el producto de uno de los dos genes, por lo que esto puede estar favoreciendo una tasa constante en el sitio 2-1 confiriéndole un mayor costo informacional a este lugar y permitiendo que la selección negativa influya en la secuencia.

Si la restricción evolutiva está basada en que tanto se permite cambiar un nucleótido sin alterar un producto y observamos que el marco de lectura en que se encuentran permite que se presenten cambios en la posición 2 y 1 que corresponden a el sitio 3 sin alterar uno de los productos, aunado al hecho de que las sustituciones en la posición 1 del traslape afecta el producto final del portador en su posición 2, podemos suponer que las restricciones evolutivas radican en el nucleótido de la posición 1 del traslape y 2 del portador cargando así con todo el peso de la restricción evolutiva que pudiese

encontrarse en la secuencia. Una propuesta por la cual no se tenga evidencia de selección negativa en las regiones por separado, pudiera ser por el hecho de que hay poca distancia evolutiva dentro de la filogenia, por lo que tal vez, no pueda registrar el método un efecto de selección, lo que remarcaría el hecho de que es susceptible a la cohesión dentro del grupo.

Tabla 6.- valores de *máxima verosimilitud* de M0 y M2, valores relativos, chi^2 y omegas utilizadas en el análisis, en cada uno de las regiones en las que se dividió el gen portador cosA.

cosA.	Gen Completo	Región No- traslap.	Región traslapada.	Región No-traslap.
Numero de bases.	2247pb.	1227	1020	
fase.	-1			
lnL-M0	11523.138577	6270.17547	5162.477811	
lnL-M2	11523.118971	6270.173475	5161.469962	
LRT [(2)(M2-M0)]	0.039212	0.156848	2.015698	
Chi2 - 1g.l. (P)	843.000	0.6921	0.1557	
omega M0	0.04694	0.04412	0.03527	
omega M2	0.04719/ 0.04507*	0.04401/0.04486*	0.03201/ 0.05631*	

4.4.9 .- ***cosA.***

El gen portador de cosA es una proteína *fusaric acid resistance protein* consta de 2247 pb. y se divide en dos regiones una sección no traslapada de 1227 pb. que representa el 54% y otra traslapada de 1020 pb. cosA que corresponde al 46% del total de la secuencia del portador, es un traslape convergente y se encuentra en fase -1 con respecto a su portador (tabla.-6).

El portador de cosA no presenta diferencias significativas entre los modelos en ninguno de los 3 casos analizados, el traslape ha sido comprobado por métodos experimentales lo cual descarta el hecho de que no sea real, los tamaños de los fragmentos en que se divide son relativamente similares y grandes por encima de las 1000 pb., si analizamos con detalle la región compartida como se puede observar en la imagen (figura.- 11) cambio en la posición 1 afecta la posición 2 en ambos casos, el sitio 3 tiene la característica de que en ambos genes se localizan en el mismo sitio, por lo tanto los cambios

presentado en este sitio con un costo informacional bajo alteran de igual manera a ambas caras de las secuencias lo que nos dice que probablemente la tasa de sustitución en ambos genes se dé a la misma velocidad y por consiguiente en el análisis no se pueda observar una restricción evolutiva por efecto del marco de lectura del traslape. Si además tomamos en cuenta que la región compartida se encuentra en la región 5´ que se ha propuesto como una zona en donde hay menos sustituciones por sitio debido al costo que tiene el sitio de inicio de la traducción del producto nos permite suponer que la selección negativa si actúa sobre la secuencia el efecto es mínimo y no se puede detectar.

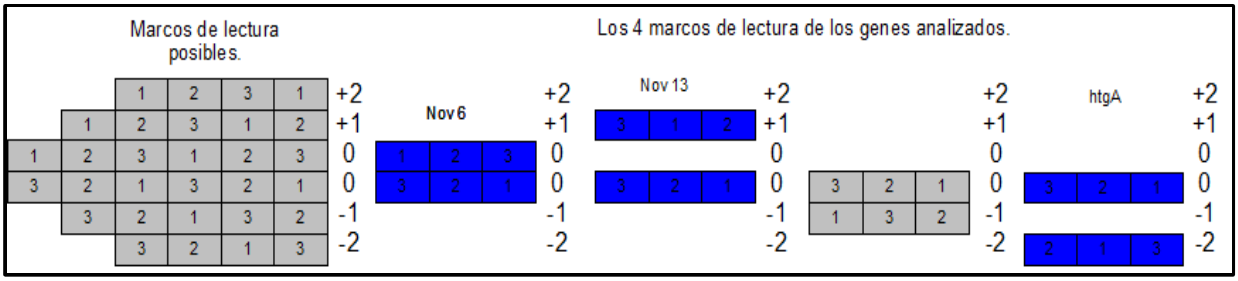

Figura 11.- Representación de los marcos de lectura posibles para dos secuencias antiparalelas, en el análisis se encontraron 4 tipos de traslapes, el traslape en fase 0 Nov 6, en fase +1 Nov13, en fase -1 y en fase -2 htgA, de los cuales se marcan en azul aquellos en donde se encontró selección negativa.

Ahora bien si resumimos los resultados obtenidos, tenemos que de los 9 casos tenemos solo 3 con evidencia de selección negativa, encontramos que los efectos se encontraron en Nov 6 en el análisis del portador y en la región no traslapada; en Nov 13 se observó en el gen portador además, en la secuencia compartida por ambos genes y por último, el caso de htgA donde se encontró evidencia solo el análisis del gen portador.

De los genes que presentaron efectos de selección negativa Nov6 presentaban un traslape en fase 0 en la región que correspondería al carboxilo terminal, Nov13 en fase +1 también en la sección del carboxilo terminal y htgA en fase -2 que se localiza en la región central del portador. Las longitudes de los diferentes vemos que Nov6 tiene solo un 27% de la secuencia del portador se encuentra compartida por ambos genes, el 73% aproximado no se encuentra traslapado; Nov13 tiene el 54% pero el portador se divide en 2 regiones no traslapadas de 7% y 39% aproximadamente y por último htgA que tiene el fragmento compartido más grande con un 68% contra 26% y 6% de las regiones no traslapadas.

Las fases en las que se encuentran los genes no dan pauta para decir que las restricciones evolutivas presentes en las secuencias, son debido al costo informacional presente en ellas, tomando mas a detalle los 3 casos, observamos como se ha mencionado antes que la posición de nucleótidos permite o da pauta para que la selección actúe sobre los distintos codones que conforman las secuencias transcripcionalmente activas.

Nov 6 presenta un traslape en fase 0, lo que significaría que la posición 1 de ambos genes corresponden a la posición 3 de su correspondiente, el costo informacional es alto para la posición 1 pero no para la 3, por lo que la restricción que presenta el tercer nucleótido es muy tenue, por lo que cambios el efecto se ve sobre la posición 1, por lo que la posición 3 no puede cambiar porque si esto ocurriese la posición 1 se vería afectada y por tanto el producto lo que provocaría posible pérdida del producto. Los costos presentes en la posición 2 son los mismo para ambos genes, lo cual mantiene constante las sustituciones en este sitio. El tamaño de la región traslapada es pequeña con respecto a la no traslapada y por lo tanto no se puede registrar un efecto en esta fracción de secuencia, proponemos que el tamaño es el factor que evita registrar alguna restricción evolutiva debido a que en otros casos donde se presentaba un traslape sucede lo mismo, si suponemos además que la región donde se localiza se encuentra muy conservada, el método pudiera tener un sesgo como se ha mencionado antes a el tamaño de la secuencia y la consistencia en la filogenia.

El caso de Nov 13 consta de una fase en la cual las posiciones de los nucleótidos se ven más comprometidas debido a la carga informacional que tienen cada uno de sus sitios se puede observar que los nucleótidos 1 y 2 se encuentran compartiendo el sitio 2 y 1 correspondientemente, por lo que una sustitución en la posición 3 tiene el mismo costo informacional para ambos genes, como se ha mencionado anteriormente, los sitios 1 y 2 se ven comprometidos a permanecer sin cambios evitando así la pérdida de uno de los productos lo que aumenta la posibilidad de que la tasa de sustitución se vea disminuida tanto en todo gen portador como en la región traslapada. Además el tamaño del traslape es considerable en comparación a las otras dos regiones en las que se dividió el portador, lo que puede facilitar al método la detección de selección negativa.

El gen htgA tiene una particularidad en la que la secuencia compartida es mucho mayor a la que no se encuentra traslapada, además la fase en la que se encuentra es una de las más comunes entre las predichas por Krakauer por lo tanto se esperaría que hubiese evidencia de selección negativa, tanto el análisis del portador como en la región traslapada, caso que no fue así. Si analizamos la fase en la que

se encuentran los nucleótidos podemos ver en la figura 11 que el portador presenta las posiciones 3, 2 y 1 que corresponden a las posiciones 2, 1 y 3, lo cual no habla de que cualquier cambio en cualquier posición de ambos genes alteran sin duda el producto complementario, los costos informacionales se encuentran divididos en las 3 posiciones de igual manera los cambios permitidos en la posición 3 del portador alteran el sitio 2 del traslape, las sustituciones en el sitio 2 del portador afectan el nucleótido 1 y el sitio 1 del portador es aquel que puede cambiar sin alterar el producto del traslape pero por la carga de información que subyace en este sitio las sustituciones están limitadas por que se alteraría el producto del mismo portador, por lo tanto el traslape se encuentra en una posición que permite además de estar bajo selección negativa la permanencia de ambos genes por estar bajo dos fuertes efectos selectivos, lo que nos permitiría proponer que el aumento en las restricción evolutiva están fuertemente relacionadas con el tipo de fase en las que se encuentran los traslapes y a su vez la detección de dichas restricciones están influenciadas además por la longitud y posición del traslape.

Con respecto a la posición de los 3 casos que se encontraron con evidencias de selección negativa podemos decir que pudiera favorecer su observación si no se encuentra en alguna región en donde se encuentren *per se* un fuerte efecto selectivo, porque podría enmascarar un efecto al provocado por la aparición de un traslape. En Nov6 y Nov13 se encuentran proximales al carboxilo terminal que presenta un costo informacional relativamente alto, en el caso del gen htgA se encuentra en el centro por lo que no podríamos definir si es un factor que aumente o disminuya las restricciones evolutivas. Se necesitaría un muestreo más grande de genes traslapados para suponer que los genes que se encuentran en las regiones proximales al carboxilo o amino terminal promueven o disminuyen las restricciones evolutivas que los traslapes provocan.

5.- Métodos Experimentales.

La parte experimental fue realizada con las secuencias reportadas de los genes *htgA y cosA* en trabajos anteriores y obtenidas de la base de datos de NCBI-GenBank (Delaye et al., 2008; Kim et al., 2009), se encontró que el gen *htgA* tenia dos secuencias reportadas las cuales se ocuparon para la síntesis automática la cual fue realizada por la compañía Genscript, se nombraron *htgA-l* de 597 pb., *htgA-s* de 492 pb. y *cosA* una longitud de 1026 pb.,

5.1 Optimización de secuencias *cosA y htgA* en sus dos versiones.

Cada una de las secuencias fueron optimizadas cambiando nucleótidos que fueran compatibles con el uso de codón de *E. coli,* sin alterar el producto y así tener mayor rendimiento al momento de la inducción y purificación, como se muestra en el anexo 6A.

5.2 Cepas *E. coli dh5α, E. coli* BL21, plásmidos puc57 y pet-19b.

La bacteria *E. coli* es la especie más empleada y estudiada tanto genética como fisiológicamente. Entre las ventajas que brinda este microorganismo como hospedero están: rápida generación de biomasa (elevada velocidad de crecimiento), fácil manipulación genética, no posee requerimientos costosos asociados a medios de cultivo o equipamiento, alta eficiencia en la incorporación de material genético foráneo, gran variedad de vectores de expresión y variantes mutantes **(Jonasson y col., 2002)**. Los parámetros importantes para una exitosa producción de proteínas recombinantes en *E. coli* incluyen eficiencia transcripcional y traduccional, estabilidad del vector de expresión y de los mRNA transcritos, estabilidad proteolítica, localización y plegamiento de la proteína **(Jonasson y col., 2002)**.

Las cepas utilizadas en este trabajo fueron *E. coli* BL21 con el genotipo [*(DE3) pLysE; hsdS, gal, lcIts857, ind1, Sam7, Nin5,lacUV5-T7genel]* y *E. coli dh5-α* con el genotipo [*fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*].

La cepa *E.coli BL21* es deficiente en proteasa lon y carece de una de membrana externa ompT, esta proteasa puede degradar proteínas durante la purificación. Además, presenta altos niveles de expresión y la proteína recombinante es fácilmente inducida (lactosa o IPTG). La ventaja más grande se asocia a la expresión basal de T7 RNA polimerasa. La T7 RNA polimerasa es altamente selectiva y activa, cuando se induce fuertemente casi todos los recursos de la célula se dirigen hacia la producción de la proteína recombinante, los productos pueden corresponder hasta en un 50% a la proteína de interés.

La cepa *dh5-α* tiene características que permiten su transformación con un alto grado de eficiencia lo que permite que sea utilizado en una gran cantidad de métodos de DNA recombinación, presenta una mutación en la regiones *endA1* que inactiva endonucleasas intracelulares, que podrían

degradar plásmidos invasivos, además tiene una mutación *hsdR17* que elimina endonucleasas del sistema RMS de *E.coli k12* lo que permite al sistema hsdR metilar el DNA para su degradación, presenta *recA* que elimina recombinaciones homologas, permite que los plásmidos permanezcan estables y evita la multimerización de plásmidos, por ultimo presenta un supresor de tipo *amber* el cual restaura una mutación sin sentido en las secuencias, restaurando la función proteica.

Los plásmidos utilizados en el trabajo son principalmente pet-19b y puc57 las secuencias se encuentran en el anexo 6A plásmidos.

5.3 Medios y antibióticos.

Se utilizó como medio de cultivo para el crecimiento de las cepas LB (LB 1litro: Bacto-triptona 10g; Extracto de levadura 5g; $Nacl_2$ 10g; aforar a 1 litro de agua miliQ), además se suplementaron con ampicilina de una solución concentrada conservada a -20ºC, se preparó una solución concentrada a 100 mg/ml en agua bidestilada y se esterilizó.

5.4 Transformación de *E. coli dh5α y E. coli BL21*.

Las cepas *E. coli dh5-α y E.coli BL21* fueron transformadas a partir de cultivos competentes hechos con el protocolo de quimio-competencia :

1.-De un preinóculo de 10 ml de medio SOB crecido a 37°C toda la noche.
2.-Tomar los 10 ml e inocular 500ml de medio LB, incubar a 37°C en agitación entre 1hr. 30min. - 2hr. hasta tener una densidad óptica de .5 - .7.
3.-Transferir el medio de cultivo a botellas de centrifuga estériles de 250 ml previamente enfriadas a -20°C.
4.-Incubar 30 min. en hielo (cubrir completamente las botellas).
5.-Centrifugar a 2500 rpm por 15min a -4°C.
6.-Resuspender la pastilla en 20 ml de solución RF1 [RF1 (por 100ml): 1,2g de RbCl;0,99g $MnCl_2$ x4H_2O;0.294 g de acetato potásico; 0;15 g de $CaCl_2$ X H_2O;11,9 ml de agua destilada hasta 100ml, pH

5,8]

7.-Incubar 15 min en agua con hielo.

8.-Centrifugara 2500 rpm por 15 min a -4°C.

9.- Resuspender la pastilla en 4 ml de solución RF2 a -4°C [RF2 por 50ml: 0,1046 g de ácido morfolino propanosulfónico (MOPS); 0,06 g de RbCl; 0,55 g de $CaCl_2$ X H_2O; 5,95 ml de glicerol; agua destilada hasta 50 ml. PH6,8]

10.-Incubar en hielo 15 min.

11.-Separar en alícuotas de 100 µl en tubos de 1.5 ml, inmediatamente congelar hielo seco.

12.-Almacenar a -80°C.

Las cepas competentes se utilizaron en todos los experimentos, las transformaciones se realizaron siguiendo el protocolo de choque térmico:

1.-Descongelar un dial en hielo (20min.-30min).

2.-Agregar de 1µl -9µl de DNA (plásmido) al dial y mezclar por inversión 3 - 4 veces.

3.-Incubar en hielo por 30min.

4.-Dar choque térmico a las células a 42°C por 1 min.

5.-Poner en agua con hielo por 2 min.

6.-Agregar 1 ml de medio SOC a temperatura ambiente.

7.-Incubar a 37°C por 1 hr. en agitación

8.-Plaquear 100µl en placas con LB-ampicilina.

5.5 Extracción de vectores.

Para la extracción de los vectores se utilizó el protocolo de QIAgen miniprep **(cita)**. El protocolo es diseñado para la purificación de hasta 20 µg de plásmido con alto numero de copias a partir de entre 1-5 ml de cultivo de *E. coli* en LB.

1.- De un medio de cultivo centrifugar 1ml a 6000rpm por 1min.

2.- Resuspender la pastilla de células en 250 µl de buffer P1 en un tubo de microcentrifuga.

3.- Adicionar 250 µl de P2 y mezclar por inversión de 3 a 4 veces, (hasta tener una mezcla viscosa).

4.- Adicionar 350 µl de buffer N3 y mezclar por inversión de 3-4 veces.

5.- Centrifugar por 10 min a 13,000rpm.

6.- Tomar el sobrenadante en una columna QIAprep por pipeteo, centrifugar por 1min.

7.- Adicionar 500 µl de buffer PB y centrifugar 1 min adicional.

8.- Lavar la columna con buffer PE y centrifugar 1min.

9.- Poner la columna en un tubo de centrifuga nuevo, en el centro de la columna adicionar 50 µl de agua miliQ o 50 µl de buffer EB dejar reposar 1min y centrifugar por 1min.

5.6 Digestiones con las enzimas NedI, BamHI, pcr y electroforesis de gel de agarosa.

El protocolo utilizado para las restricciones de BamHI y NdeI fue el recomendado en el kit de cada uno de las enzimas, se realizaron para cada gen una digestión sencilla como control lo que permitía comprobar que los genes se encontraban en el vector y también se realizaron dobles digestiones para extraerlos de los vectores correspondientes.

El primer vector digerido fue Puc57 donde se encontraban los genes insertados directamente de la secuenciación automática, se hizo una electroforesis de gel de agarosa al .7 % y fueron purificados los fragmentos por el kit gel extraction QIAgen descrito *a posteriori.*

5.7 Digestión enzimática.

1.-Se hizo una mezcla de acuerdo a las proporciones adecuadas:

Agua libre miliQ	16µl
Buffer-3 10X (BamHI) o (NdeI)	2µl
DNA (0.5-1 µg/µl)	1µl
Enzima (BamHI) (NdeI)	0.5-2 µl

2.- Mezclar dando un *spin* por 5 segundos.

3.- Incubar a 37°C de 3 a 4 horas.

4.- Detener la reacción con un choque térmico a 98°C por 1 min.

Para el caso de pet-19b se hicieron las mismas restricciones teniendo el vector listo para ser ligado con los fragmentos, a las cepas transformadas se les realizo pcr's de colonia para comprobar que las transformaciones estaban hechas adecuadamente y contenían los genes *cosA y htgA*.

5.8 Reacción en cadena de la polimerasa (pcr) de colonias.

Los pcr's de colonia se realizaron mezclando el toque de una colonia con una punta de micropipeta de 10 µl con la mezcla de los componentes recomendados por el kit de *Taq polymerase start fusion*:

38 µl H_20 miliQ
5 µl 10X PCR buffer (500 mM KCl, 100 mM Tris-HCl (pH 9.0), 1.0% Triton X 100)
3 µl 25 mM $MgCl_2$
1 µl 10 mM dNTPs (10 mM each dATP, dTTP, dGTP. dCTP)
1 µl 20 µM forward primer T7
1 µl 20 µl reverse primer T7
0.2-1 µl *Taq polymerase*

Todos los pcr's fueron programados con una serie de ciclos que dependían del tamaño y las recomendaciones de los plásmidos utilizados.

Los protocolos de purificación de DNA en gel de agarosa fueron hechos a partir de los recomendados por el kit de QIAgen:

1.- Cortar con una navaja el fragmento de gel de agarosa al .7 % intentando tomar solo lo necesario de la muestra.
2.- Pesar la banda de gel, adicionar 3 volúmenes de buffer QG por 1 volumen del gel.
3.- Incubar a 50°C por 10 min (hasta que se disuelva el gel) eventualmente mezclar por inversión 2 -3 veces cada 2 minutos.
4.- Checar que la mezcla sea amarilla sin trazas.
5.- Adicionar 1volumen de isopropanol frío y mezclar por inversión.

6.- Pasar a una columna de purificación el contenido de la mezcla.

7.- Centrifugar por 1min.

8.- Repetir el proceso hasta no tener mezcla en el tubo de micro-centrifuga.

9.- Adicionar .5ml de buffer QG y centrifugar por 1 min.

10.- Lavar la columna con .75ml de buffer PE y centrifugar por 1min.

11.- Centrifugar por 1min a 13000 rpm.

12.- en un tubo de 1.5ml poner la columna y adicionar 50µl de buffer EB incubar por 1 minuto y centrifugar por 2 minutos.

5.9 Ligación de los genes y transformación de *E. coli dh5α y E. coli BL21* con el vector pet19b pcr.

La ligación de pet19b con los genes se realizó con el protocolo estándar:

1.- Se mezcló en un tubo de micro-centrifuga en hielo para una reacción final de 20µl.

T4 DNA ligasa	1µl
10x buffer DNA-ligasa	2µl.
Vector linearizado	50ng (0.025 pmol)
DNA (gen)	50ng (0.076 pmol)
Agua miliQ (hasta tener)	20µl

2.- Mezclar con pipetas o vortex incubar a 16°C toda la noche o a temperatura ambiente por 2 horas.

3.- mantener en hielo y transformar 5-10 µl para una reacción de 100µl de células competentes.

Fueron comprobadas cada una de las transformaciones por resistencias al antibiótico, PCR de colonia de los fragmentos y por dobles digestiones.

5.10 Análisis de características físico-químicas de *cosA y htgA.*

Se realizaron análisis físico-químicos de los posibles productos directamente de la pagina http://expasy.org/tools/ una base de datos que tiene una serie de herramientas que permiten predecir

comportamientos físico-químicos a partir de la archivos de secuencias, se realizaron para todos los genes dichos análisis buscando perfiles hidropáticos, probables plegamientos, la cantidad de citocinas contenidas en las secuencias debido a que esto permite plegamientos estables, lo que nos permitiría decidir *a posteriori* si sería conveniente la cristalización de los genes traslapados en los trabajos reportados de importancia para este proyecto.

Los resultados de cada uno de los genes son reportados en el **anexo 7A** físico-químicas propiedades. En el caso de *cosA y htgA* en sus dos versiones se buscaron los puntos isoelectricos debido a que a partir de ellos podríamos predecir el tipo de buffers y columnas para su purificación.

5.11 Estandarización de métodos de inducción y purificación de las proteínas por columnas de níquel o cobalto con buffers de fosfatos y de sodio, electroforesis de sds-page desnaturalizante.

Se probaron distintos protocolos de inducción de la proteína, los cuales variaban en tiempo, 3 concentraciones de IPTG y 3 temperatura de inducción:

La inducción fue escalonada para disminuir el tiempo de crecimiento hasta las densidades ópticas adecuadas (.5 D.O - .6 D.O) se escalonaron 3 veces de 1:10ml, 10:100ml, 100:1000ml, todas con el respectivo antibiótico en proporción 1000x.

1.- Inocular 10 ml de medio LB ampicilina toda la noche con la cepas de *E. coli BL21* transformada son los genes *cosA y htgA* en sus dos versiones se creció toda la noche a 37°C.
2.- Inocular 90 ml de medio con los 10 ml del paso anterior durante toda la noche 37°C.
3.- Inocular 900ml de medio del crecimiento anterior toda la noche por 5 a 6 horas o D.O. de .5 a 6.
4.- Mantener en hielo por al menos 30 min. tomar una muestra como control negativo.
- En este paso se separaron alícuotas para probar los diferentes volúmenes final de IPTG.
5.- Adicionar IPTG a un volumen final (.5mM; 1mM; 1.5mM o 2mM).
6.- Incubar los diferentes tubos de medio con el inductor a diferentes temperaturas según el caso
-Se utilizaron 3 temperaturas 20°C, 30°C y 37°C.
7.- Se tomaron muestras a las 2 hr., 4hr. y 20hr.
8.- Se realizó una electroforesis de acrilamida SDS-PAGE con los productos.

5.12 Electroforesis de poliacrilamida SDS-PAGE.

Los productos de la inducción se corrieron en un gel de acrilamida SDS-page desnaturalizantes, con ello pudimos ver la cinética de inducción de las 3 construcciones, cada muestra cargada constaba de 15µl de una dilución 1:10 del total de la inducción.

El Gel se preparo a 15% del gel separador

Gel separador (15%):

H_2O	- 7.2 ml.
Acrilamida/Bis	- 15 ml.
Tris-HCl pH 8.8	- 7.5 ml. a 1.5 M
SDS 20%	- .15 ml.
TEMED	- .02 ml.
PSA (100 mg/1 ml)	- .15 ml.

Gel concentrador :

H_2O	- 3.075 ml.
Acrilamida/Bis	- .67 ml.
Tris-HCl pH 6.8	- 1.25 ml. a .5 M
SDS 20%	- .025 ml.
TEMED	- .005 ml.
PSA (100 mg/1 ml)	- .025 ml.

Se utilizó el marcador molecular de invitro gen o de Bio-Rad en cada una de las corridas de los productos, las muestras fueron de 8µl de una dilución 1:10 del producto de la centrifugación.

5.13 Purificación de proteínas.

Con los métodos de inducción estandarizados se realizaron pruebas de purificación de la productos de los genes *htgA-l, htgA-s y cosA*, se siguieron los protocolos recomendados de las columnas de purificación QIAgen (Níquel y Cobalto).

1.- Teniendo un pellet de celulas inducidas con los protocolos anteriormente mencionados. Resuspender la pastilla en buffer de lisis (2-5ml x gr) durante 15 minutos en hielo.

2.- Adicionar 1mg/ml de lisozima he incubar por 30 minutos.

3.- Sonicar en hielo, 6 pulsos de 10 ciclos de (10 seg. x 10 seg.) de 200W – 300W.

4.- Si el lisado es muy viscoso, adicionar RNAsa (10µg/ml) y DNAsa (5µg/ml) e incubar en hielo por 15 min.

5.- Centrifugar el lisado a 10,000 x *g*.durante 30 min. a 4^0C, guardar la pastilla (P) y tomar el sobrenadante (S).

6.- Adicionar 5 µl SDS-buffer a 5 µl sobrenadante y guardar a -20^0C.

7.- Mezclar 4 ml del sobrenadante con 1 ml de buffer de Níquel o Cobalto, mezclar por a 200rpm a 4^0C por 60 min.

8.- Cargar la mezcla en una columna de Níquel o cobalto y rescatar la muestra que atravesó la columna (FW).

9.- Lavar la con buffer de Lavado de 10mM, 20mM y 40mM de imidazol y colectar cada fracción.

10.- Agregar el buffer de elución (50mM) a la columna y analizar en un gel desnaturalizante SDS-page. (protocolo anteriormente mencionado)

6. Discusión y Resultados Experimentales.

El trabajo experimental realizado con las cepas de *E. coli* y los vectores producto de síntesis automática dieron buen resultado, todos fueron comprobados con controles para cada uno de los pasos desde la extracción de los vectores hasta las purificaciones de las proteínas.

El reporte que se agregó como anexo 6A menciona que la optimización se llevo a cabo para corregir algunos parámetros como lo son:

- El uso de codón parcial
- Contenido de GC
- Contenido de dinucleótidos
- Estructuras de mRNA

- Sitios prematuros de Poli-A
- Estabilidad de mRNA
- Secuencias repetidas y sitios de restricción que interfieran con al clonación.

Como se puede observar en el anexo 6A las modificaciones del uso de codón parcial aumenta el índice de adaptación por codón (CAI), lo que permite que se tenga un óptimo nivel de expresión del gen modificando el uso de codón de la secuencia delimitando las regiones en las cuales las enzimas de restricción actúan. Los cambios en la secuencia se emplean para reducir el numero de codones en tándem que suelen ser raros o poco usados por los organismos modelos. Si tomamos en cuenta que se utilizó a *E. coli* para el trabajo se adecuó el uso de codón de los genes que provienen de *P. fluorescens* aumentando el índice de CAI de .7 a .89 para el gen *cosA,* de .65 a .87 para *htgA-l* y para *htgA-s*, el contenido de GC también se modificó prolongando la vida media del mRNA y a su vez evitaba que aparecieran asas en los productos de RNA que afectaran los niveles de expresión de los genes, los tres fragmentos que se mandaron a sintetizar presentaban estas modificaciones las cuales se encuentran reportadas en el anexo 6A.

Con las modificaciones hechas en los genes aseguramos que la expresión sea óptima en nuestras cepas, las transformaciones de las cepas con los vectores puc57 tuvieron buenos resultados, en las placas de LB-ampicilina (100µg:1µl) se creció en 100µl y 150µl de cultivo y hay una gran cantidad de clonas para descartar que fuera contaminación resistente se tomaron al azar clonas a las cuales se les realizaron pruebas de pcr de colonias y digestiones con cada una de ellas, en la figura 12 se muestra la serie de pasos para comprobar que las transformaciones eran adecuadas resultado las digestiones, se puede ver que hay aparición de las bandas que corresponden a los tamaño esperado de los productos de 1026 pb. para *cosA,* 492pb. para *htgA-s* y de 597 pb. para *htgA-l.*

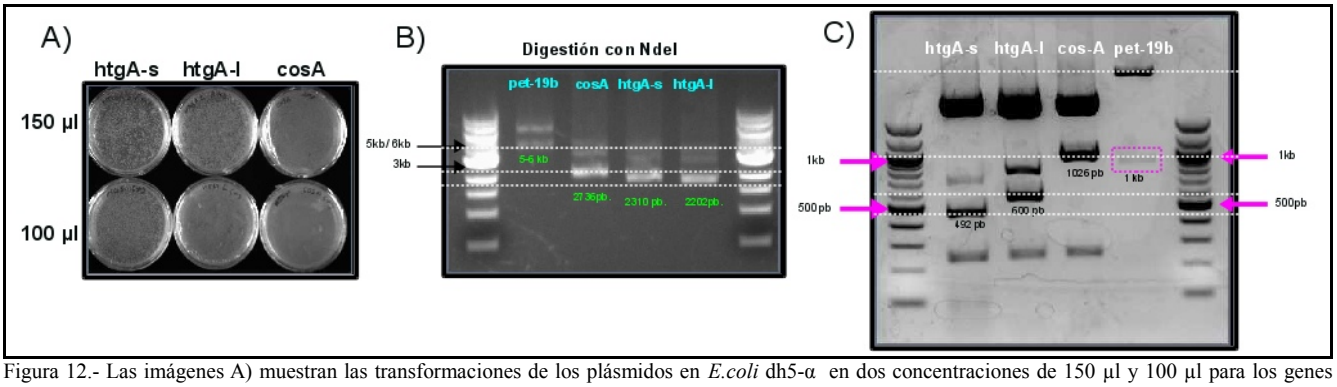

Figura 12.- Las imágenes A) muestran las transformaciones de los plásmidos en *E.coli* dh5-α en dos concentraciones de 150 µl y 100 µl para los genes htgA-s, htgA-l y cosA en puc-57, B) digestiones con NdeI para linearizar los plásmidos que resultan con los pesos esperados pet-19b con un peso entre 5 kb y 6 kb, cosA con un peso aproximado de 2736 pb, htgA-s de 2310 pb y htgA-l de 2202 pb, en un gel de agarosa al .7%, con un marcador 1kb de *New-england-BIOlabs* como controles la imagen solo muestra la digestión con una enzima NdeI. C) Doble digestión con NdeI y BamHI para extraer los genes con los pesos aproximados: htgA-s de 492 pb, htgA-l 597pb y cosA 1026 pb, el vector pet-19b al estar digerido con ambas enzimas se libera un fragmento intermedio de 1kb marcado en un recuadro punteado.

Se transformó la cepa *E.coli dh5-α* con el vector puc57 portador de los genes sintéticos (*cosA/htgA-l/htgA-s*), una vez transformada se amplificaron los plásmido creciendo las cepas en LB-ampicilina como se puede observar en la figura 12A, el resultado de la transformación fue un crecimiento óptimo de las cepas, a cada una se extrajo el plásmido el cual fue digerido con las enzimas de restricción BamHI y NdeI (figura.-12B) con sus respectivos controles, se corrieron electroforesis en gel de agarosa como se puede ver en la imagen, cada uno de los fragmentos presenta los pesos esperados para cada fragmento, se realizaron las pruebas con una sola enzima y con ambas, se digirió al mismo tiempo el vector pet-19b en donde cada uno de los genes serían insertados (figura.-12C).

Con el producto de las digestiones se ligaron en el plásmido pet-19b. Se transformaron de nuevo en *E.coli dh5-α* de las cuales se obtuvieron un total de 8 colonias (imagen no mostrada), 1 para *cosA,* 1 para *htga-l* y 6 para *htgA-s* todas fueron amplificadas en medio LB-ampicilina, se les extrajo DNA plasmidico (figura.- 13A) y se realizaron PCR`s de colonia con los primer`s T7 *(forward-reverse)* para corroborar que tenían el vector insertado, se eligieron 4 colonias que representaban las mejores amplificaciones obteniendo de *cosA* una*, htgA-l* una *y htga-s* dos. Con el producto restante se transformó *E.coli BL21,* para comprobar que los genes se insertaron adecuadamente se tomaron muestras al azar de las que se hicieron pcr's de colonias, de DNA extraído sin kit y con kit QiaGen. (figura.-13)

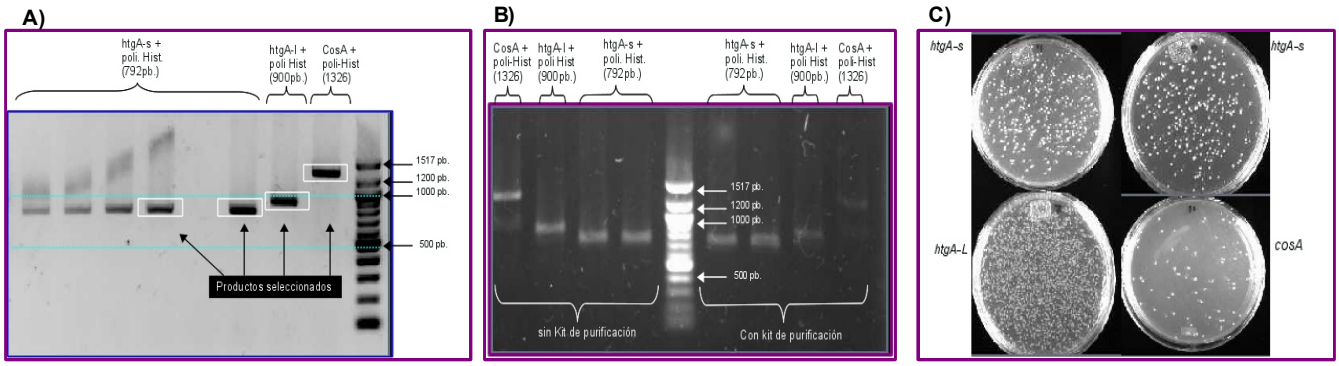

Figura 13.- Secuencia de pasos para la extracción, transformación y amplificación de los plásmidos con los 3 genes. A) se muestra pcr de 8 colonias obtenidas con las ligaciones de pet-19b + gen, se eligieron las 4 mejores amplificaciones de los genes. se utilizo dos técnicas de purificación y extracción de plásmido en B) se muestra una amplificación el plásmido pet-19b + los genes con los primers T7, con un kit de purificación y sin el kit. C) Las colonias se crecieron con LB-ampicilina (100µg:1µl), muestran las 4 cepas amplificadas.

Se realizaron diferentes búsquedas de características físico-químicas, para *cosA y htgA* en sus dos versiones para obtener alguna clave que nos permitiera tener un punto de partida a la hora de purificación de los productos, la búsqueda de la composición de aminoácidos se hizo directamente de la pagina (http://web.expasy.org/compute_pi/) donde se encuentra la paquetería necesaria para determinar el punto isoeléctrico y el peso molecular que nos permiten elegir el tipo de buffers utilizaremos para la purificación de la proteína los resultados se encuentran en el anexo 7A.

Se encontraron los tamaños relativos, pesos moleculares y punto isoeléctrico para cada una de las proteínas, el numero de residuos de una cadena define el tamaño y deduce el peso molecular de los productos, el gen *cosA* que tiene 338 a.a. con un peso molecular aproximado de 38.655 kDa, pI de 10.11, *htgA-l* tiene 195 a.a. con un peso de 21.09 kDa, pI de 9.44, *htga-s* tiene 160 a.a., un peso aproximado de 17.5 kDa y pI de 9.15.

La inducción de las cepas se realizó con el protocolo antes mencionado, se probaron 4 concentraciones de IPTG a .5, 1, 1.5, 2 mM, 3 temperaturas y 3 tiempos. En las imágenes de las figuras 14-16 se pueden observar los geles de SDS-PAGE para cosa, htga-l y htga-s.

6.1 *htgA-s*

Para el registro de la inducción del gen htgA-s, se tomaron muestras de 3 tiempos y 4 concentraciones de IPTG como se menciono anteriormente (figura.-14), los tiempos 2 no se muestran, para ningún caso se encontró registro de inducción a este tiempo.

A la temperatura de 20^0C - tiempo 1 (2hr.) no se observa el efecto del IPTG en ningún caso en el control del tiempo 0 se puede observar el patrón de bandeo con una mínima expresión, no encontramos evidencia de algún producto.

En el tiempo 3 tampoco se puede apreciar el producto esperado, la intensidad para todas las banda son difusas, la concentración de IPTG no aumenta la producción de la proteína.

La inducción a 30°C en el tiempo 1 no revela inducción, las bandas que corresponderían al peso esperado se ven muy tenues, el control no inducido presenta un patrón de bandeo más intenso debido a la cantidad de muestra cargada en este caso se utilizó 20µl con una dilución 1:5 figura 14-B.

El tiempo 3 tampoco muestra algún efecto de inducción, las concentraciones de IPTG no promueven la producción de la proteína esperada.

No se encontraron en los pesos esperados un aumento en la inducción para ninguna de las 4 concentraciones el aumento de la intensidad del control no inducido es debido a la cantidad de muestra cargada y la dilución en que se encuentra aun con ese aumento en la muestra cargada no se observa una banda que sea más intensa con el peso esperado.

El último registro para la inducción de *htgA-s* fue a 37°C, en todos los casos tanto de concentraciones de IPTG como a lo largo del tiempo, no se encontró registro de inducción, los patrones de bandeo fueron tenues en cada uno de los casos, las inducción no dio resultado alguno para este gen.

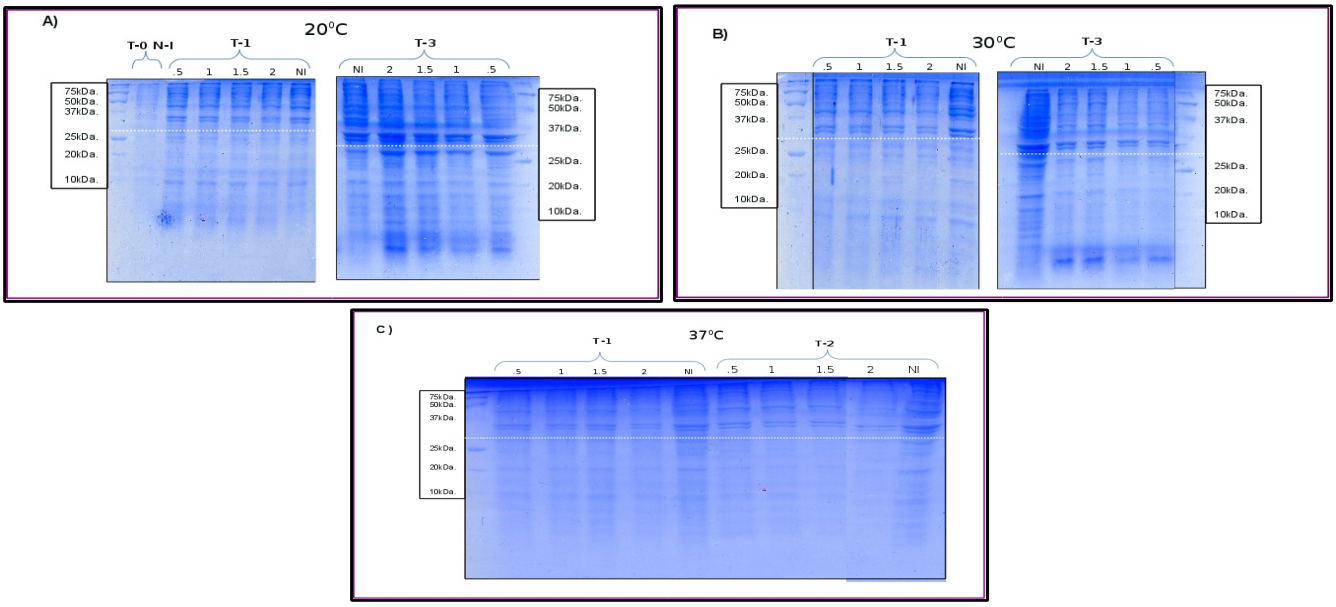

Figura 14 .- Muestra los geles de SDS-page para *htgA-s,* en A) se puede observar las 4 concentraciones de IPTG a una temperatura de 20^{0}C, a tiempo 0 (N-I), 1 y 3 con sus respectivos controles, la linea punteada se;ala en donde se localizaría el producto esperado de 17.64 kDa. + His-tail 8kDa.=25/28 kDa. B) muestra el siguiente paso de la inducción a 30^{0}C, y C) muestra la misma inducción a 37^{0}C.

6.2 *htgA-l.*

En el caso de *htga-l* los experimentos no revelaron que hubiese inducción en ningún caso como se muestra en la figura.- 15, los productos esperados fueron de 21.09 kDa, en los diferentes tiempos y con las distintas concentraciones de IPTG el patrón de bandas fue el mismo en todos los casos, no hay producto del gen *htgA-l*.

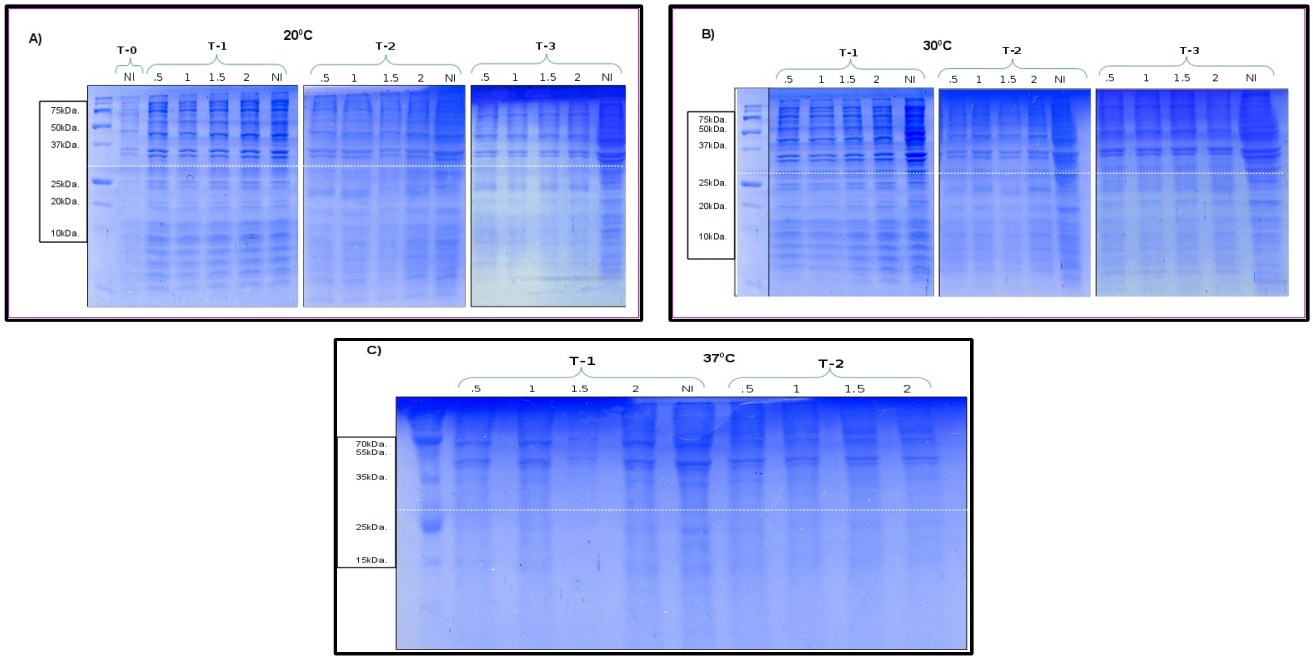

Figura 15 .- Muestra los geles de SDS-page para *htgA-l,* en A) se puede observar las 4 concentraciones de IPTG a una temperatura de 20^{0}C, a tiempo 0 (N-I), 1 y 3 con sus respectivos controles, la linea punteada se;ala en donde se localizaría el producto esperado de HtgA-l 21.23 kDa. + His-tail 8kDa. =29/30 kDa. B) muestra el siguiente paso de la inducción a 30^{0}C, y C) muestra la misma inducción a 37^{0}C.

6.3 *cosA.*

El gen *cosA* fue aquel en donde se encontraron buenos resultados con el método experimental realizado, se obtuvo que en todas las temperaturas y concentraciones de IPTG hubo inducción, con respecto a los controles no inducidos.

La inducción a 20°C muestra una banda que corresponde al producto del gen que tiene un peso aproximado de 38.6 kDa., el patrón de bandas de los productos se compararon con el control tiempo 0 no-inducido.

En el tiempo 1 es evidente la inducción de los productos al compararse con el control que no presenta una banda con la intensidad que presentan las diferentes concentraciones en el peso esperado para la misma, entre las concentraciones se puede ver que no hay diferencias entre las cantidades del inductor de cada muestra, en otras palabras no es determinante para la inducción la cantidad utilizada de IPTG, a partir de .05mM.

Al tiempo 2 la inducción se vuelva más evidente las bandas que corresponden al producto de gen tiene una intensidad mayor que el control no inducido, con respecto al tiempo de inducción no se ve una gran diferencia entre los tiempos 1, 2 y 3.

El fenómeno de degradación de las proteínas después de un largo periodo de tiempo es común en los métodos de purificación la inducción después de 21hr. mantiene el producto estable. La concentración de IPTG mínima es de .5, el tiempo tampoco altera la sobre producción ni la estabilidad de la proteína.

Las pruebas a 30ºC en el tiempo 1 no mostraron un aumento en la inducción con respecto a lso diferentes tiempos pero si con el control no inducido tiempo 0, la muestra del control fue de 20µl por lo que se ve más intenso el patrón de bandas lo cual no permite concluir que esté presente el producto esperado. El tiempo 2 mostró buenos resultados, se puede apreciar el aumento del producto esperado en comparación al control en la imagen se puede ver que la cantidad de IPTG no altera la producción de la proteína, en ningún caso ni se ve alterada la expresión a través del tiempo. En el tiempo 3 después de 21 hr. se aprecia una mayor cantidad de producto, se puede apreciar que a 1.5 y 2 mM se obtienen una mayor cantidad, lo que nos permite proponer que el T-3 sería el máximo para la sobre expresión del producto.

Las pruebas a 37ºC en los 3 tiempos se puede observar una disminución de producto con respecto a las otras temperaturas, no hubo un aumento en los dos primeros tiempos, al tiempo 3 se aprecia una banda que correspondería al peso esperado por lo que suponemos que la inducción es mínima en esta temperatura. No hay diferencias entre las distintas cantidades de IPTG que mantiene el mismo patrón anteriormente mencionado en cada uno de los tiempos analizados. El control en el tiempo 0 no inducido, nos permite ver que hay una baja inducción, debido a la temperatura, por lo que descartamos la posibilidad de estos parámetros para la sobre expresión del gen *cosA*.

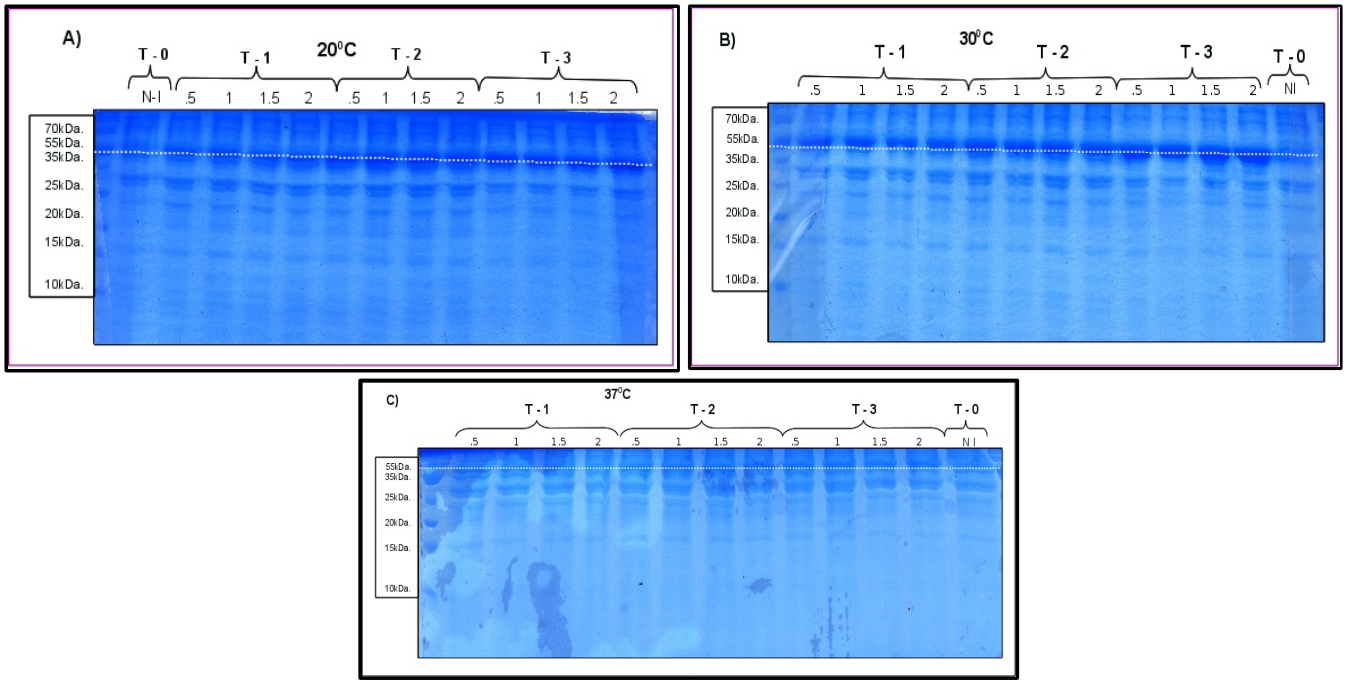

Figura 16 .- Muestra los geles de SDS-page para cosA, en A) se puede observar las 4 concentraciones de IPTG a una temperatura de 20^{0}C, a tiempo 0 (N-I), 1 y 3 con sus respectivos controles, la linea punteada señala en donde se localizaría el producto esperado de cosA 38.8 kDa. + His-tail 8 kDa. = 45/49 kDa. B) muestra el siguiente paso de la inducción a 30^{0}C, y C) muestra la misma inducción a 37^{0}C.

La inducción de los tres genes nos permitió elegir tiempos y temperaturas para las pruebas de purificación, basándonos en los resultados podemos decir que para *htgA-s* no hubo inducción con ninguna concentración, se podrían variar ciertos parámetros, como es la temperatura hacer un gradiente, modificar el medio que podría estar afectando el crecimiento, las concentraciones de IPTG fueron las estándares y con los resultados del gen *cosA* descartamos alguna modificación de concentración del inductor.

El gen *htgA-l* no hubo registro de inducción con ninguna concentración por lo que descartamos seguir algún procedimiento con este gen.

El gen *cosA* es el mejor prospecto para seguir analizando, presenta una muy buena inducción con diferentes condiciones tenemos inducción a 20ºC y 30ºC, se puede observar que las bandas entre estas dos temperaturas a pesar de ser muy buenas, hay una ligera diferencia en 30ºC las bandas se ven mas intensas. Las concentraciones de IPTG pueden ser desde lo mínimo .5mM a 2mM, lo que es un muy buen prospecto para una sobre expresión de la proteína, las muestras cargadas del producto

inducido en el gel fue una dilución de 1:10, la intensidad mostrada en los geles es adecuada tomando esto en cuenta nos proponemos para obtener una adecuada sobre expresión del producto se recomienda tomar las temperaturas de 30ºC en un lapso de tiempo de entre 4 a 8 horas en trabajos posteriores.

6.4 Pruebas de purificación de proteínas.

Las pruebas de purificación nos permitirían tener los parámetros necesarios para obtener la proteína en condiciones adecuadas para su cristalización.

Las pruebas de purificación estuvieron enfocadas en utilizar diferentes tipos de buffers, la primera prueba (figura.-17 A) se realizó en columnas de níquel con *buffers* de Tris-Hcl a pH 7 durante 6 hrs. a 30ºC de inducción con una concentración de .5mM de IPTG, para el registro del método se cargo 15µl en un gel desnaturalizante SDS-PAGE, como se puede observar en el gel, la inducción fue adecuada con respecto a la no inducida, se encuentra que la purificación no se realizó adecuadamente, en los distintos paso no se encontró un producto desde el primer paso, se pude observar una banda muy tenue en carril que correspondería a la inducción (Ind), no se encontró productos en los siguientes pasos, el producto de los diferentes pasos fueron diluciones de 1:10 volúmenes por lo que en los diferentes carriles se puede ver que las bandas son tenues a comparación de las hechas en las pruebas de inducción. (figura.- 17)

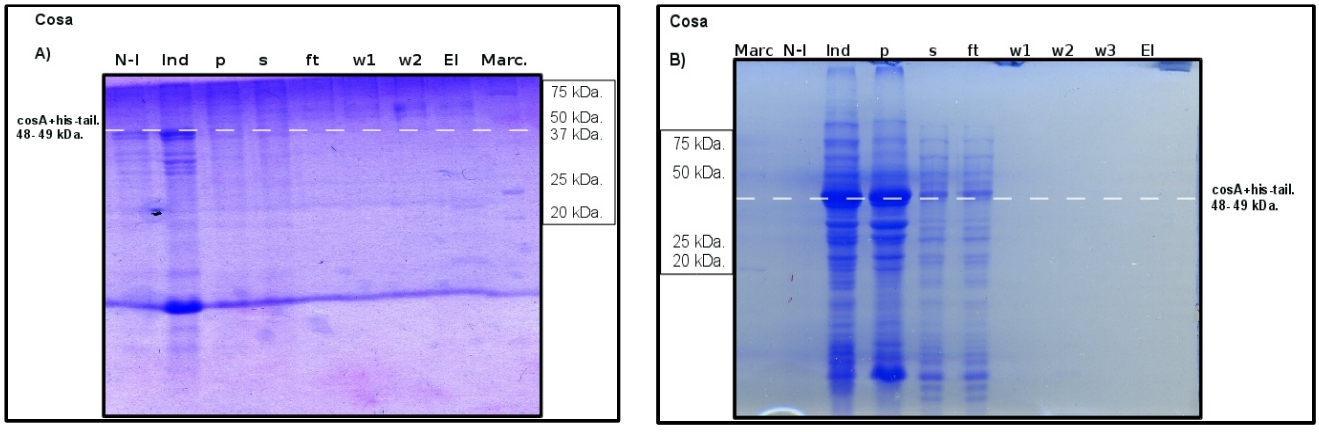

Figura 17.- Muestra dos geles con los resultados de la prueba de inducción en A) se utilizo buffer tris-HCl y columna de níquel, la linea punteada señala el peso para el producto esperado de entre 48 y 49kDa. para *cosA,* los carriles de izquierda a derecha están divididos en No-inducción, Inducido, Pastilla, Sobrenadante (S), Flow-through(ft), lavado 1 (W1), lavado 2 (W2), elución (E) y marcador, la muestra cargada presenta una dilución de 1:10. En B) se muestra la segunda prueba de purificación en columna de cobalto, con buffers de fosfatos para *cosA,* en la linea punteada se puede observar el producto en los carriles Ind, p, s y ft, las muestras se cargaron con una dilución de 1:10.

En la segunda prueba de purificación se modificó el *buffer* de Tris-HCl a uno de fosfatos con pH 8 y se cambio a una columna de cobalto durante 6 hrs. a 30ºC, se realizaron diluciones 1:10 para cargar las muestras que nos permiten observar que el producto se queda en la pastilla donde se encuentran los fragmentos insolubles, se puede observar una tenue banda en el sobrenadante, y en el ft, lo que nos muestra que la proteína no se esta pegando adecuadamente a la columna debido a su alto punto isoelectrico, por el momento podemos mencionar que el protocolo para la columna de cobalto nos ha dado un mejor funcionamiento.

Se realizaron 3 pruebas mas (datos no mostrados) con los mismos resultados, el producto de las pruebas se quedaba en las primeras fracciones (p, s y ft) y se utilizaron rangos de entre pH de entre 7.5 y 8.5, se mantuvo el tipo de buffers de Fosfatos.

Los resultados fueron similares todo el producto se queda en la parte insoluble de la muestra, no hay una unión del producto con la columna, se intentó implementar diferentes protocolos como la purificación de la proteína en su forma desnaturalizada, el cual podría darnos mejores resultados aumentando la afinidad del producto unido a la columna pero debido a problemas técnicos y del tiempo requerido para la parte experimental se detuvieron los experimentos.

7. Resultados generales.

En este trabajo se analizaron los genes cosA, htgA (en sus dos versiones) y el grupo Nov's de *Pseudomonas fluorescens*, los cuales presentan características que ponen en evidencia los efectos sobre sus portadores entre ellas se encuentran longitud, cohesión, localización y tipo de traslape.

Se han documentado los fenómenos que permiten la aparición y la permanencia de los nuevos genes en muy pocos grupos de organismos, en el analizado encontramos varios de estos factores, podemos comprobar el costo informacional de la posición del traslape y suponer la facilidad de que el gen se fije, lo cual se puede traducir en restricciones evolutivas impuestas por selección natural, lo que afecta directamente al portador y al traslape.

Encontramos en los métodos bioinformaticos que hay una correlación entre la longitud de la secuencia traslapada y la cohesión del grupo, lo que nos indica que es difícil rastrear evidencia de selección negativa en secuencias cortas en un grupo muy conservado como lo fue en los diferentes casos en donde el traslape era menor al 27% con respecto a las porciones no traslapadas.

Como se ha mencionado anteriormente la localización del traslape es otro parámetro importante como lo fue en el caso de *nov' 11, 14, 15 y cosA,* debido a la importancia que existe en las regiones terminales e iniciales de un gen; en las regiones iniciales de un gen por lo general se encuentra las región reguladora que en la mayoría de los casos es altamente conservada y por lo tanto las restricciones que ahí radican son muy altas, en menor medida pero sin dejar de ser importante, las regiones terminales de un gen restringen la aparición de una secuencia codificante traslapada y a su vez aumenta su posibilidad de permanencia en el genoma por estar bajo dos presiones selectivas: a) la propia por ser una secuencia activa y b) la de la región donde se localiza, en este trabajo no se encontró selección negativa en las regiones traslapadas de los genes antes mencionados, lo cual nos podría señalar que no hay evidencia de que estas restricciones sean claras y contundentes para predecir la permanencia de un traslape, tampoco se observó el efecto que ejerce el gen traslapado sobre el portador lo cual habla de la importancia en estas regiones las cuales minimizan los efectos por una doble restricción impuesta por algún traslape, debido principalmente por el tamaño del traslape, y su localización respecto a su portador.

Los tipos de fases en que se encuentran los traslapes se han analizado en distintos trabajos (Fukuda et al., 2003; Kaessmann, 2010; D C Krakauer, 2000) y como se ha mencionado en aquí las fases son determinantes para la permanencia de los traslapes y las causas se ven reflejadas en las secuencias portadoras, se ve una mayor presión selectiva en las regiones en donde se encuentran secuencias traslapadas en algunos casos y a su vez, dentro de la filogenia como en los genes *htgA, nov 6 y nov 13*, en las regiones no traslapadas no se encontró evidencia de algún efecto ejercido en las filogenias, por la presencia del traslape con respecto al portador, con los resultados obtenidos en este análisis nos da una idea que nos permite suponer que los traslapes alteran la dinámica evolutiva de los genes portadores, aumentando la presión selectiva sobre las regiones codificantes, el tamaño de la muestra que se utilizó en este trabajo fue pequeña por lo que haría falta realizar un análisis mas complejo y con un numero mayor de genes traslapado. Lo que si podemos asegurar es el hecho de que el tamaño y la fase en donde se encuentran los genes traslapados determinan el grado de presión ejercida sobre si y sobre su portador, aunado al hecho de que las regiones que conforman un dominio funcional o estructural y presentan un traslape, aumentan la posibilidad de permanencia de este nuevo gen.

Como punto final, la función de los genes traslapados se han predicho en algunos casos, esto lo tomamos con renuencia debido al hecho de que los métodos con los que los han encontrado no han sido del todo claro, los productos encontrados en el trabajo de (Kim et al., 2009) están basados en análisis proteómicos donde localizan fragmentos de péptidos, los cuales se ubican en la secuencia de nucleótidos lo que hace una búsqueda un poco ambigua y a su vez se traduce en que posiblemente no tengan una función definida, por lo que la importancia del análisis a nivel de secuencia se vuelve vaga. Determinar la función a partir de un análisis bioinformático es sumamente complicado y por lo tanto no podríamos decir si existe una función definida, por las mismas fallas en el sistema como se demostró en este trabajo, así aunque tuviésemos evidencia de selección de cualquier tipo, podríamos solo intuir si por estar bajo presión negativa tuviera una relevancia funcional, pero no podríamos asegurarlo.

8. Conclusiones.

Por lo tanto, podemos concluir que la selección negativa que es evidente en *nov 6, nov 13 y htgA*, son resultado de un traslape y sus diferentes efectos que producen en los portadores que a su vez permiten la estabilidad, permanencia de las secuencias traslapadas y ¿porque no? del propio portador, como lo es el caso de los distintos virus los cuales ocupan la compactación genómica (fenomeno poco conocido) como adaptación a las altas tasas de mutación que presentan resultado de su dinámica génica, las cuales disminuirían por la presencia de un traslape.

Se ha encontrado en varios casos de virus que las estructuras de los productos de traslapes tienen conformaciones poco usuales y con características que suponen pudieran tener una estructura que sale del promedio de las encontradas en las bases de datos, varios trabajos mencionan que los virus toman como una característica adaptativa este fenómeno lo cual nos permite pensar pudiera ser parte de un mecanismo evolutivo que aumenta la posibilidad de aparición y permanencia de productos novedosos, aunque por otro lado las formas encontradas en las diferentes bases de datos muestran que entre mas estructuras se describen, menos formas se encuentran, lo que pudiera ser reflejo de que las restricciones físico-químicas, juegan un papel importante en la evolución de las formas, tomando en cuenta que los productos descritos en su mayoría provienen de genes con una historia evolutiva larga y hay pocos análisis sobres genes de aparición reciente o con una historia evolutiva relativamente corta, la idea de que las formas están reducidas a unos pocos conjuntos pierde validez y puede ser discutida,

ya que como se ha mencionado, los genes encontrados en las bases de datos son genes bien definidos y caracterizados, de entre ellos se encuentran pocos que no se les ha determinado función, lo que nos lleva a pensar si lo que observamos en los conjuntos de formas es resultado de que solo se tienen las formas de genes con historias evolutivas bien definidas. En otras palabras no hay un registro suficientemente amplio y distribuido en los tres dominios de estructuras *de novo* que nos permitan tomar una postura sobre la importancia de las restricciones históricas *v.s.* restricciones físico-químicas.

Sin los efectos de las restricciones físico-químicas la gran cantidad de formas que pudieran aparecer serían tantas y tan variadas que la selección natural se vería diezmada para actuar, por otro lado, si las restricciones históricas fuesen el factor principal la reducción de las formas existentes por selección natural llevarían a la extinción de las mismas y la permanencia de los productos por efecto físico-químicos tendrían una enorme barrera que no podrían atravesar, debido a que un efecto selectivo las eliminaría inmediatamente, en cualquiera de los dos casos las modificaciones que pudieran existir en los productos tendrían que ser parte de una ruta evolutiva que no podría ser recorrida sin la ayuda de la selección natural. Saber cual de ellas tiene un mayor peso en la evolución de las estructuras proteicas es una discusión con argumentos muy complejos y pocas formas de resolverlo, las tecnologías bioinformáticas no son suficientes para llegar a una conclusión sobre el tema, la descripción de las estructuras han tenido un aumento exponencial en las ultimas dos décadas, pero aun no se tiene el suficiente conocimiento sobre estas y en menor medida sobre los que son resultado de nuevos genes que están constantemente apareciendo. Pero lo que consideramos es la mejor idea sobre el tema es el hecho de que la evolución de estos, esta ligada a la interacción entre las restricciones físico-químicas y las evolutivas probablemente en igual medida, pero en diferente tiempo de acción y todo bajo el supuesto de que la selección natural influye sobres las interacciones de ambas, la cual permite la permanencia de los productos, volviendo así la discusión sobre la relevancia entre las dos tipos de restricciones, la relevancia temporal del efecto de cada una de ellas.

La parte experimental que ha quedado inconclusa en este trabajo presenta adelantos importantes para determinar la estructura de, al menos, unos productos que en un principio se propuso como objetivo de este trabajo, determinar las condiciones de purificación de los productos es un gran reto que no puede ser dejado de lado, las condiciones de temperatura y tiempo de inducción se encuentran bien definidas ahora, el trabajo a seguir es descubrir qué condiciones y qué tipo de columnas son ideales

para la purificación de los productos basados en sus propiedades, si observamos las secuencias de los productos podremos ver que tiene un alto punto isoeléctrico, lo que no permite que interactúe con las columnas que se probaron en el trabajo, por lo que se puede realizar un protocolo de purificación de la proteína de forma desnaturalizada para después ser reconstituida, el uso de otro tipo de columnas puede ser otra alternativa y el aumento del pH de los distintos buffer's utilizados, incrementaría así la afinidad del producto por la columna, disminuyendo los efectos del punto isoeléctrico de los productos sobre el sustrato, estos experimentos no se llevaron a cabo por falta de tiempo, pero sería recomendable terminar con la parte experimental permitiendo la continuidad del proyecto.

Como lo hemos estado comentado los genes traslapados forman parte de las grandes incógnitas de la biología evolutiva, molecular y estructural, lo cual abre las puertas de una gran cantidad de oportunidades para su estudio, la aparición de nuevos genes por mecanismos de sobreimpresión es un hecho que debería ser estudiado más a detalle por sus implicaciones tanto evolutivas como funcionales, los genes traslapados son una linea de investigación poco estudiada y con una relevancia enorme en las diferentes áreas de la biología que no hay que olvidar y mucho menos dejar a un lado.

"*Nuestro conocimiento de cómo se originan las partituras de esta perfecta sinfonía es mínimo. Todos los acordes, notas y tiempos serán revelados en su momento, por ahora, se ha disfrutado un molto, allegro, andante, minuento o un presto más, o al menos he escuchado cantar al destino.*"

Gracias.

III. Bibliografia.

- Boi, S., Solda, G., & Tenchini, M. L. (2004). Shedding Light on the Dark Side of the Genome: Overlapping Genes in Higher Eukaryotes. *Current Genomics*, *5*(6), 509–524.
- Brosius, J. (1991). Retroposons-Seeds of Evolution. *Perspective*, *4*(February), 1991.
- Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. *Gene*. 179:487-496.
- Carthew, R. W., & Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, *136*(4), 642–55.
- Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings. Biological sciences / The Royal Society*, *277*(1701), 3809–17.
- Cyrus Chothia. (1992). One thousand families for the molecular biologist. *Nature*, *vol357*.
- Delaye, L., Deluna, A., Lazcano, A., & Becerra, A. (2008). The origin of a novel gene through overprinting in Escherichia coli. *BMC evolutionary biology*, *8*, 31.
- Denton, M., & Marshall, C. (2001). Laws of form revisited Protein folds. *Nature*, *410*(March), 2001.
- Finnegan, D. J. (1990). Transposable elements and DNA transposition in eukaryotes. *Current opinion in cell biology*, *2*(3), 471–7.
- Fukuda, Y., Nakayama, Y., & Tomita, M. (2003). On dynamics of overlapping genes in bacterial genomes, *323*, 181–187.
- Helen Pearson. (2006). WHAT IS A GENE ? *Nature*, *441*(May).
- James., R., & Deyrick, D. (1991). Identification of a gene, closely linked to dnaK, which is required for high-temperature growyh of Escherichia coli. *journal of general microbiology*, (137), 1271–1277.
- Johnson, Z. I., & Chisholm, S. W. (2004). Properties of overlapping genes are conserved across microbial genomes. *Genome research*, *14* (11), 2268–72.
- Jones, C. D., & Begun, D. J. (2005). Parallel evolution of chimeric fusion genes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(32), 11373–8.
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome research*, *20*(10), 1313–26.
- Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature reviews. Genetics*, *10*(1), 19–31.
- Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics*, *9*(8), 605–18.
- Keese, P. K., & Gibbs, a. (1992). Origins of genes: “big bang” or continuous creation? *Proceedings of the National Academy of Sciences of the United States of America*, *89*(20), 9489–93.

- Kidwell, M G, & Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution; international journal of organic evolution*, *55*(1), 1–24.
- Kidwell, Margaret G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49–63.
- Kim, W., Silby, M. W., Purvine, S. O., Nicoll, J. S., Hixson, K. K., Monroe, M., Nicora, C. D., et al. (2009). Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens Pf0-1. *PloS one*, *4*(12),
- Kleckner, N. (1981). TRANSPOSABLE ELEMENTS IN PROKARYOTES Ann. Rev. Genet.15:341-404
- Koonin, E. V., Makarova, K. S., & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual review of microbiology*, *55*, 709–42.
- Krakauer, D C. (2000). Stability and evolution of overlapping genes. *Evolution; international journal of organic evolution*, *54*(3), 731–9.
- Krakauer, David C. (2002). Evolutionary Principles of Genomic Compression. *Comments on Theoretical Biology*, *7*(4), 215–236.
- Levitt, M. (2009). Nature of the protein universe. *PNAS*, *106*(27).
- Long, M., & Langley, C. H. (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science (New York, N.Y.)*, *260*(5104), 91–5.
- Makalowska, I., Lin, C.-F., & Makalowski, W. (2005). Overlapping genes in vertebrate genomes. *Computational biology and chemistry*, *29*(1), 1–12.
- Makałowska, I., Lin, C.-F., & Hernandez, K. (2007). Birth and death of gene overlaps in vertebrates. *BMC evolutionary biology*, *7*, 193.
- Mathias, S. L., Scott, a F., Kazazian, H. H., Boeke, J. D., & Gabriel, a. (1991). Reverse transcriptase encoded by a human transposable element. *Science (New York, N.Y.)*, *254*(5039), 1808–10.
- Missiakas, D., Georgopoulos, C., & Raina, S. (1993). The Escherichia coli Heat Shock Gene htpY : Mutational Analysis , Cloning , Sequencing , and Transcriptional Regulation, *175*(9), 2613–2624.
- Moran, J. V., DeBerardinis, R. J., & Kazazian, H. H. (1999). Exon shuffling by L1 retrotransposition. *Science (New York, N.Y.)*, *283*(5407), 1530–4.
- Ochman, H., Lawrence, J. G., & Groisman, E. a. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*(6784), 299–304.
- Pavesi, a, De Iaco, B., Granero, M. I., & Porati, a. (1997). On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. *Journal of molecular evolution*, *44*(6), 625–31.
- Pavesi, A. (2000). Detection of Signature Sequences in Overlapping Genes and Prediction of a Novel Overlapping Gene in Hepatitis G Virus. *Journal of Molecular Evolution*, 284–295.

- Rancurel, C., Khosravi, M., Dunker, a K., Romero, P. R., & Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *Journal of virology*, *83*(20), 10719–36.
- Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V., Wolf, Y. I., Jordan, I. K., Tatusov, R. L., & Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. *Trends in genetics : TIG*, *18*(5), 228–32.
- Sabath, N. (2009). Molecular Evolution of Overlapping Genes. *PHD thesis*, (December).
- Scherbakov, D. V., & Garber, M. B. (2000). Overlapping Genes in Bacterial and Phage Genomes. *Molecular Biology*, *34*(4).
- Silby, M. W., & Levy, S. B. (2008). Overlapping protein-encoding genes in Pseudomonas fluorescens Pf0-1. *PLoS genetics*, *4*(6).
- Zhang, J., Zhang, Y., & Rosenberg, H. F. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature genetics*, *30*(4), 411–5.
- Zhang, Y. wu, Liu, S., Zhang, X., Li, W.-B., Chen, Y., Huang, X., Sun, L., et al. (2009). A functional mouse retroposed gene Rps23r1 reduces Alzheimer's beta-amyloid levels and tau phosphorylation. *Neuron*, *64*(3), 328–40.
- Zhou, Q., & Wang, W. (2008). On the origin and evolution of new genes -a genomic and experimental perspective. *Journal of genetics and genomics*, *35*(11), 639–48.