UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

Facultad de Ciencias

IMPACTO DE LA HIBRIDACIÓN VIRTUAL EN LA DETERMINACIÓN DE SECUENCIAS DE DNA

TESIS

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS BIOLÓGICAS (BIOLOGÍA EXPERIMENTAL)

P R E S E N T A

FABIÁN REYES PRIETO

DIRECTOR DE TESIS: DR. ROGELIO MALDONADO RODRÍGUEZ

MÉXICO, D.F. NOVIEMBRE, 2007

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

AGRADECIMIENTOS

Al Posgrado en Ciencias Biológicas, por la oportunidad de estudiar una maestría de excelencia.

Al CONACYT, por la beca que me otorgó durante el período de febrero de 2005 a diciembre de 2006.

Al Dr. Rogelio Maldonado Rodríguez, por brindarme el espacio, el respaldo y la tutoría para emprender mi desarrollo profesional.

Al Dr. Alfonso Méndez Tenorio, pues gracias a su invaluable y atenta asesoría fue posible en gran medida realizar este proyecto.

A la Dra. María Alicia González Manjarrez, al Dr. Mauricio Salcedo Vargas y al Dr. Enrique Merino Pérez, por sus valiosos aportes y sugerencias para la consolidación de este trabajo.

A los amigos del Laboratorio de Biotecnología y Bioinformática Genómica, que hacen del lugar de trabajo un sitio inspirador y confortable.

A la UNAM, ese ente maravilloso en cuyos espacios he tenido la fortuna de avanzar, y por la que de una u otra forma siempre mantendré una devoción inalterable.

Al IPN, por los primeros años de comunión, por lo grandioso que será estar vinculados.

ÍNDICE

ÍNDICE DE FIGURAS

RESUMEN

Algunas técnicas de secuenciación de DNA como la secuenciación por hibridación (SBH, por las siglas del inglés ‘sequencing by hybridization’) son susceptibles de mejoras, para su posible “revigorización” y amplia difusión. En la presente investigación se atendió el perjuicio que representa para la SBH el manejo de conjuntos de sondas extremadamente numerosos, a través del uso de un conjunto de 511 sondas 8-mer con características especiales. Tales sondas fueron sometidas junto con blancos de secuencia conocida a una rutina de hibridación virtual sin restricciones termodinámicas. Los patrones de hibridación sirvieron para un análisis de cobertura y superposición, y posteriormente para la reconstrucción de las secuencias de los blancos empleando un algoritmo diseñado también durante este trabajo. Las reconstrucciones obtenidas fueron satisfactorias para el caso de un blanco modelo hipotético que hibridó con un conjunto de sondas modelo, aunque en el caso del resto de los blancos analizados los inconvenientes que no pudo solucionar el algoritmo fueron varios. Se proponen como perspectiva perfeccionamientos a nivel del conjunto de sondas y del algoritmo de reconstrucción.

ABSTRACT

Some DNA sequencing techniques, such as the sequencing by hybridization (SBH), are susceptible of practical improvements to reinforce their essential virtues and extend their use. A major practical inconvenient of the SBH is the high number of probes required in a single sequencing event. In this work, I analyzed the use of 511 specifically-designed 8-mer probes to address the efficiency and practical use of probes for SBH. A routine of virtual hybridization with no thermodynamic restrictions was carried out with the 511 8-mer probes against 68 known DNA sequence targets. The hybridization patterns were analyzed to quantify cover, superposition, and ultimately to reconstruct the sequence of the DNA targets using an algorithm designed in this work. The obtained sequence reconstructions were satisfactory only for the particular case when a hypothetical model was hybridized against a model probe set. However, the algorithm was not able to reconstruct the sequences for the remaining analyzed DNA targets. Several events inherent to the hybridization process were not deciphered by the algorithm. The results indicate that future investigations about probe set design and reconstruction algorithms are required to explore the potential applications of the SBH.

1. INTRODUCCIÓN

1.1. ESTRATEGIAS PARA SECUENCIAR DNA

Según Drmanac et al. (2002), existen dos estrategias posibles para secuenciar DNA (Fig. 1): métodos directos, en los cuales la posición de cada base en la cadena del DNA a secuenciar (DNA blanco) es determinada individualmente [e.g. secuenciación por división química (Maxam y Gilbert 1977), secuenciación por terminación controlada de la replicación (Sanger et al. 1977), y pirosecuenciación (Ronaghi et al. 1998)], y métodos indirectos, en los cuales la secuencia de DNA blanco es ensamblada a partir de la determinación experimental de su contenido oligonucleotídico [e.g. secuenciación por hibridación (SBH, por las siglas del inglés ‘sequencing by hybridization’; Drmanac y Crkvenjakov 1987)].

1.2. MÉTODOS DIRECTOS DE SECUENCIACIÓN DE DNA

La secuenciación por división química se efectúa al marcar con fósforo radiactivo (^{32}P) los extremos 5’ de fragmentos en los que es segregado el DNA blanco. Posteriormente estos fragmentos marcados son subdivididos en una posición secuencia-específica,

obteniéndose cuatro conjuntos diferentes [fragmentos subdivididos en adenina (A), guanina (G), timina (T), y citosina (C)]. Finalmente éstos son sometidos a una electroforesis en gel de poliacrilamida (los fragmentos más cortos se desplazan mayores distancias a través del gel), lo que permite separar fragmentos que se diferencian tan sólo en un nucleótido. La lectura de los patrones de bandas obtenidos en autoradiografías permite establecer la secuencia de los fragmentos y así reconstruir la secuencia total del DNA blanco.

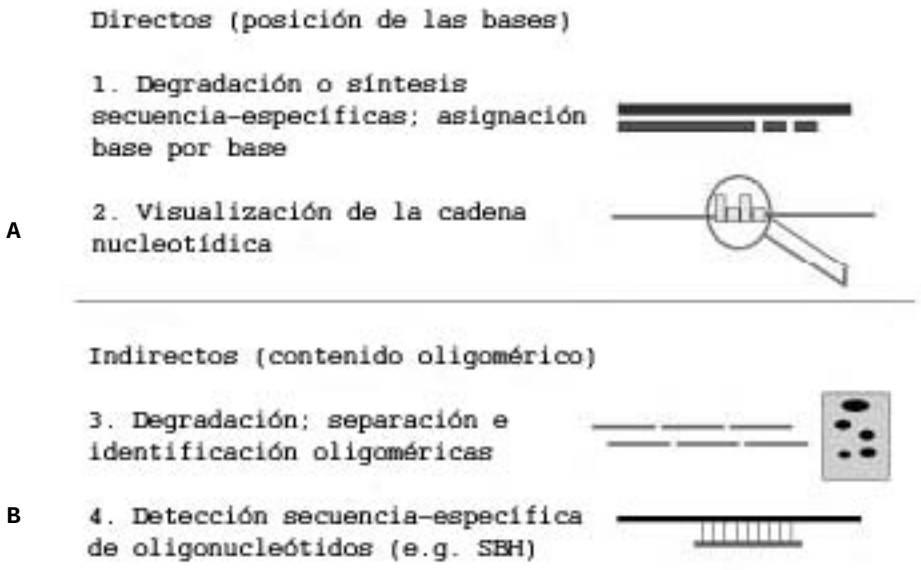

Fig. 1. Estrategias de secuenciación de DNA. Los métodos de secuenciación son directos o indirectos, dependiendo del tipo de datos experimentales obtenidos del DNA de interés (DNA blanco). **A** Los métodos directos implican la determinación del carácter o base que está presente en una posición particular en la cadena de DNA. Hay dos tipos principales de secuenciación directa. La primera implica la degradación, síntesis, o separación secuencia-específicas seguida por una asignación base por base; la segunda implica la visualización directa de la cadena. **B** Los métodos de secuenciación indirectos implican la determinación de las subcadenas de bases que están presentes o ausentes en el DNA blanco. El primero de dos métodos indirectos implica la fragmentación del

DNA blanco, seguida de la separación e identificación de los oligómeros resultantes. La SBH, que es el segundo método indirecto, implica el apareamiento secuencia-específico del DNA blanco con oligómeros (sondas) complementarios de secuencia conocida o determinable. (Modificada a partir de Drmanac et al. 2002).

La secuenciación por terminación controlada de la replicación es el proceso inverso de la secuenciación por división química. En el método original, cuando la DNA polimerasa está sintetizando una nueva cadena de DNA, se agregan en reacciones independientes los análogos 2’,3’-didesoxi de cada base. La síntesis se detiene cuando la DNA polimerasa incorpora los análogos, lo que ocasiona la obtención de fragmentos sintetizados de diferentes largos (como aquellos fragmentos obtenidos por división química). Los fragmentos se marcan radiactivamente para finalmente, y como ocurre en el método de secuenciación por división química, ser sometidos a una electroforesis en gel de poliacrilamida. A partir de la lectura de los patrones de bandeo obtenidos en autoradiografías se reconstruye la secuencia. Algunas mejoras fueron hechas en los 1980s con el uso de diferentes tintes fluorescentes para marcar los terminadores (Smith et al. 1986; Prober et al. 1987), de manera que todos pudieron ser incorporados en una sola reacción. Posteriormente el empleo de gel fue sustituido por capilares, lo que simplificó la etapa de separación e incrementó los largos de lectura. La expansión masiva en la secuenciación de DNA que se ha

dado en los últimos 10 años tiene como fundamento químico la técnica antes descrita, propuesta por Sanger hace 30 años. Tal progreso ha sido conducido por grandes centros de secuenciación que han implementado industrialmente dicho método, concretamente a través del método de secuenciación aleatoria conocido como ‘shotgun’ (Sanger et al 1980). Un efecto colateral de esa automatización es que la secuenciación a gran escala está alejada de los pequeños laboratorios de investigación, y la gran mayoría de los datos secuenciales son generados ahora por esos grandes complejos.

La pirosecuenciación es una técnica bioluminométrica de secuenciación de DNA que está basada en la secuenciación por síntesis. Su fundamento es la detección en tiempo real del pirofosfato inorgánico (PPi), el cual es liberado en la incorporación exitosa de nucleótidos durante la síntesis de DNA. El PPi es inmediatamente convertido en adenosín trifosfato (ATP) por la ATP sulfurilasa, y el nivel de ATP generado es detectado por los fotones producidos por la acción ATP-dependiente de la luciferasa. El ATP y los desoxinucleótidos son degradados por la apirasa, una enzima degradante de nucleótidos. La presencia o ausencia de PPi, y por lo tanto la incorporación o no incorporación de cada nucleótido agregado, es evaluada en última instancia por la detección o no de fotones. En 2005 Margulies et al. propusieron un método integral que emplea la

pirosecuenciación para la obtención de secuencias. El método, también conocido como secuenciación 454 [después de que 454 Life Sciences (Brandford, CT, EUA) lo ha comercializado], actualmente ha despertado un gran interés entre los grupos de trabajo que requieren métodos de secuenciación con reducciones en tiempo y costos.

Recientemente la compañía Illumina (San Diego, CA, EUA) ha desarrollado otro método de secuenciación paralela masiva por síntesis de fragmentos amplificados, dicha tecnología se llama Secuenciación Solexa (Bennett et al. 2005). En ésta, la amplificación del DNA es sobre una superficie sólida, y la secuenciación por síntesis se efectúa a través de la incorporación de nucleótidos modificados que están marcados con tintes coloreados.

Otros métodos directos de secuenciación dependen de la visualización de bases empleando microscopía de efecto túnel (‘scanning tunneling microscopy’; Beebe et al. 1989; Lindsay y Philipp 1991; Woolley et al. 2000) o nano-poros (Church et al. 1998; Yan y Xu 2006). Las dificultades técnicas principales asociadas con la visualización directa de las bases son la diferenciación incorrecta de las mismas y la baja velocidad de secuenciación.

1.3. SBH, UN MÉTODO INDIRECTO PARA SECUENCIAR DNA

Como se mencionó anteriormente la SBH es un método indirecto para secuenciar DNA, el cual se basa en determinar el contenido oligonucleotídico de un DNA blanco sin requerir la asignación experimental de la posición de cada base. El contenido oligonucleotídico es obtenido por la hibridación específica entre oligómeros (sondas) de secuencia conocida o determinable y el DNA blanco. La secuenciación por hibridación fue propuesta primero en 1987 por Drmanac y Crkvenjakov. Cada uno de los oligonucleótidos que hibrida indica la presencia de una o más secuencias complementarias en el DNA blanco, revelando una pequeña parte de información de su secuencia. Entonces la secuencia completa se determina compilando los resultados de muchas de estas pruebas de hibridación. El método de SBH tiene el potencial para secuenciar con precisión y bajo costo ácidos nucleicos muy largos (>10,000 bases), todo en una sola reacción (Drmanac et al. 2002).

1.3.1. HISTORIA DE LA SBH

Científicos de muchas disciplinas contribuyeron en la construcción del marco teórico que posibilitó el desarrollo de la SBH (Drmanac y Crkvenjakov 1987). En su celebérrimo artículo de

1953, Watson y Crick propusieron un modelo estructural para el DNA, el cual implicaba, entre otras consideraciones, apareamientos específicos de A con T y de G con C. Al paso del tiempo su modelo fue corroborado, y el concepto de complementariedad, referente a los apareamientos específicos, fue acuñado. Doty (1960) observó que cuando se calienta en solución, el DNA de doble cadena se “funde” (desnaturaliza) para dar origen a cadenas de hebra sencilla, las cuales por complementariedad renaturalizan espontáneamente cuando la solución es enfriada. Este resultado permitió considerar la posibilidad de reconocer un tramo de DNA usando otro. Posteriormente algunos investigadores descubrieron formas para sintetizar eficientemente grandes cantidades de pequeñas moléculas secuencia-específicas de DNA, esto a través de agregar paso a paso nucleótidos en columnas (Beaucage y Caruthers 1980). Wallace et al. (1979) emplearon oligómeros con la finalidad de confirmar la presencia de secuencias complementarias en blancos de DNA, y demostrar que eran capaces de detectar con precisión mutaciones de una sola base en moléculas de DNA definidas y esparcidas puntualmente sobre membranas. Poustka y Lehrach (1986) propusieron la obtención de la huella genómica (‘fingerprinting’) clonal de cósmidos usando un conjunto pequeño de sondas. En 1987 Mullis y Faloona desarrollaron la técnica de la reacción en cadena de la polimerasa (PCR). Los trabajos mencionados ofrecieron las

herramientas básicas de investigación que más tarde fueron fundamentales para el desarrollo de los métodos de secuenciación y huella genómica por hibridación.

La SBH fue primero propuesta y patentada por Drmanac y Crkvenjakov en 1987. Ellos propusieron que la secuencia de un fragmento de DNA podría determinarse por un experimento de hibridación en el cual el DNA blanco fuera expuesto a un conjunto de oligonucleótidos en condiciones que favorecieran un apareamiento completo. Entonces las secuencias del subconjunto de sondas que hibridaron positivamente podrían usarse para determinar la secuencia del blanco. Un formato propuesto por estos investigadores para secuenciar muestras complejas de DNA fue la hibridación en pozos que contenían sondas marcadas (también el uso de grandes arreglos de clonas). Ellos presentaron en 1989 un algoritmo para la reconstrucción de secuencias de DNA tan complejas como el genoma humano. Las investigaciones en el área de la SBH continuaron. Bains y Smith (1988) demostraron la reconstrucción de secuencias en un proceso de ensamble simulado que empleaba 256 sondas 6-mer diseñadas especialmente con dos posiciones degeneradas que se ubicaban en la región media de cada sonda (sondas espaciadas). En una aplicación de patente de 1988, Southern propuso un método combinatorio en vidrio para la síntesis *in situ* de arreglos oligonucleotídicos complejos. Él propuso que muestras de DNA hibridarían con arreglos de sondas,

permitiendo la secuenciación total de aquellas, así como la detección de mutaciones. Lysov et al. (1988) también propusieron la secuenciación de DNA por medio de hibridación con arreglos de sondas, en este caso usando arreglos de oligonucleótidos depositados en un soporte sólido. Macevicz (1989) propuso usar un conjunto pequeño de sondas con un diseño binario, el cual implicaba el uso de una base específica, e.g. A, y una base degenerada, que en el caso del ejemplo no debería ser A (i.e., una mezcla de T, G, y C). En 2004 Cowie et al. desarrollaron un enfoque basado en arreglos llamado secuenciación combinatoria por hibridación (‘combinatorial sequencing by hybridization’), en el cual se emplean dos conjuntos de sondas universales cortas, uno de ellos está fijo a un soporte sólido y el otro está libre en solución y marcado con un fluoróforo. El DNA blanco sin marcar es generado por PCR, después se mezcla con las sondas marcadas y con DNA ligasa, y finalmente se hibrida con las sondas fijas al soporte sólido. Cuando las sondas en ambos conjuntos hibridan con posiciones contiguas y complementarias del DNA blanco, se unen covalentemente por la DNA ligasa creando una larga sonda marcada y fija a la superficie del arreglo. El proceso combinatorio identifica todas las posibles sondas (secuencias) que son complementarias al blanco. Un lector estándar verifica las señales fluorescentes en cada posición del arreglo, para que posteriormente con el uso de software especializado se interprete

la imagen escaneada y se obtenga una lectura completa de la secuencia del templado obtenido por PCR.

Otra propuesta técnica que incorpora la hibridación es la secuenciación “polonial” (del inglés ‘polony’, una contracción de ‘polymerase colony’), la cual fue propuesta por Shendure et al. (2005), y que actualmente se conoce como SOLiD [Applied Biosystems (Foster City, CA, EUA)]. En ésta, la fase de secuenciación depende de la hibridación y ligado de oligonucleótidos (9-mer) marcados con fluorescencia.

1.3.2. PRINCIPIOS DE LA SBH

Puede considerarse que una característica común de la mayoría de los procedimientos que emplearon y emplean SBH, es el uso de arreglos de oligonucleótidos. Éstos, también conocidos comúnmente como microarreglos o chips de DNA, fueron propuestos simultánea e independientemente por Bains y Smith (1988), Lysov et al. (1988), Southern (1988), y Drmanac et al. (1989). La SBH consiste en poner en contacto un arreglo de sondas con una solución del DNA blanco, posteriormente un método bioquímico determina el subconjunto de sondas que se unen al blanco, i.e., el patrón de hibridación. Finalmente empleando un algoritmo combinatorio se reconstruye la secuencia del blanco a partir de su patrón de hibridación.

Cuando una sonda de longitud *n* hibrida específicamente con un DNA blanco, se revela la existencia dentro de la cadena de una o más secuencias *n*-mer complementarias. De acuerdo a Drmanac et al. (2002), una secuencia *n*-mer particular ocurre aproximadamente una vez cada $4^n/2$ pares de bases. Esta proporción estadística ayuda a definir parámetros de diseño apropiados para experimentos de SBH. Moléculas largas de DNA blanco requieren sondas largas, de manera que se eviten situaciones en las cuales cada sonda hibride con una o más secuencias complementarias del blanco, lo que ocasiona la pérdida de información específica. Las sondas demasiado cortas (<5-mer) se unen tan frecuentemente a la mayoría de los blancos que proveen información insuficiente o inútil. Por otro lado, sondas muy largas hibridan tan raramente que, de no existir información previa que guíe el diseño de las mismas, su utilidad es demasiado ineficiente para la mayoría de los procedimientos estándares de SBH. Pueden emplearse conjuntos muy grandes de sondas largas, pero este acercamiento resulta ser técnicamente demasiado complicado ya que deben ser sintetizadas y microarregladas decenas de miles de sondas individuales. Tales conjuntos grandes de sondas pueden requerir métodos especiales muy costosos de síntesis e hibridación en paralelo. En general, los experimentos de SBH pueden usar conjuntos de sondas con longitudes de 5 a 25-mer, aunque algunas aplicaciones emplean sondas más largas. Los métodos de SBH pueden abordar blancos de

largos muy diferentes, situados en un rango que va desde unas pocas bases hasta cientos de megabases, dependiendo del diseño experimental. Los llamados pozos de sondas, que contienen desde unas pocas hasta decenas de miles de sondas, pueden emplearse para minimizar el número requerido de experimentos de hibridación, un método especialmente efectivo para blancos cortos (Bains y Smith 1988; Drmanac et al. 2002; Macevicz 1989; Pevzner y Lipshutz 1995; Southern 1988).

En un experimento de SBH que use un conjunto completo de sondas *n*-mer (i.e., todas las sondas posibles de largo *n*), cada base de DNA es leída redundantemente por *n* sondas superpuestas. Con la figura 2 se ilustra como la “superposición” de las sondas puede usarse para ensamblar la secuencia del DNA blanco en un experimento de SBH que emplea sondas 8-mer. En este ejemplo, cada sonda que da hibridación positiva se superpone siete bases con su vecina más cercana, seis bases de su segunda vecina, y así sucesivamente. Este principio de superposición permite la determinación de secuencias que son mucho más largas que el largo de cada sonda, lo cual se efectúa al comparar y alinear *n*-1 o menos bases superpuestas que son compartidas por las sondas. Es importante señalar que no se requiere una superposición completa de *n*-1 bases para determinar las secuencias de los blancos; pueden ser suficientes superposiciones cortas, permitiéndose el uso de conjuntos incompletos de sondas (i.e., no todas las

secuencias posibles para sondas de un mismo largo). Otra ventaja adicional de la superposición de las sondas es la minimización de errores aleatorios, porque cada base es “leída” por múltiples sondas. Drmanac et al. (2002) sostienen que las ambigüedades que resultan de la existencia de grandes repetidos en tándem [e.g. $(CA)_{30}$], y regiones repetidas *n*-1 o mayores, puede resolverse por tratamientos experimentales adicionales, tales como el uso de oligonucleótidos más largos., entre otros.

Fig. 2. Superposición de sondas y ramas de ensamblaje. **A** Un blanco que contiene cebadores (‘primers’) en cada extremo de secuencia PPPPPPP y dos secuencias 7-mer repetidas. **B** El ensamble usando sondas superpuestas se inicia en el extremo 5′ empleando una sonda positiva que corresponda a la secuencia del cebador. La extensión de la secuencia continúa en dirección 3′ hasta el primer punto de ramificación, donde existen dos diferentes posibilidades para continuar la extensión (GGTCCCTc y GGTCCCTa). El proceso de ramificación (**C**)

genera muchas secuencias alternativas. Ya que la secuenciación debe terminar con el cebador 3', y los sub-fragmentos de secuencia ocurren sólo una vez entre los puntos de ramificación, entonces resulta el ensamblaje de dos posibles secuencias (**D**), las cuales difieren en la ubicación de las sub-secuencias CAA y AAT. (Modificada a partir de Drmanac et al. 2002).

Vale la pena hacer hincapié en dos conceptos relacionados con la SBH: 1) la ya mencionada superposición, que es el número de bases con el que dos o más sondas reconocen la misma secuencia de un DNA blanco (Fig. 2A); y 2) la "cobertura", que es el porcentaje de la secuencia del DNA blanco que es reconocida por el conjunto de sondas.

1.4. HIBRIDACIÓN VIRTUAL

La hibridación virtual (VH, por las siglas del inglés 'virtual hybridization') se fundamenta en algunas consideraciones relacionadas con el DNA, y que anteriormente fueron expuestas para la SBH, viz. complementariedad, desnaturalización y renaturalización, síntesis secuencia-específica de oligonucleótidos, entre otras. El concepto general de la VH implica la simulación por computadora (*in silico*) de eventos de hibridación entre moléculas de DNA. En el caso particular de este trabajo la VH corresponde al empleo de software (desarrollado ad hoc) para analizar la predicción de los patrones de hibridación que resultan al simular la interacción entre conjuntos de sondas

previamente diseñados y conjuntos de blancos de DNA, todo en condiciones de incubación preestablecidas (Méndez-Tenorio 2006; Maldonado-Rodríguez et al. 2007). El software predice la fuerza de enlace en sitios de hibridación alternativos a lo largo de las secuencias de los blancos, y lo hace calculando la energía libre de Gibbs ($\triangle G^o$) que se presenta cuando alguna sonda hibrida con una región específica de cualquier blanco (Reyes-López et al. 2003). Todas y cada una de las posiciones a través de las secuencias de los blancos son evaluadas. Enseguida el software identifica y señala aquellas posiciones de hibridación que son termodinámicamente más estables con relación a un limite dado de $\triangle G^o$, el cual simula la temperatura y condiciones de incubación reales. Los resultados obtenidos constituyen los patrones de hibridación virtual, y es a partir de estos patrones que se considera factible una reconstrucción de las secuencias de los blancos empleando combinatoria matemática. De esta manera se estaría llevando a cabo una secuenciación de DNA por hibridación virtual.

La VH fue descrita y empleada por Reyes-López et al. (2003), permitiéndoles verificar una correlación razonable entre los valores de $\triangle G^o$ calculados por VH para los rDNA 16S de 7 diferentes especies microbianas (1 *Bacillus* y 6 *Pseudomonas*), y la intensidad de las señales de hibridación obtenidas para estos mismos a través de un sistema de hibridación por chips de DNA. Lo

anterior les permitió concluir que las hibridaciones tanto perfectas como ambiguas contribuyen a la identificación microbiana a través de la obtención de la huella genómica por hibridación.

Otro trabajo reciente que utiliza la VH es el de Méndez-Tenorio et al. (2006), en el cual esta herramienta permitió la identificación exitosa de tipos y subtipos de papillomavirus humano (HPV, por las siglas del inglés ‘human papillomavirus’).

Belcaid et al. (2006) analizaron rearreglos cromosómicos a partir de la VH entre sondas y ocho diferentes genomas de plantas.

El uso de huellas genómicas generadas por VH, permitió a Casique-Almazán et al. (2007) identificar con alta sensibilidad tiroides, además de establecer también posibles relaciones filogenéticas entre ellos.

La figura 3 pretende ilustrar los ámbitos de una SBH clásica y de lo que sería un ciclo virtual de secuenciación, evidenciando el paralelismo de su desarrollo y fundamentos teóricos.

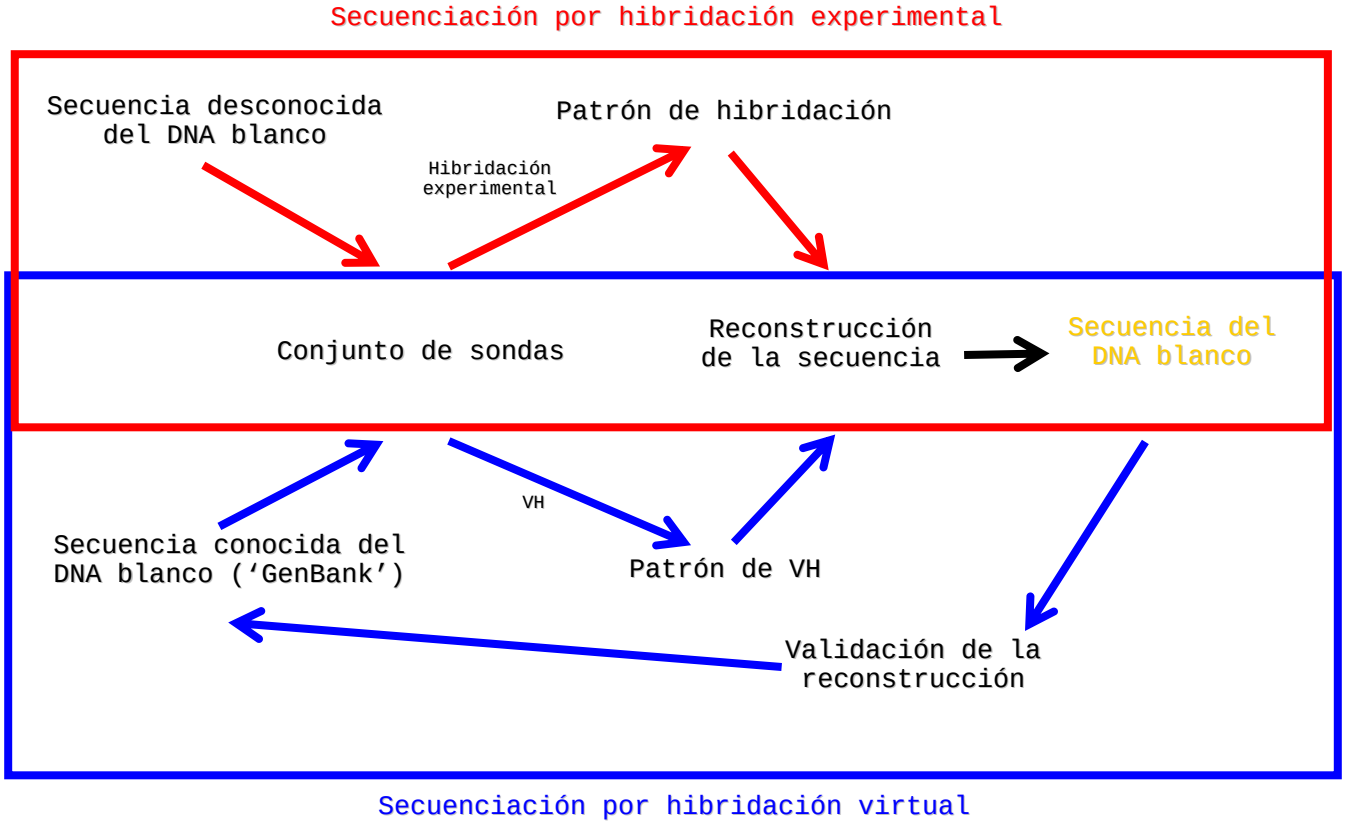

Fig. 3. Ámbitos independientes e interfase entre la secuenciación por hibridación experimental y la secuenciación por hibridación virtual. Se pretende ilustrar lo que se consideró como dos procedimientos paralelos en su desarrollo y fundamentos teóricos.

2. JUSTIFICACIÓN

Hoy en día un gran número de ciencias requieren de la secuenciación de DNA, e.g. Arqueología, Antropología, Genética, Biotecnología, Biología Molecular, Ciencias Forenses, entre otras. En muchas disciplinas está ocurriendo una revolución importante y silenciosa; la secuenciación de DNA promueve la obtención de nuevos conocimientos, lo que está originando un replanteamiento de los fundamentos conceptuales. Simultáneamente están emergiendo nuevos e importantes cuestionamientos, como aquellos de carácter bioético o los relacionados con la seguridad y salud públicas.

Entonces, ¿Por qué hacer mejoras a un método de secuenciación de DNA como la SBH? He aquí algunas respuestas: el conocimiento del genoma humano no es suficiente; todo aporte teórico-experimental que favorezca la alta productividad de secuenciación en los laboratorios de investigación, llevará este proceso más allá de un necesario y relevante ejercicio de descubrimiento, lo colocará como un ensayo rutinario para proponer y probar hipótesis. En la actualidad se hacen numerosos estudios orientados al desarrollo de nuevas metodologías de secuenciación, las cuales permitirán lo que se ha denominado “genómica personal”, i.e., el estudio rutinario de nuestros genomas individuales; para la identificación de patógenos nuevos

y conocidos; para la exploración de la diversidad microbiana hacia metas terapéuticas, ambientales y agrícolas.

Las tecnologías, y sus mejoras asociadas, que nos llevarán a alcanzar las metas mencionadas serán aquellas que provean mejorías en tres aspectos: longitud de lectura, rendimiento total, y costo.

El método de SBH presenta dificultades, en primera instancia las que tienen que ver con los algoritmos de reconstrucción de las secuencias de los blancos, y en segunda instancia las experimentales, que se deben tanto a la hibridación ambigua como a la estabilidad sumamente variable de las sondas de una misma longitud. La mayoría de las investigaciones de SBH con microarreglos consideran conjuntos muy grandes de sondas de un tamaño determinado, incluso todas las posibles secuencias para ese tamaño. Por tanto, y dado que la tecnología limita el número de sondas en un microarreglo, una cuestión importante sería el diseño de conjuntos más pequeños con capacidad para secuenciar cadenas de DNA.

3. OBJETIVOS

3.1. OBJETIVO GENERAL

Diseñar un programa de cómputo que permita efectuar la reconstrucción de secuencias conocidas de DNA a partir de la hibridación virtual que ocurra entre estás y un conjunto especial de sondas.

3.2. OBJETIVOS ESPECÍFICOS

Diseñar rutinas computacionales que efectúen el análisis de superposición y cobertura de las secuencias conocidas de DNA empleando un conjunto de sondas 8-mer.

Usando los resultados del análisis antes mencionado, implementar un procedimiento de cómputo que permita la reconstrucción de las secuencias.

Hacer una valoración de las características que presentan las reconstrucciones generadas.

4. MATERIALES Y MÉTODOS

4.1. PROGRAMACIÓN

Todos los procedimientos computacionales desarrollados en este trabajo fueron escritos de forma estructurada en el entorno de programación integrado Borland Delphi 6 (Borland Software Corporation 2002).

4.2. DNA BLANCOS

Se emplearon 68 secuencias de HPV que fueron obtenidas por Internet de la base de datos *Entrez Nucleotide* (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide). Estrictamente cualesquiera secuencias reales hubieran sido útiles, pues el único carácter considerado es que fueran conocidas en su totalidad, lo que permitiría una evaluación más objetiva de su reconstrucción. El rango de largo de las secuencias fue de 50 a 7,909 nucleótidos. Las secuencias fueron archivadas en formato FASTA.

Se diseñó un DNA blanco hipotético (DNA blanco modelo) 44 nt cuya secuencia provenía de las secuencias de un conjunto de 10 sondas hipotéticas (sondas modelo) totalmente diferentes, de tal

manera que al ser procesados empleando las rutinas computacionales que se describirán posteriormente, podrían analizarse con mayor facilidad los resultados. La finalidad principal de utilizar el DNA blanco y las sondas modelo, fue vislumbrar la implementación del procedimiento para la reconstrucción de las secuencias. A continuación se muestra la secuencia del DNA blanco modelo:

ACTCTTTACTCAAGGCTACGTGGAAGCCGCTAATCAGTTCATAC

4.3. SONDAS

Empleando el software Universal Fingerprint Chip (UFC) Designer (Beattie K.L. et al. 2003) se obtuvo un conjunto de 511 sondas 8-mer. Éste se diseñó con la única consideración de que entre todas y cada una de las sondas habría dos diferencias en la secuencia. El conjunto proviene de todas las secuencias posibles 8-mer (65,536), y fue obtenido a través de los siguientes tres pasos de agrupamiento que efectúa el programa de cómputo: 1) agrupamiento por sustitución (‘substitution clustering’), que sustituye bases para obtener secuencias con dos diferencias entre sí; 2) agrupamiento por bloque (‘block clustering’), elimina las dos diferencias de los extremos de cada sonda; y 3) agrupamiento por refinamiento (‘refine clustering’), que separa las dos

diferencias. Antes de cada paso de agrupamiento se efectuó una aleatorización de los subconjuntos considerando un número diferente como semilla. 5673, 7899 y 653, fueron las semillas. A continuación se detalla como fue siendo depurado el conjunto inicial por medio de los tres pasos de agrupamiento del software UFC Designer: el agrupamiento por sustitución encontró 8,033 sondas con dos diferencias entre sí, esto a partir de las 65,536 secuencia 8-mer posibles; el agrupamiento por bloque eliminó los casos en los que las diferencias ocurrían en los extremos, quedando 1,062; y finalmente el agrupamiento por refinamiento seleccionó 511 casos en los que las dos diferencias se encontaron espaciadas.

En la sección anterior se hizo referencia al diseño del DNA blanco y las sondas modelo. A continuación se enlistan las secuencias de las sondas modelo:

AGGCTACG
ATCAGTTC
TGGAAGCC
ACTCTTTA
GCTAATCA
CTCAAGGC
GTTCATAC
TACGTGGA
TTTACTCA
AGCCGCTA

4.4. VH

Cada uno de los 68 DNA blancos fue hibridado virtualmente con el conjunto de 511 sondas empleando el módulo Virtual Hybridization del software UFC Designer. Lo anterior se realizó con solo una de las cadenas. La VH sólo tuvo como limitante la permisividad máxima de un apareamiento imperfecto (‘mismatch’), además de no considerar valores termodinámicos restrictivos.

4.5. COBERTURA Y SUPERPOSICIÓN A PARTIR DE UNA VH SIN RESTRICCIONES TERMODINÁMICAS

Debido a las consideraciones de experimentación *in silico* expuestas para la VH (un solo apareamiento imperfecto, y ninguna restricción termodinámica), en el caso de esta investigación los resultados de una VH sin restricciones termodinámicas (VHX) fueron los mismos que los producidos por la VH. Por lo tanto toda metodología y resultado presentados en adelante hacen referencia a la VHX.

En la figura 4 se presenta de manera general el diagrama de flujo de las rutinas computacionales que efectuaron primeramente la VHX, en segunda instancia un ordenamiento de los resultados de esa hibridación, y finalmente el análisis de cobertura y

superposición. Para este último ténganse en cuenta los conceptos mencionados en la sección 1.3.2.

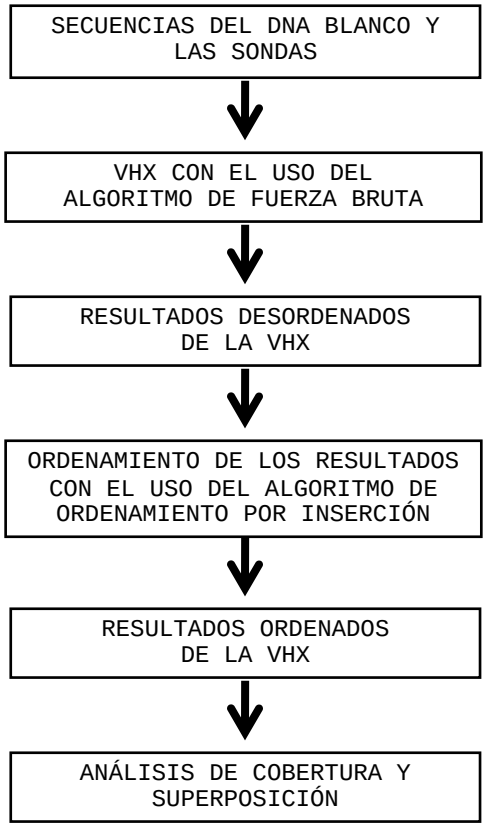

Fig. 4. Diagrama de flujo de las rutinas que realizan VHX, ordenamiento de los resultados de la VHX, y análisis de cobertura y superposición. La secuencia del DNA blanco es comparada con todas y cada una de las secuencias de las 511 sondas 8-mer usando el algoritmo de la fuerza bruta (esto estrictamente es la VHX). Con la finalidad de facilitar la comprensión de los resultados de la VHX y simplificar el análisis posterior de cobertura y superposición, los resultados desordenados de la VHX se organizan con el algoritmo de ordenamiento por inserción, esto en función de la posición del DNA blanco en el las sondas son idénticas al menos 87.5%. Los resultados ordenados de la VHX son procesados con otro algoritmo para analizar la cobertura y la superposición.

4.5.1. VHX

Se diseñó un procedimiento que permitiera obtener, a partir de las secuencias de los DNA blancos (una sola cadena) y las sondas, los casos de las posiciones en las que las dos secuencias (DNA blanco/sonda) son idénticas en al menos siete posiciones (un solo apareamiento imperfecto). Lo mencionado es el razonamiento detrás de la VHX.

Para poder ejecutar la tarea antes descrita se empleó lo que en programación se conoce como un algoritmo de búsqueda de texto (‘string searching’), específicamente el algoritmo llamado de la fuerza bruta (‘brute-force algorithm’). Este es un algoritmo de búsqueda exhaustiva, pues funciona al comparar cada una de las secuencias de las sondas con cada una de las posiciones de las secuencias de los DNA blancos. Con este programa se buscaron las secuencias de las sondas que fueran idénticas con alguna sub-secuencia del DNA blanco al menos en 7 de las 8 posiciones comparables (87.5% idénticas).

4.5.2. ORDENAMIENTO DE LOS RESULTADOS DE LA VHX

El conjunto de sondas obtenido de la rutina anterior fue ordenado con un algoritmo de ordenamiento llamado de ordenamiento por

inserción (‘insertion sort’). El orden fue en función de la posición del DNA blanco en que cada una fue idéntico al menos 87.5%.

4.5.3. ANÁLISIS DE COBERTURA Y SUPERPOSICIÓN

Los datos de salida de la rutina antes descrita, se emplearon como entrada para el procedimiento que analizó la superposición y la cobertura resultantes de la VHX entre cada uno de los DNA blancos de HPV y el conjunto de 511 sondas 8-mer. A continuación se enlistan los casos de superposición entre sondas que fueron considerados, además del razonamiento matemático subyacente incluido en el algoritmo. Considérese que el análisis de superposición se hizo por pares de sondas, y para todos los pares la primera posición de la sonda (igual a la posición del DNA blanco en la que hibridó) que hibridó en una posición del DNA blanco más cercana al inicio de éste se denominó posA, mientras que la primera posición de la otra sonda (igual a la posición del DNA blanco en la que hibridó) del par en cuestión fue denominada posB. El valor de n es la longitud de las sondas, i.e., 8.

Caso 1. No existe superposición entre sondas y se presenta entre ellas una región descubierta del DNA blanco, e.g.:

```
  CCCGTAGT   ACCCTAAC        sondas separadas
CACCCGTAGTACTAACCTAACAT      DNA blanco
  ********   * ******        identidades absolutas
```

Si posB - posA >= (n + 1) entonces la superposición es 0 y el caso es el 1.

Caso 2. No existe superposición entre sondas, aunque al estar contiguas no se presenta entre ellas una región descubierta del DNA blanco, e.g.:

```
   AACATAGTGTGCTCCA          sondas contiguas
CCTAACATTGTGTGCTACAGCAT      DNA blanco
   ***** ******* **          identidades absolutas
```

Si posB - posA = n entonces también la superposición es 0 y el caso es el 2.

Caso 3. Existe superposición de una sola posición, e.g.:

```
         TGTGCTAC            sonda
   AACATAGT                  sonda
CCTAACATTGTGTGCTACAGCAT      DNA blanco
   ***** *********           identidades absolutas
```

Si posB - posA = (n - 1) entonces la superposición es de 1 y el caso es el 3.

Caso 4. Existe superposición de más de una sola posición y de menos de n-1 posiciones, donde n es la longitud de las sondas, e.g.:

```
     TTAGGCAG              sonda
  GACGTTAG                 sonda
CTGACTTTAGGGAGTATATTATT    DNA blanco
  *** ***** **             identidades absolutas
```

Si posB - posA < (n - 1) y también posB - posA > 0 entonces la superposición es igual a (posA + n) – posB y el caso es el 4. Este caso es altamente informativo, e idealmente mientras mayor

sea el número de hibridaciones positivas de este tipo mayor sería la probabilidad de reconstrucciones exitosas.

Caso 5. Existe superposición de n posiciones, e.g.:

```
      TGTGGATG             sonda
      TGTAGAGG             sonda
TTAGACATGTGGAGGAATATGCA    DNA blanco
      *** ** *             identidades absolutas
```

Si posB - posA < (n - 1) y también if posB - posA = 0 entonces la superposición es igual a n y el caso es el 5.

4.6. RECONSTRUCCIÓN DE SECUENCIAS A PARTIR DE LA IDENTIDAD ENTRE LOS PREFIJOS Y LOS SUFIJOS OBTENIDOS DE LAS SONDAS QUE HIBRIDARON

4.6.1. PREFIJOS Y SUFIJOS

Con los resultados desordenados del patrón de VHX (sección 4.5.1) como entrada, se implementó una rutina que obtuvo de las sondas todos los prefijos y sufijos posibles con longitudes desde 2 hasta 7 posiciones (las sondas son 8-mer).

4.6.2. COMPARACIÓN ENTRE PREFIJOS Y SUFIJOS

Con el uso de dos variantes esencialmente idénticas de una rutina, se efectuó una tarea relevantísima para la reconstrucción: primero se encontraron identidades al comparar todos y cada uno de los prefijos con todos los sufijos, y posteriormente se encontraron las identidades al comparar todos y cada uno de los sufijos con todos los prefijos.

4.6.3. RECONSTRUCCIÓN DE SECUENCIAS

A partir del análisis de los resultados obtenidos con el DNA blanco y las sondas modelo se desarrollaron otros dos procedimientos cuyo objetivo común fue reconstruir las secuencias de los DNA blancos, esto empleando como entrada las listas de identidades generadas por las rutinas descritas en la sección 4.6.2. El primero de ellos empleó como inicio para la reconstrucción todas y cada una de las sondas presentes en la lista de identidades entre prefijos y sufijos, lo cual produjo un número de reconstrucciones igual al número de sondas en el listado. El segundo procedimiento efectuó lo mismo que el anterior, pero usando las identidades entre sufijos y prefijos. Esto también generó una cantidad de reconstrucciones igual al número de sondas en el listado. Las secuencias reconstruidas aparecen en pantalla, junto con las sondas que, a pesar de aparecer en el listado de identidades que se usó, no fueron empleadas en la reconstrucción.

La ruta general del algoritmo de reconstrucción puede representarse con el diagrama de flujo de la figura 5.

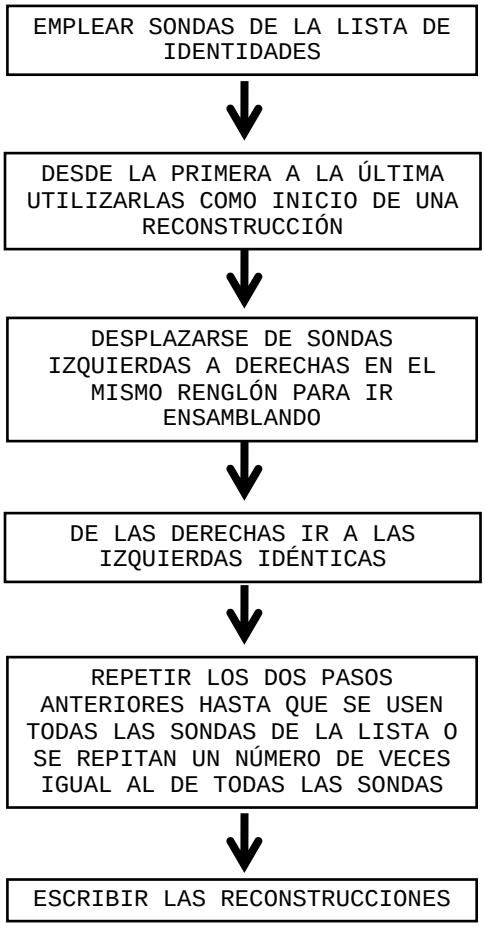

Fig. 5. Diagrama de flujo de la rutina que devino en la reconstrucción de secuencias. Las rutinas de la sección 4.6.2 generan listas de identidades entre afijos (prefijos y sufijos). Todas y cada una de las sondas en estas listas fungieron como inicio de una secuencia reconstruida. El algoritmo de reconstrucción funciona a partir de que la sonda de inicio continúa con la siguiente en función del afijo idéntico compartido. Posteriormente está segunda sonda se busca y encuentra en la columna de donde provino la primera, continuándose la reconstrucción empleando la secuencia de la sonda con la que comparte un afijo idéntico. Esta rutina se repite un número de veces igual al del número total de sondas que hay en las listas de afijos idénticos. Finalmente cada secuencia reconstruida es escrita en pantalla, junto con las secuencias de las sondas no empleadas en dicha reconstrucción.

Debe recalcarse que la implementación de las rutinas de reconstrucción fue facilitada enormemente por el diseño del DNA blanco y las sondas modelo.

4.6.4. EVALUACIÓN DE SECUENCIAS RECOSNTRUIDAS

Para evaluar el funcionamiento del algoritmo de reconstrucción se emplearon primero el DNA blanco y las sondas modelo. Posteriormente se usaron como ejemplos reales dos DNA blancos de HPV, el primero 245 nt (sin redundancia en las señales de VH) y el segundo 1,000 nt (redundante en las señales de VH), junto con el conjunto de 511 sondas 8-mer. En los casos de las reconstrucciones de las secuencias provenientes de HPV, fueron elaboradas a partir del anañisis de una sola cadena.

5. RESULTADOS Y DISCUSIÓN

Téngase en cuenta lo mencionado en la sección 4.5.

5.1. VHX

Algo indispensable para abordar la reconstrucción de secuencias, fue evaluar previamente la cobertura y la superposición de las sondas que resultaron de efectuar la VHX.

Al emplear el procedimiento de VHX (4.5.1) se obtuvieron archivos de texto con tablas de tres columnas, en ellas se muestran: las secuencias de las sondas que son similares al DNA blanco en al menos 87.5%; la posición del DNA blanco en donde se inicia esta similitud; y el porcentaje de identidad (87.5 o 100%). La figura 6 presenta como ejemplo de este tipo de tablas, la obtenida para el DNA blanco y las sondas modelo (secciones 4.2 y 4.3).

Secuencia (5′ → 3′)	Posición	identidad (%)
AGGCTACG	13	100.0
ATCAGTTC	33	100.0
TGGAAGCC	21	100.0
ACTCTTTA	1	100.0
GCTAATCA	29	100.0
CTCAAGGC	9	100.0
GTTCATAC	37	100.0
TACGTGGA	17	100.0
TTTACTCA	5	100.0
AGCCGCTA	25	100.0

Fig. 6. Texto del archivo resultante de la rutina de VHX al utilizar el DNA blanco y las sondas modelo. Las columnas de izquierda a derecha muestran la

secuencia de la sonda que hibridó, la posición del DNA blanco donde lo hizo, y el porcentaje de identidad entre la sonda y la sub-secuencia del DNA blanco con la que hibridó.

Como puede observarse las características del DNA blanco hipotético permitieron hibridaciones virtuales perfectas (100% de identidad) con todas y cada una de las sondas modelo. Al emplear las secuencias de HPV las tablas resultantes permitieron observar, en primera instancia, que no todas las sondas 8-mer dieron señal, y que en casos donde la hubo, la hibridación fue imperfecta. Esto se ejemplifica en la figura 7 con el uso de un DNA blanco de 71 nt, cuya secuencia también se anota.

TGTACATGAACTAGAGTAAACCTTTTTTATACAGTGTGTGGTGACGTTTAGTTATATATAATGAAACCTAG

Secuencia (5′ → 3′)	Posición	identidad (%)
GACCTTTT	19	87.5
ATAAATAA	54	87.5
GTATGGTG	36	87.5
ATGTAATG	56	87.5
AGTGATAT	50	87.5
GTTATACA	26	87.5
GTTATACA	51	87.5

Fig. 7. Texto del archivo resultante de la rutina de VHX al utilizar un DNA blanco 71 nt de HPV y el conjunto de 511 sondas 8-mer. Las columnas de izquierda a derecha muestran la secuencia de la sonda que hibridó, la posición del DNA blanco donde lo hizo, y el porcentaje de identidad entre la sonda y la sub-secuencia del DNA blanco con la que hibridó.

5.2. ORDENAMIENTO DE LOS RESULTADOS DE LA VHX

Con la finalidad de facilitar la comprensión de los resultados de la VHX, así como la implementación posterior del algoritmo que

efectuaría el análisis de cobertura y superposición, las tablas de VHX fueron ordenadas en función de la posición del DNA blanco en que cada sonda fue idéntica al menos 87.5% a la secuencia de éste.

La figura 8 muestra el ordenamiento antes explicado para el DNA blanco y las sondas modelo.

Secuencia (5′ → 3′)	Posición	identidad (%)
ACTCTTTA	1	100.0
TTTACTCA	5	100.0
CTCAAGGC	9	100.0
AGGCTACG	13	100.0
TACGTGGA	17	100.0
TGGAAGCC	21	100.0
AGCCGCTA	25	100.0
GCTAATCA	29	100.0
ATCAGTTC	33	100.0
GTTCATAC	37	100.0

Fig. 8. Texto del archivo resultante de la VHX ordenada al utilizar el DNA blanco y las sondas modelo. Las columnas de izquierda a derecha muestran la secuencia de la sonda que hibridó, la posición del DNA blanco donde lo hizo, y el porcentaje de identidad entre la sonda y la sub-secuencia del DNA blanco con la que hibridó.

5.3. ANÁLISIS DE COBERTURA Y SUPERPOSICIÓN

La rutina implementada para analizar superposición y cobertura, generó una pantalla que indica el número de bases cubiertas en relación al número total de bases del DNA blanco, la frecuencia de cada caso de superposición (mencionados en la sección 4.4.2), el número total de señales (hibridaciones) y el número de sondas diferentes que hibridan. La figura 9 muestra una reproducción del

texto de la pantalla obtenida para el DNA blanco y las sondas modelo.

La importancia de conocer la cobertura estribó indirectamente en la posibilidad de afirmar o negar la capacidad del conjunto de 511 sondas para generar reconstrucciones completas del DNA blanco.

```
Se cubrieron 44 bases de un total de 44

Sondas en el Caso 1 = 0
Sondas en el Caso 2 = 0
Sondas en el Caso 3 = 0
Sondas en el Caso 4 = 10
Sondas en el Caso 5 = 0

El total de hibridaciones fue 10

El número de sondas diferentes que hibridan es 10
```

Fig. 9. Texto de la pantalla que resulta de utilizar la rutina para analizar superposición y cobertura a partir de la VHX del DNA blanco y las sondas modelo.

El diseño del DNA blanco y las sondas modelo permitió observar una cobertura total, así como todas las superposiciones en el caso considerado como ideal para una reconstrucción facilitada y confiable.

La figura 10 muestra los resultados de superposición y cobertura para el DNA blanco de 71 nt (anotado anteriormente) que se hibridó virtualmente con el conjunto de 511 sondas 8-mer antes mencionado. Obsérvese que se presentan casos no ideales para la reconstrucción.

```
Se cubrieron 37 bases de un total de 71

Sondas en el Caso 1 = 3
Sondas en el Caso 2 = 0
Sondas en el Caso 3 = 1
Sondas en el Caso 4 = 3
Sondas en el Caso 5 = 0

El total de hibridaciones fue 7

El numero de sondas diferentes que hibridan es 6
```

Fig. 10. Texto de la pantalla que resulta de utilizar la rutina para analizar superposición y cobertura con un DNA blanco 71-mer de HPV y el conjunto de 511 sondas 8-mer.

Como pudo observarse para el caso del DNA blanco 71-mer, y la mayoría del resto de las 68 secuencias de HPV, la cobertura parcial vaticinaba, en caso de poder obtenerlas, reconstrucciones también parciales.

Para poder hacer una interpretación del análisis de cobertura y superposición, se elaboraron las gráficas presentadas en las figuras 11, 12, y 13. Para la primera fueron empleados los resultados obtenidos con las 68 secuencias de HPV y el conjunto de sondas 8-mer. Con las dos restantes se usaron los resultados de las primeras 60 secuencias de HPV (con un rango de longitud de 50 a los 1,000 nt) y el conjunto de sondas 8-mer.

En primera instancia la figura 11 permite observar que a medida que la longitud del DNA blanco se incrementa, el porcentaje en que es cubierto tiende a estabilizarse en aproximadamente 80%, lo que confirmaba gráficamente lo mencionado respecto a que sólo podrían esperarse reconstrucciones parciales

de los DNA blancos, esto usando para la VHX el conjunto de 511 sondas. En segunda instancia se observa que existe una relación lineal entre el tamaño del DNA blanco y el número de señales en el caso ideal de superposición (Caso 4), dato que parece alentador, sin embargo debe considerarse que dicha relación no considera la "redundancia" en la aparición de sondas, i.e., el caso de que una misma sonda hibride positivamente en varias regiones del DNA blanco.

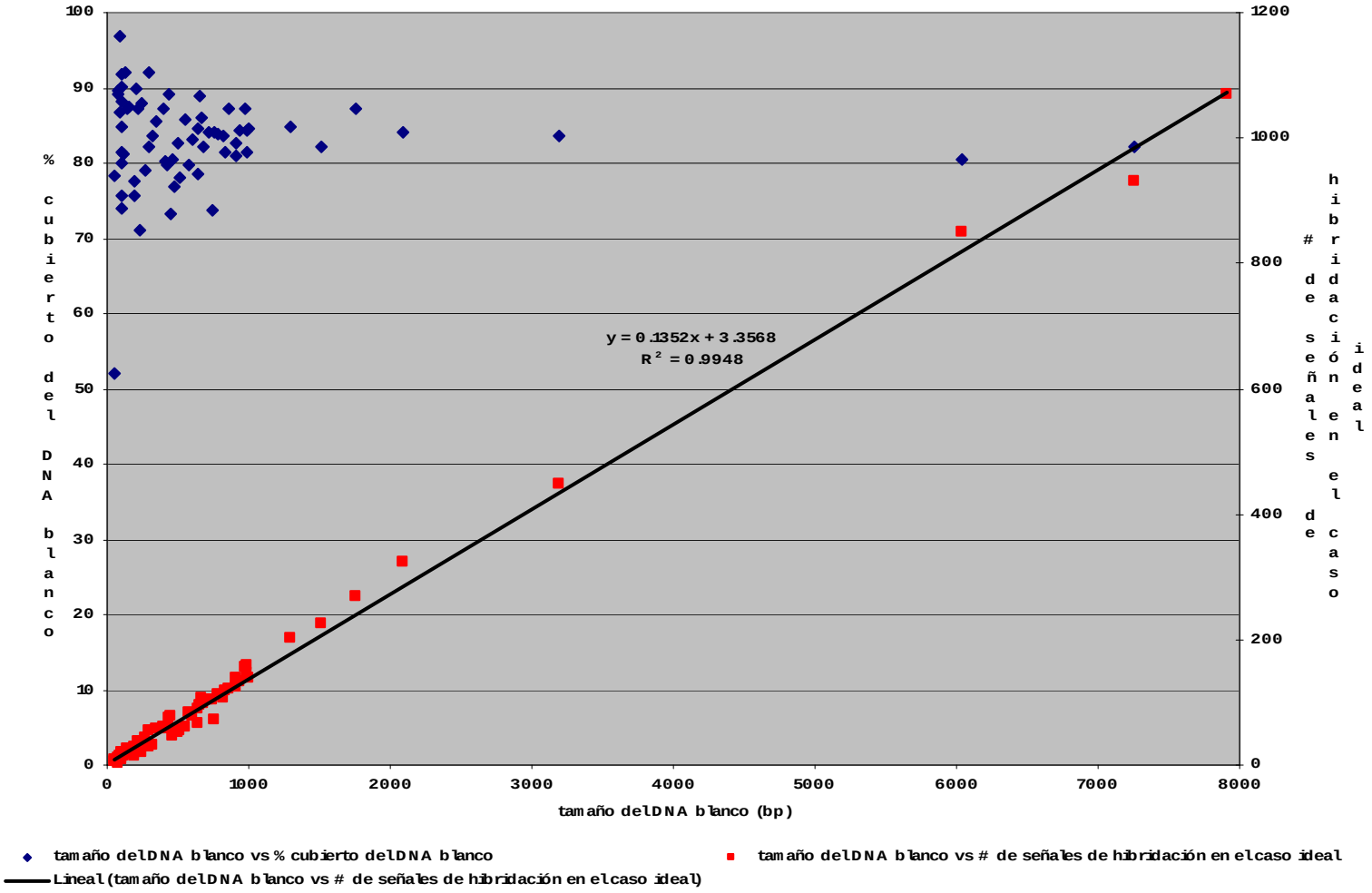

Fig. 11. Variación del porcentaje cubierto del DNA blanco y del número de sondas que hibridaron en el caso de superposición ideal (Caso 4) respecto a la

longitud del DNA blanco. A medida que la longitud del DNA blanco se incrementa, el porcentaje en que es cubierto con el conjunto de 511 sondas tiende a estabilizarse en aproximadamente 80%, lo que confirma gráficamente lo mencionado respecto a que sólo podrían esperarse reconstrucciones parciales de los DNA blancos. Existe una relación lineal entre el tamaño del DNA blanco y el número de señales de hibridación en el caso de superposición ideal (Caso 4), dato que parece alentador, sin embargo debe considerarse que dicha relación no considera la redundancia en la aparición de sondas

La figura 12 permite concluir que en la mayoría de los casos es un hecho la redundancia de las señales de hibridación, característica que presentó un inconveniente reconstructivo que el algoritmo de reconstrucción no podría manejar.

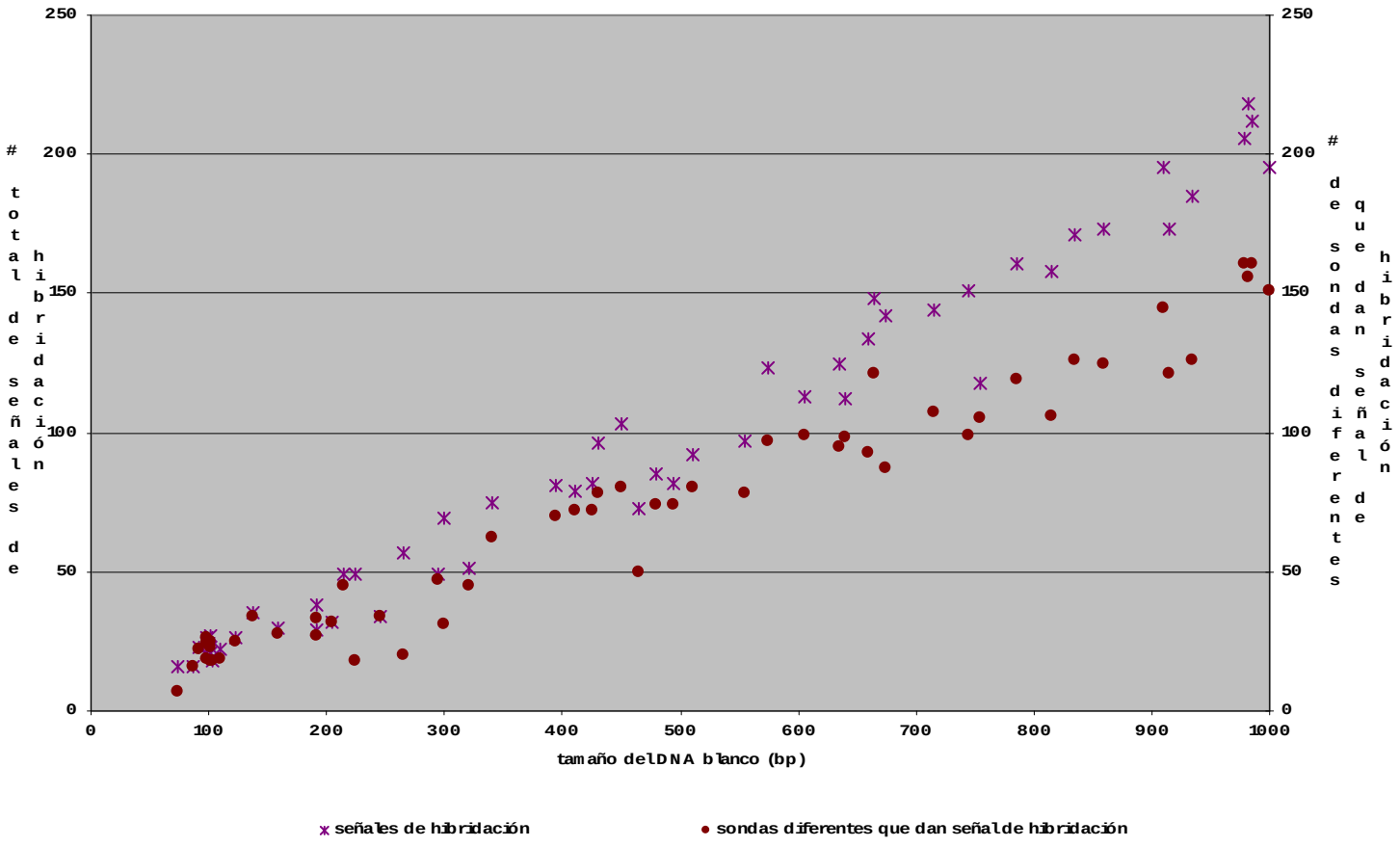

Fig. 12. Variación del número total de señales de hibridación y del número de sondas diferentes dentro de esas señales con respecto a la longitud del DNA

blanco. Puede apreciarse que en la mayoría de los casos para cada DNA blanco el número total de señales de hibridación implicaba redundancia variable en las sondas.

La relación entre el número total de señales y el número de sondas diferentes que dan señal se muestra gráficamente en la figura 13. La relación en la mayoría de los casos nunca fue de uno a uno, lo que representa la existencia de la redundancia antes mencionada.

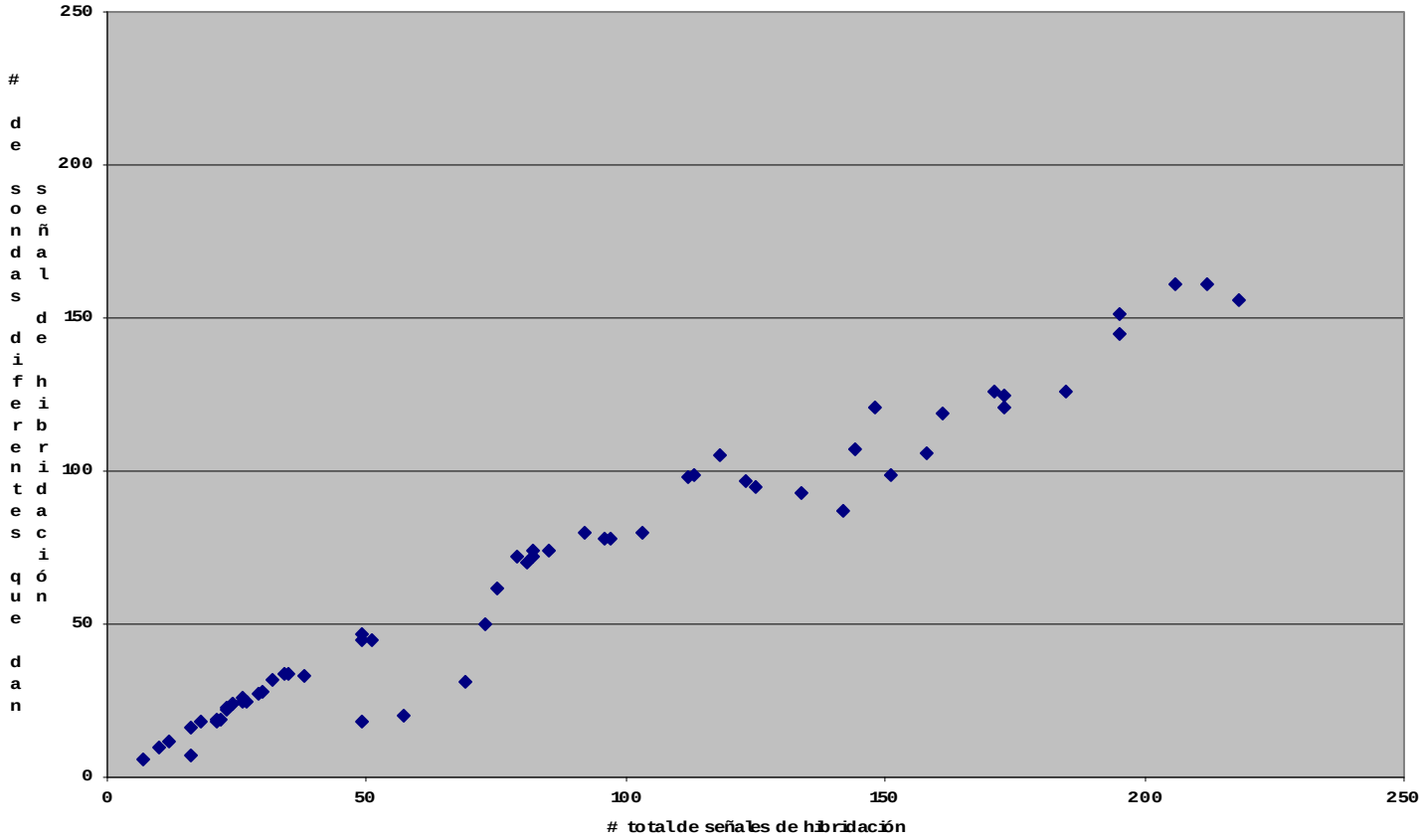

Fig. 13. Número de sondas diferentes que dan señal de hibridación con respecto al número total de señales de hibridación. La relación en la mayoría de los casos nunca fue de uno a uno, lo que representa la existencia de la redundancia antes mencionada.

Así, extrapolando a las longitudes correspondientes de los DNA blancos, pudo concluirse que teóricamente para elaborar una posible reconstrucción de secuencias (a partir de los patrones de VHX con las 511 sondas 8-mer) debieron involucrarse longitudes de DNA blancos de hasta aproximadamente 1,000 nt.

5.4. PREFIJOS Y SUFIJOS

El procedimiento que emplea como entrada los resultados desordenados obtenidos de la rutina de VHX (4.5.1.), genera una tabla con los siguientes datos (por columna de izquierda a derecha): el número de sonda de acuerdo a la posición en la tabla de resultados desordenados; la secuencia de la sonda; el prefijo extraído de la sonda; el número de prefijo o sufijo; y el sufijo extraído de la sonda. La figura 14 presenta los resultados obtenidos para el DNA blanco y las sondas modelo. La rutina obtiene de cada una de las sondas 8-mer que hibridaron en la VHX, seis diferentes prefijos y seis diferentes sufijos, los cuales van desde 2 hasta 7-mer.

1	AGGCTACG	AG	1	CG	6	CTCAAGGC	CT	31	GC
1	AGGCTACG	AGG	2	ACG	6	CTCAAGGC	CTC	32	GGC
1	AGGCTACG	AGGC	3	TACG	6	CTCAAGGC	CTCA	33	AGGC
1	AGGCTACG	AGGCT	4	CTACG	6	CTCAAGGC	CTCAA	34	AAGGC
1	AGGCTACG	AGGCTA	5	GCTACG	6	CTCAAGGC	CTCAAG	35	CAAGGC
1	AGGCTACG	AGGCTAC	6	GGCTACG	6	CTCAAGGC	CTCAAGG	36	TCAAGGC
2	ATCAGTTC	AT	7	TC	7	GTTCATAC	GT	37	AC
2	ATCAGTTC	ATC	8	TTC	7	GTTCATAC	GTT	38	TAC
2	ATCAGTTC	ATCA	9	GTTC	7	GTTCATAC	GTTC	39	ATAC
2	ATCAGTTC	ATCAG	10	AGTTC	7	GTTCATAC	GTTCA	40	CATAC
2	ATCAGTTC	ATCAGT	11	CAGTTC	7	GTTCATAC	GTTCAT	41	TCATAC
2	ATCAGTTC	ATCAGTT	12	TCAGTTC	7	GTTCATAC	GTTCATA	42	TTCATAC
3	TGGAAGCC	TG	13	CC	8	TACGTGGA	TA	43	GA
3	TGGAAGCC	TGG	14	GCC	8	TACGTGGA	TAC	44	GGA
3	TGGAAGCC	TGGA	15	AGCC	8	TACGTGGA	TACG	45	TGGA
3	TGGAAGCC	TGGAA	16	AAGCC	8	TACGTGGA	TACGT	46	GTGGA
3	TGGAAGCC	TGGAAG	17	GAAGCC	8	TACGTGGA	TACGTG	47	CGTGGA
3	TGGAAGCC	TGGAAGC	18	GGAAGCC	8	TACGTGGA	TACGTGG	48	ACGTGGA
4	ACTCTTTA	AC	19	TA	9	TTTACTCA	TT	49	CA
4	ACTCTTTA	ACT	20	TTA	9	TTTACTCA	TTT	50	TCA
4	ACTCTTTA	ACTC	21	TTTA	9	TTTACTCA	TTTA	51	CTCA
4	ACTCTTTA	ACTCT	22	CTTTA	9	TTTACTCA	TTTAC	52	ACTCA
4	ACTCTTTA	ACTCTT	23	TCTTTA	9	TTTACTCA	TTTACT	53	TACTCA
4	ACTCTTTA	ACTCTTT	24	CTCTTTA	9	TTTACTCA	TTTACTC	54	TTACTCA
5	GCTAATCA	GC	25	CA	10	AGCCGCTA	AG	55	TA
5	GCTAATCA	GCT	26	TCA	10	AGCCGCTA	AGC	56	CTA
5	GCTAATCA	GCTA	27	ATCA	10	AGCCGCTA	AGCC	57	GCTA
5	GCTAATCA	GCTAA	28	AATCA	10	AGCCGCTA	AGCCG	58	CGCTA
5	GCTAATCA	GCTAAT	29	TAATCA	10	AGCCGCTA	AGCCGC	59	CCGCTA
5	GCTAATCA	GCTAATC	30	CTAATCA	10	AGCCGCTA	AGCCGCT	60	GCCGCTA

Fig. 14. Texto de la pantalla que resulta de utilizar la rutina para extracción de prefijos y sufijos después de realizar una VHX con el DNA blanco y las sondas modelo. Se obtienen de cada una de las sondas 8-mer que hibridaron seis diferentes prefijos y seis diferentes sufijos, los cuales van desde 2 hasta 7-mer.

Las figuras 15 y 16 ejemplifican, en el caso del DNA blanco y las sondas modelo, las tablas generadas al realizar la búsqueda de identidades entre prefijos y sufijos, y viceversa. Dichos procedimientos se mencionaron en la sección 4.6.1.

Las tablas ejemplificadas con la figura 15 obedecieron al siguiente formato, en un mismo renglón y de izquierda a derecha: sonda que aporta el prefijo; número de prefijo de acuerdo a la posición en la tabla de prefijos y sufijos; el prefijo; número de

sufijo idéntico de acuerdo a la posición en la tabla de prefijos y sufijos; y sonda que aporta el sufijo idéntico.

RESULTADOS DE LA BÚSQUEDA DE INDENTIDAD ENTRE
PREFIJOS Y SUFIJOS

```
AGGCTACG, Prefijo 3  AGGC => Sufijo(s) 33 CTCAAGGC,
ATCAGTTC, Prefijo 9  ATCA => Sufijo(s) 27 GCTAATCA,
TGGAAGCC, Prefijo 15 TGGA => Sufijo(s) 45 TACGTGGA,
ACTCTTTA, Prefijo 19 AC   => Sufijo(s) 37 GTTCATAC,
GCTAATCA, Prefijo 25 GC   => Sufijo(s) 31 CTCAAGGC,
GCTAATCA, Prefijo 27 GCTA => Sufijo(s) 57 AGCCGCTA,
CTCAAGGC, Prefijo 33 CTCA => Sufijo(s) 51 TTTACTCA,
GTTCATAC, Prefijo 39 GTTC => Sufijo(s)  9 ATCAGTTC,
TACGTGGA, Prefijo 43 TA   => Sufijo(s) 19 ACTCTTTA, 55 AGCCGCTA,
TACGTGGA, Prefijo 44 TAC  => Sufijo(s) 38 GTTCATAC,
TACGTGGA, Prefijo 45 TACG => Sufijo(s)  3 AGGCTACG,
TTTACTCA, Prefijo 51 TTTA => Sufijo(s) 21 ACTCTTTA,
AGCCGCTA, Prefijo 57 AGCC => Sufijo(s) 15 TGGAAGCC,
```

Fig. 15. Tabla que resulta de utilizar el procedimiento para buscar identidades al comparar prefijos con sufijos. Obtenida con el DNA blanco y las sondas modelo. En un mismo renglón y de izquierda a derecha: secuencia de la sonda que aporta el prefijo; número de prefijo de acuerdo a la posición en la tabla de prefijos y sufijos; el prefijo; número de sufijo idéntico de acuerdo a la posición en la tabla de prefijos y sufijos; y sonda que aporta el sufijo idéntico. Obsérvese que en el noveno renglón el prefijo de la sonda TACGTGGA se identificó con el sufijo de dos diferentes sondas, esto implicaría una bifurcación en la reconstrucción.

La figura 16 muestra el siguiente formato en cada renglón de izquierda a derecha: sonda que aporta el sufijo; número de sufijo de acuerdo a la posición en la tabla de prefijos y sufijos; el sufijo; número de prefijo idéntico de acuerdo a la posición en la tabla de prefijos y sufijos; y sonda que aporta el prefijo idéntico.

RESULTADOS DE LA BÚSQUEDA DE INDENTIDAD ENTRE
SUFIJOS Y PREFIJOS

```
AGGCTACG, Sufijo 3  TACG => Prefijo(s) 45 TACGTGGA,
ATCAGTTC, Sufijo 9  GTTC => Prefijo(s) 39 GTTCATAC,
TGGAAGCC, Sufijo 15 AGCC => Prefijo(s) 57 AGCCGCTA,
ACTCTTTA, Sufijo 19 TA   => Prefijo(s) 43 TACGTGGA,
ACTCTTTA, Sufijo 21 TTTA => Prefijo(s) 51 TTTACTCA,
GCTAATCA, Sufijo 27 ATCA => Prefijo(s) 9 ATCAGTTC,
CTCAAGGC, Sufijo 31 GC   => Prefijo(s) 25 GCTAATCA,
CTCAAGGC, Sufijo 33 AGGC => Prefijo(s) 3 AGGCTACG,
GTTCATAC, Sufijo 37 AC   => Prefijo(s) 19 ACTCTTTA,
GTTCATAC, Sufijo 38 TAC  => Prefijo(s) 44 TACGTGGA,
TACGTGGA, Sufijo 45 TGGA => Prefijo(s) 15 TGGAAGCC,
TTTACTCA, Sufijo 51 CTCA => Prefijo(s) 33 CTCAAGGC,
AGCCGCTA, Sufijo 55 TA   => Prefijo(s) 43 TACGTGGA,
AGCCGCTA, Sufijo 57 GCTA => Prefijo(s) 27 GCTAATCA,
```

Fig. 16. Tabla que resulta de utilizar el procedimiento para buscar identidades al comparar sufijos con prefijos. Obtenida con el DNA blanco y las sondas modelo. En cada renglón de izquierda a derecha: secuencia de la sonda que aporta el sufijo; número de sufijo de acuerdo a la posición en la tabla de prefijos y sufijos; el sufijo; número de prefijo idéntico de acuerdo a la posición en la tabla de prefijos y sufijos; y sonda que aporta el prefijo idéntico.

El efectuar las dos rutinas comparativas arrojó diferentes opciones de resultados, las cuales fueron consideradas para mejorar el diseño del algoritmo reconstructivo. Es a partir de las secuencias de las sondas que comparten afijos (prefijos y sufijos) idénticos que se intentaría la reconstrucción.

5.5. RECONSTRUCCIÓN Y EVALUACIÓN DE SECUENCIAS

Desde hace 25 años el uso de programas computacionales ha estado asociado a la reconstrucción de secuencias de DNA (Larson y Messing 1982). Desde aquel entonces y hasta la actualidad

(Sundquist et al. 2007), una gran cantidad de esos programas se han “alimentado” con datos experimentales obtenidos a través del método de secuenciación aleatoria conocido como ‘shotgun’. Como se comentó anteriormente, este método tan ampliamente utilizado está directamente involucrado en el alto grado de automatización que posee actualmente la secuenciación, sin embargo esto no lo hace infalible. En programas ensambladores para protocolos ‘shotgun’[PHRAP(http://www.phrap.org/phredphrapconsed.html);PCAP (http://seq.cs.iastate.edu);TIGR(http://www.tigr.org/software/assembler);CELERA(http://sourceforge.net/projects/wgs-assembler], no hay solución para el notorio “problema de los repetidos” (secuencias repetidas consecutivas o no), el cual debería resolverse para poder hablar de ensamblajes completos y correctos (Pevzner et al. 2001). Los programas mencionados supuestamente ensamblan genomas completos, lo que es estrictamente imposible, pues están obviando el hecho de que hoy en día no hay un solo ensamblador libre de errores (entendiendo por errores aquellas regiones de las secuencias de los DNA blancos que son reconstruidas incorrectamente). En este sentido el algoritmo de reconstrucción propuesto en este trabajo, difiere en principio del origen de los datos que emplea, pues como se ha mencionado, usa los obtenidos de una rutina de VHX. Además el ensamblaje es a partir de fragmentos (las sondas que hibridaron) mucho más pequeños en longitud.

Por otro lado con respecto a la SBH, se mencionó que fue propuesta como un método prometedor para leer secuencias de DNA en poco tiempo (Bains y Smith 1988; Lysov et al. 1988; Southern 1988; Drmanac et al. 1989). No obstante, debido a dos tipos de errores asociados con la hibridación nucleotídica, la SBH ha sido aplicada menos ampliamente de lo que sus creadores esperaban. Con el primer tipo de error, se observa un conjunto muy pequeño de sondas que hibridan, esto en función de lo que se esperaría por la longitud de los DNA blancos. El segundo tipo de error implica un conjunto demasiado grande de sondas que hibridan en función de lo que podría esperarse. El primero se conoce como falla negativa, y el segundo como falla positiva. Estos errores se deben muy probablemente a condiciones experimentales particulares, tales como la temperatura de alineamiento, o la estructura de los DNA blancos. Gracias a estos fallos es muy difícil ensamblar en un correcto orden los fragmentos. En relación a esto, y al carácter de que la gran mayoría de los proyectos de SBH emplean todas las secuencias posibles para sondas de un tamaño determinado, este trabajo innovó porque en él uso de un conjunto especial de sondas, el cual es una muestra del total posible. Es a partir de este conjunto, de la simulación *in silico* de su interacción con los DNA blancos, y del uso del algoritmo combinatorio, que se pretendió reducir la influencia de las fallas de tipo negativo y positivo. Existen programas como DNA-SPECTRUM (Belyi y Pevzner

1997) que permiten analizar el poder de resolución y los parámetros de la SBH, pero siempre desde la consideración de conjuntos muy grandes de sondas.

Debido a su diseño el empleo de las rutinas de reconstrucción con el DNA blanco y las sondas modelo, tuvieron un funcionamiento y generación de resultados ideales (Fig. 17).

```
**********************************************************************

Reconstruccion empleando las identidades entre "prefijos" y "sufijos":


1> ACTCTTTACTCAAGGCTACG
   ATCAGTTC,TGGAAGCC,GCTAATCA,GTTCATAC,TACGTGGA,AGCCGCTA,

2> ACTCTTTACTCAAGGCTACGTGGAAGCCGCTAATCAGTTC
   GTTCATAC,

3> ACTCTTTACTCAAGGCTACGTGGAAGCC
   ATCAGTTC,GCTAATCA,GTTCATAC,AGCCGCTA,

4> GTTCATACTCTTTA
   AGGCTACG,ATCAGTTC,TGGAAGCC,GCTAATCA,CTCAAGGC,TACGTGGA,TTTACTCA,AGCCGCTA,

>5 CTCAAGGCTAATCA
   AGGCTACG,ATCAGTTC,TGGAAGCC,ACTCTTTA,GTTCATAC,TACGTGGA,TTTACTCA,AGCCGCTA,

>6 ACTCTTTACTCAAGGCTACGTGGAAGCCGCTAATCA
   ATCAGTTC,GTTCATAC,

>7 ACTCTTTACTCAAGGC
   AGGCTACG,ATCAGTTC,TGGAAGCC,GCTAATCA,GTTCATAC,TACGTGGA,AGCCGCTA,

>8 ACTCTTTACTCAAGGCTACGTGGAAGCCGCTAATCAGTTCATAC


>9 GTTCATACTCTTTACGTGGA
   AGGCTACG,ATCAGTTC,TGGAAGCC,GCTAATCA,CTCAAGGC,TTTACTCA,AGCCGCTA,

>10 AGCCGCTACGTGGA
    AGGCTACG,ATCAGTTC,TGGAAGCC,ACTCTTTA,GCTAATCA,CTCAAGGC,GTTCATAC,TTTACTCA,

>11 GTTCATACGTGGA
    AGGCTACG,ATCAGTTC,TGGAAGCC,ACTCTTTA,GCTAATCA,CTCAAGGC,TTTACTCA,AGCCGCTA,

>12 ACTCTTTACTCAAGGCTACGTGGA
    ATCAGTTC,TGGAAGCC,GCTAATCA,GTTCATAC,AGCCGCTA,

>13 ACTCTTTACTCA
    AGGCTACG,ATCAGTTC,TGGAAGCC,GCTAATCA,CTCAAGGC,GTTCATAC,TACGTGGA,AGCCGCTA,
```

```
>14 ACTCTTTACTCAAGGCTACGTGGAAGCCGCTA
    ATCAGTTC,GCTAATCA,GTTCATAC,

**********************************************************************

Reconstruccion empleando las identidades entre "sufijos" y "prefijos":


>1 AGGCTACGTGGAAGCCGCTAATCAGTTCATAC
   ACTCTTTA,CTCAAGGC,TTTACTCA,

>2 ATCAGTTCATAC
   AGGCTACG,TGGAAGCC,ACTCTTTA,GCTAATCA,CTCAAGGC,TACGTGGA,TTTACTCA,AGCCGCTA,

>3 TGGAAGCCGCTAATCAGTTCATAC
   AGGCTACG,ACTCTTTA,CTCAAGGC,TACGTGGA,TTTACTCA,

>4 ACTCTTTACGTGGA
   AGGCTACG,ATCAGTTC,TGGAAGCC,GCTAATCA,CTCAAGGC,GTTCATAC,TTTACTCA,AGCCGCTA,

>5 ACTCTTTACTCAAGGCTACGTGGAAGCCGCTAATCAGTTCATAC


>6 GCTAATCAGTTCATAC
   AGGCTACG,TGGAAGCC,ACTCTTTA,CTCAAGGC,TACGTGGA,TTTACTCA,AGCCGCTA,

>7 CTCAAGGCTAATCA
   AGGCTACG,ATCAGTTC,TGGAAGCC,ACTCTTTA,GTTCATAC,TACGTGGA,TTTACTCA,AGCCGCTA,

>8 CTCAAGGCTACGTGGAAGCCGCTAATCAGTTCATAC
   ACTCTTTA,TTTACTCA,

>9 GTTCATACTCTTTACGTGGA
   AGGCTACG,ATCAGTTC,TGGAAGCC,GCTAATCA,CTCAAGGC,TTTACTCA,AGCCGCTA,

>10 GTTCATACGTGGA
    AGGCTACG,ATCAGTTC,TGGAAGCC,ACTCTTTA,GCTAATCA,CTCAAGGC,TTTACTCA,AGCCGCTA,

>11 TACGTGGAAGCCGCTAATCAGTTCATAC
    AGGCTACG,ACTCTTTA,CTCAAGGC,TTTACTCA,

>12 TTTACTCAAGGCTACGTGGAAGCCGCTAATCAGTTCATAC
    ACTCTTTA,

>13 AGCCGCTACGTGGA
    AGGCTACG,ATCAGTTC,TGGAAGCC,ACTCTTTA,GCTAATCA,CTCAAGGC,GTTCATAC,TTTACTCA,

>14 AGCCGCTAATCAGTTCATAC
    AGGCTACG,TGGAAGCC,ACTCTTTA,CTCAAGGC,TACGTGGA,TTTACTCA,
```

Fig. 17. Texto de la pantalla que resulta de utilizar las rutinas para reconstrucción de secuencias alternativas. Usando el DNA blanco y las sondas modelo. Se muestran las secuencias alternativas que reconstruyó el programa. Debajo de cada una están las sondas no empleadas en cada reconstrucción.

La salida de las rutinas de reconstrucción mostradas obedece al siguiente formato:

1) En primera instancia y con el encabezado - Reconstrucción empleando las identidades entre "prefijos" y "sufijos":- están todas las reconstrucciones hechas a partir de los resultados de tablas como la ejemplificada con la figura 15, debajo de cada secuencia reconstruida se indican las sondas que no fueron empleadas en la reconstrucción. En el caso del DNA blanco modelo, fueron catorce diferentes reconstrucciones, un número igual al de las sondas cuyos sufijos tuvieron identidad con algún prefijo (Fig. 15). Obsérvese con atención que en este caso la reconstrucción en la octava posición es aquella que empleó todas las sondas que aportaron prefijos. Cabe señalar que una reconstrucción ideal debería emplear todas las sondas que hibridaron positivamente, por lo que mientras el número de sondas no empleadas en la reconstrucción se aproximara más a cero, la asignación de mayor confiabilidad sería para las reconstrucciones con dicho carácter.

Una de las reconstrucciones de la secuencia del DNA blanco modelo se esquematiza con la figura 18, donde puede apreciarse como a partir de la secuencia de la octava sonda mostrada en la figura 15 (GTTCATAC), son usadas las identidades entre prefijos y sufijos hasta alcanzar la reconstrucción total del DNA blanco.

También vale la pena destacar que las ramificaciones en la reconstrucción resultantes de sondas cuyo afijo aportado se identifica con el afijo de dos o más sondas diferentes, fue un factor no considerado en el algoritmo reconstructivo implementado en este trabajo.

Fig. 18. Esquema de la reconstrucción del DNA blanco modelo teniendo como inicio la octava sonda de la figura 15 (GTTCATAC). Obsérvese que la reconstrucción fue a partir de las identidades de prefijos con sufijos.

2) En segundo lugar y con el encabezado -Reconstrucción empleando las identidades entre "sufijos" y "prefijos":- aparecen las reconstrucciones elaboradas con los datos de la tabla que se muestra en la figura 16. También debajo de cada secuencia reconstruida se indican las sondas que no fueron empleadas en cada reconstrucción. Fueron catorce reconstrucciones, al igual que el número de sondas cuyos prefijos tuvieron identidad con algún sufijo. En este caso la quinta reconstrucción fue la ideal, ello a partir del criterio de evaluación mencionado anteriormente (Fig. 19).

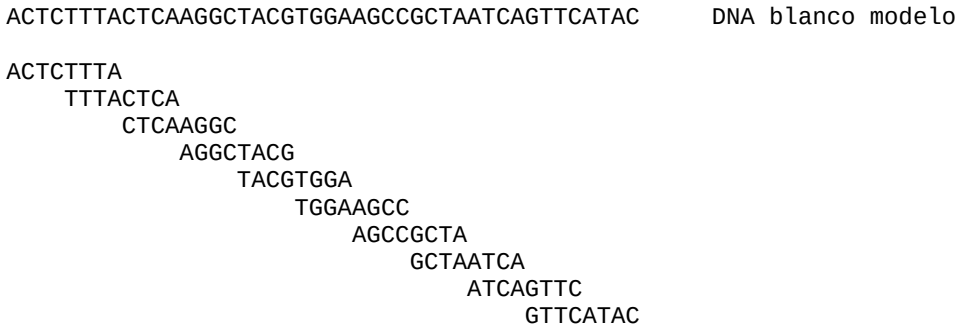

Fig. 19. Esquema de la reconstrucción del DNA blanco modelo teniendo como inicio la quinta sonda de la figura 16 (ACTCTTTA). Obsérvese que la reconstrucción fue a partir de las identidades de sufijos con prefijos.

Con el DNA blanco de 245 nt, a pesar de conocer de antemano que las sondas que hibridaron no eran redundantes (criterio por el que fue seleccionado para analizar la reconstrucción bajo esta consideración), se generaron 136 reconstrucciones parciales posibles a partir de la búsqueda de identidad entre prefijos y sufijos, y un número igual de reconstrucciones parciales posibles empleando la búsqueda de identidad entre sufijos y prefijos. Fue abundante la presencia de reconstrucciones estructuradas a partir de la repetición múltiple de sub-secuencias (‘loops’).

Para el DNA blanco de 1,000 nt proveniente de HPV, empleando la tabla de identidad entre prefijos y sufijos, el número de reconstrucciones posibles fue de 3,459. Con el uso de las identidades entre sufijos y prefijos el número de reconstrucciones posibles fue igual al mencionado. En este caso también fueron excesivamente abundantes las reconstrucciones estructuradas a partir de la repetición múltiple de sub-secuencias.

6. CONCLUSIONES

Las siguientes conclusiones son vertidas a la luz del éxito que para este trabajo representó el análisis de cobertura y superposición, tanto del DNA blanco y las sondas modelo, como de las 68 secuencias de HPV (una sola cadena) y el conjunto de 511 sondas 8-mer.

El algoritmo de reconstrucción que fue implementado usando como referente la VHX del DNA blanco y las sondas modelo, generó una reconstrucción completa de aquél. Claro está que éstos fueron ideales en cuanto a la cobertura, los casos de superposición y la redundancia, pues su diseño obedeció a una simplificación en aras de facilitar la escritura del algoritmo de reconstrucción.

El algoritmo permitió reconstruir sin ambigüedad la secuencia del DNA blanco ideal debido en primera instancia a que las superposición de las sondas modelo fue de tipo ideal, i.e., constante en el número de posiciones superpuestas, y más de una posición superpuesta (caso 4). En el caso de los DNA blancos de HPV la superposición era variable en cuanto el número de posiciones superpuestas, además de que en muchas ocasiones los casos más numerosos eran de los tipos no ideales (casos 1, 2, 3, o 5).

En el caso del DNA blanco modelo la superposición que hicieron las sondas modelo fue total, factor que permitió una

reconstrucción completa de la secuencia. Por otro lado el análisis de cobertura de las 68 secuencias de HPV empleando el conjunto de 511 sondas 8-mer, evidenció que dichas sondas al hibridar (con la permisividad de un apareamiento imperfecto) sólo realizaban coberturas parciales de los DNA blancos, lo que implicaba que en caso de poder obtenerse reconstrucciones confiables, éstas siempre serían parciales.

La ausencia de redundancia en la hibridación de las sondas modelo fue otro carácter que facilitó la reconstrucción ideal del DNA blanco modelo. La redundancia de señales de hibridación no es por si misma un artificio, pues ocurre frecuentemente que una sub-secuencia (en este caso representada por una sonda) se repita a lo largo de la secuencia de un DNA blanco. No obstante, en el caso del diseño del algoritmo esto implicó una consideración que no fue posible tener en cuenta.

Entre y dentro de muchas de las reconstrucciones se observaron sub-secuencias repetidas, en principio porque como punto de inicio llegó a emplearse una misma sonda, debido a que en los listados de identidades ésta aportaba más de un afijo (prefijo o sufijo según la tabla comparativa empleada). En segunda instancia, y es sumamente importante señalarlo, fue abundante la presencia de reconstrucciones estructuradas a partir de la repetición múltiple de sub-secuencias. La explicación de esta repetición la encontramos en sondas donde el prefijo y el

sufijo de una misma son idénticos, e.g. CTACACTA. Este tipo de sondas ocasionaron que el algoritmo de reconstrucción entrará en ciclos que ocasionaron reconstrucciones con una sola sub-secuencia repetida muchas veces. Los posibles conjuntos de sondas que se empleen en el futuro deberán carecer de esta característica en su secuencia.

El algoritmo que se implementó generó un gran número de reconstrucciones para las secuencias de HPV. Este gran número incluye un exceso de reconstrucciones estructuradas a partir de la repetición múltiple de sub-secuencias ocasionadas por la acusa anteriormente expuesta. El proceso de inclusión reiterada de una sub-secuencia se repetía hasta un número igual al del total de secuencias en los listados de identidades. Sin embargo, aún obviando estas reconstrucciones, el número sigue siendo muy elevado, además de que pudo observarse que las reconstrucciones restantes son muy pequeñas en relación a la longitud total de las secuencias que se pretendían reconstruir. Existen dos explicaciones para esto, en primer lugar el algoritmo de reconstrucción no permite seguir todos los posibles trayectos de reconstrucción, y en segundo lugar sólo considera superposiciones perfectas, lo que dejó excluidas superposiciones imperfectas que pudieron formar parte de reconstrucciones de mayor longitud y confiabilidad.

Existen algoritmos recursivos que permitirían seguir todos y cada uno de los trayectos de reconstrucción para un listado de identidades determinado. Sin embargo, la cantidad de reconstrucciones posibles crecería en número exponencialmente, lo que ahondaría el problema para discriminar entre todas y llegar a la original.

Resulta importante reiterar que este proyecto pretendió ser sólo una evaluación teórica de una SBH con el uso de un grupo de sondas. No fueron hechas consideraciones experimentales relevantísimas en una hibridación real, pues por motivos de simplificación serían abordadas posteriormente en caso de que este proyecto resultara en reconstrucciones parciales al menos confiables.

Hasta el momento se han recapitulado las dificultades asociadas con la reconstrucción que efectuó el algoritmo implementado en esta investigación, pero a continuación se indican a manera de perspectivas, propuestas que podrían redundar en resultados reconstructivos mucho más específicos y confiables.

7. PERSPECTIVAS

El diseño de otros conjuntos de sondas con características de longitud y composición diferentes, probablemente generaría resultados de VH que complementarían los obtenidos en este estudio con el conjunto de 511 sondas 8-mer. Los nuevos conjuntos podrían cubrir las regiones descubiertas por el conjunto mencionado, lo que en principio eliminaría el perjuicio de la cobertura parcial. También podrían buscarse consensos generados por el alineamiento de las reconstrucciones provenientes de diferentes VHX. Una posibilidad más para incrementar el porcentaje de cobertura, sería efectuar la hibridación con la otra cadena de los DNA blancos.

Otro enfoque que abatiría en cierta medida los problemas señalados anteriormente, sería mejorar el algoritmo de reconstrucción de manera que puedan incluirse en las reconstrucciones las identidades producidas por superposiciones imperfectas. Otra mejora sería diseñar subrutinas que discriminaran las secuencias formadas la inserción cíclica de sub-secuencias, y permitieran sólo la consideración del resto.

Como se mencionó antes una de las limitaciones más severas de la SBH, además de las secuencias repetidas, ha sido que cuando el estudio experimental se hace con un conjunto de sondas con todas las secuencias posibles de una misma longitud, estas tienen

un espectro muy amplio de T_m. Por ejemplo, para sondas 8-mer se podría estimar que desde 2 X 8 = 16 (para sondas que sólo contengan A y T) hasta 4 X 8 = 32 (para sondas formadas exclusivamente por G y C) estaría ubicado el rango de T_m., i.e., 16°C de variación (32 – 16). Y debido a que para lograr que hibriden las sondas con menor T_m se debe incubar a una temperatura menor (al menos en 3°C) a la de su T_m, en tales condiciones las sondas con alto valor de T_m forman híbridos imperfectos, lo que genera alternativas de reconstrucción que crecen erróneamente por cada imperfección permitida. Por todo lo anterior, se propone que otra posibilidad de mejoría en las reconstrucciones la darán los resultados de hibridaciones virtuales con restricciones termodinámicas, que además usen conjuntos especiales de sondas cuyos rangos de T_m sean estrechos, lo que implicaría una aproximación más estrecha con al contexto experimental de la SBH.

Recientemente se han realizado investigaciones asociadas con los parámetros termodinámicos que sustentan la VH (Romero-Hernández et al. 2007), lo que inminentemente enriquecerá esta herramienta. En este sentido pueden vislumbrarse elementos extra muy importantes para el mejoramiento de los resultados obtenidos durante este proyecto.

Vale la pena destacar que un importante uso que podrá darse a la rutina que analiza cobertura, es en el contexto de la

obtención de la huella genómica como método para identificación de especies. Su utilización para conocer las capacidades de cobertura de un genoma empleando un conjunto de sondas (las que formarían parte del arreglo que pretendiera detectar la huella genómica), representaría posibilidades de mejorar el diseño de los prospectos para la identificación.

REFERENCIAS

Bains W., Smith G.C. (1988) A novel method for nucleic acid sequence determination. *J. Theor. Biol.* **135**:303-307.

Beattie K.L., Doktycz M.J., Mendez-Tenorio A., Maldonado-Rodriguez R., Guerra-Trejo A. (2003) Solicitud de Patente en EUA 410040.

Beaucage S.L., Caruthers M.H. (1981) Deoxynucleotide phosphoramidites — a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett*. **22**(20):1859-1862.

Beebe T.P., Wilson T.E., Ogletree D.F., Katz J.E., Balhorn R., Salmeron M.B., Siekhaus W.J. (1989) Direct observation of native DNA structures with the scanning tunneling microscope. *Science* **243**(4889):370-372.

Belcaid M., Bergeron A., Chateau A., Chauve C., Gingras Y., Poisson G., Vendette M. (2007) Exploring Genome Rearrangements using Virtual Hybridization. En: Sankoff D., Wang L., Chin F. (eds.) Proceedings of 5th Asia-Pacific Bioinformatics Conference, APBC 2007, 15-17 January 2007, Hong Kong, China en: Advances in

Bioinformatics and Computational Biology 5, Imperial College Press, Reino Unido, p.205-214.

Belyi I., Pevzner P.A. (1997) Software for DNA sequencing by hybridization. *Comput. Appl. Biosci.* **13**(2):205-210.

Bennett S.T., Barnes C., Cox A., Davies L., Brown C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**:373-382.

Borland Software Corporation. (2002) Borland Delphi 6.0. Software. http://www.borland.com.

Casique-Almazán, J., García-Chéquer A.J., Maldonado-Rodríguez R., Beattie K.J., Méndez-Tenorio A. (2007) High Sensibility Viroid Identification and Classification by Virtual Hybridization Generated Fingerprints Using the Universal Fingerprinting Chip (UFC). *Front. Biosci.* (Enviado para su publicación).

Church G., Deamer D.W., Branton D., Baldarelli R., Kasianowicz J. (1995-1998) Aplicación de Patente en EUA 5,795,782.

Cowie S., Drmanac S., Swanson D., Delgrosso K., Huang S., du Sart D., Drmanac R., Surrey S., Fortina P. (2004) Identification of APC gene mutations in colorectal cancer using universal microarray-based combinatorial sequencing-by-hybridization. *Hum. Mutat.* **24**(3):261-271.

Doty P., Marmur J., Eigner J., Schildkraut C. (1960) Strand separation and specific recombination in deoxyribonucleic acids: physical chemical studies. *Proc. Natl. Acad. Sci. USA* **46**(4):461-476.

Drmanac R., Crkvenjakov R. (1987) Aplicación de Patente Yugoslava 570/87. (1993) Otorgada posteriormente como Drmanac, R. Method of sequencing of genomes by hybridization with oligonucleotide probes. Aplicación de Patente en EUA PCT/US94/10945.

Drmanac R., Labat I., Brukner I., Crkvenjakov R. (1989) Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* **4**(2):114-128.

Drmanac R., Drmanac S., Chui G., Diaz R., Hou A., Jin H., Jin P., Kwon S., Lacy S., Moeur B., Shafto J., Swanson D., Ukrainczyk T., Xu C., Little D. (2002) Sequencing by

hybridization (SBH): advantages, achievements, and opportunities. *Adv. Biochem. Eng. Biotechnol.* **77**:75-101.

Lindsay S.M., Philipp M. (1991) Can the scanning tunneling microscope sequence DNA? *Genet. Anal. Tech. Appl.* **8**(1):8-13.

Lysov I. u P., Florent'ev V.L., Khorlin A.A., Khrapko K.R., Shik V.V. (1988) [Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. A new method] *Dokl. Akad. Nauk. SSSR* **303**(6):1508-1511. Ruso.

Macevicz, S.C. (1989) Aplicación de Patente Internacional PCT/US89/04741.

Maldonado-Rodríguez R., Méndez-Tenorio A., Jaimes-Díaz H., Flores-Cortés P., Guerra-Trejo A., Reyes-Rosales E., Espinosa-Lara M., Santiago-Hernández J.C., Reyes-López M.A., and Beattie K.L. (2007) Virtual Hybridization: A suite of bioinformatic tools to design DNA microarrays, predict genomic fingerprints and perform bacterial identification. *Cancer Biol. Ther.* (Enviado para su publicación).

Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., Dewell

S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L., Jarvie T.P., Jirage K.B., Kim J.B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F., Rothberg J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376-380.

Maxam A.M., Gilbert W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74**:560-564.

Méndez-Tenorio A. (2006) Elaboración de software relevante para el diseño y evaluación de microarreglos de DNA. Tesis de Doctorado, Dpto. de Bioquímica ENCB-IPN, México, D.F.

Méndez-Tenorio A., Flores-Cortés P., Guerra-Trejo A., Jaimes-Díaz H., Reyes-Rosales E., Maldonado-Rodríguez A., Espinosa-Lara M., Maldonado-Rodríguez R., Beattie K.L. (2006) In silico evaluation of a novel DNA chip based fingerprinting

technology for viral identification. *Revista Latinoamericana de Microbiología* **48**(2):56-65.

Larson R., Messing J. (1982) Apple II software for M13 shotgun DNA sequencing. *Nucleic Acids Res.* **10**(1):39-49.

Mullis K.B., Faloona F.A. (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**:335-350.

Pevzner P.A., Lipshutz R.J. (1995) Towards DNA sequencing chips. En: Prívara I., Rovan B., Ružička P. (eds.) Proceedings of the 19th International Symposium, MFCS'94, August 1994, Košice, Slovakia en: Lecture Notes in Computer Science 841, Springer-Verlag, Alemania, p.143-158.

Pevzner P.A., Tang H., Waterman M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**(17):9748-9753.

Prober J.M., Trainor G.L., Dam R.J., Hobbs F.W., Robertson C.W., Zagursky R.J., Cocuzza A.J., Jensen M.A., Baumeister K. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**(4825):336-341.

Poustka A., Lehrach H. (1986) Jumping libraries and linking libraries. The next generation of molecular tools in mammalian genetics. *Trends Genet.* **2**:174-179.

Reyes-López M.A., Méndez-Tenorio A., Maldonado-Rodríguez R., Doktycz M.J., Fleming J.T., Beattie K.L. (2003) Fingerprinting of prokaryotic 16S rRNA genes using oligodeoxyribonucleotide microarrays and virtual hybridization. *Nucleic Acids Res.* **31**(2):779-789.

Romero-Hernández A., Cruz-Laguna E., Méndez-Tenorio A., Benight A.S., Beattie K.L., Maldonado-Rodríguez R. (2007) Improving Virtual Hybridization by the Calorimetric Evaluation of the Thermodynamic Stability of Some DNA Bimolecular Interactions. *Cell Biochem. Funct.* (Enviado para su publicación).

Ronaghi M., Uhlén M., Nyrén P. (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**(5375):363–365.

Sanger F., Nicklen S., Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**:5463-5467.

Sanger F., Coulson A.R., Barrell B.G., Smith A.J., Roe B.A. (1980) Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* **143**(2):161-178.

Shendure J., Porreca G.J., Reppas N.B., Lin X., McCutcheon J.P., Rosenbaum A.M., Wang M.D., Zhang K., Mitra R.D., Church G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**(5741):1728-1732.

Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B., Hood L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* **321**(6071):674-679.

Southern, E. (1988) Aplicación de Patente Internacional PCT/GB89/00460.

Sundquist A., Ronaghi M., Tang H., Pevzner P., Batzoglou S. (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* **2**(5):e484. doi:10.1371/journal.pone.0000484

Wallace R.B., Shaffer J., Murphy R.F., Bonner J., Hirose T., Itakura K. (1979) Hybridization of synthetic

oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res.* **6**(11):3543-3557.

Watson J.D., Crick F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**(4356):737-738.

Woolley A.T., Guillemette C., Li Cheung C., Housman D.E., Lieber C.M. (2000) Direct haplotyping of kilobase-size DNA using carbon nanotube probes. *Nat. Biotechnol.* **18**(7):760-763.

Yan H., Xu B. (2006) Towards rapid DNA sequencing: detecting single-stranded DNA with a solid-state nanopore. *Small* **2**(3):310-312.