Análisis comparativo y exhaustivo de secuencias de longitud 12 nucleótidos en cromosomas humanos.

Vicente Arnau e Ignacio Marín *.

Resumen—**Hemos desarrollado un nuevo algoritmo que permite analizar la frecuencia de aparición de cadenas de 12 nucleótidos en secuencias de DNA. Es lo suficientemente rápido como para ser utilizado a escala genómica sobre un ordenador personal. En este trabajo y a modo de ejemplo, realizamos el estudio de la frecuencia de aparición de todas las secuencias de 12 nucleótidos presentes en los cromosomas humanos 21 y 22, cada uno de ellos con una longitud aproximada de 33 millones de nucleótidos. La comparación entre los resultados obtenidos para cada cromosoma nos permite localizar secuencias específicas de cada uno de ellos. Todo este proceso se realiza en un tiempo inferior a 2 minutos sobre un PC a 1,7 GHz, con una tasa de análisis de 45 millones cadenas de longitud 12 nucleótidos por minuto.**

Para la localización de las secuencias específicas de los cromosomas 21 y 22, seleccionamos aquellas cadenas de 12 nucleótidos que han aparecido con una frecuencias superior a 40 en uno de ellos y no están presentes en el otro cromosoma. Al final de este trabajo presentaremos algunas futuras aplicaciones de nuestro algoritmo, como son el análisis de genes y el ensamblaje del DNA.

Palabras clave—**Algorítmica, análisis de DNA.**

I. Introducción

En el análisis del genoma son utilizados generalmente algoritmos que podríamos calificar de fuerza bruta, haciendo uso de grandes supercomputadores y procesamiento paralelo de la información. Sin embargo, el aumento de prestaciones de los actuales ordenadores personales, que han alcanzado ya frecuencias de funcionamiento de 3GHz, discos de más de 100 Gbytes y memoria principal de 3 Gbytes, permite abrir nuevas posibilidades de uso de estos computadores. Y si a ello añadimos el diseño de nuevos algoritmos de análisis que optimicen el tiempo de análisis de los datos, podemos llegar a una solución de compromiso aceptable sobre este tipo de ordenadores de bajo coste. Es evidente que aun así, para muchos tipos de análisis va ha seguir siendo necesario el uso de grandes y costosos equipos informáticos, sobre los cuales aplicar masivamente paralelismo en el procesamiento de la información si se quiere alcanzar los resultados esperados.

El propósito de este artículo, que es un resumen del trabajo presentado en HiCOMB-2003 [18], es mostrar una nueva aplicación informática, que permite de forma rápida y exhaustiva, determinar el número de "palabras" de 12 nucleótidos presentes en secuencias de DNA de cualquier tamaño, incluido cromosomas enteros de cualquier genoma conocido. Hay mucho escrito sobre estimación y análisis de palabras en DNA (revisados en [1]), pero muchos de ellos se concentran en secuencias cortas. De este modo, hay algunos trabajos que analizan dinucleótidos (para ver un ejemplo reciente, ver [2]), siendo uno de los resultados más significativos encontrar la disminución en frecuencia del dinucleótido CG en genomas de vertebrados, debido a su conversión en CA o TG asociada a la metilación [1]. La composición de oligonucleótidos ha sido usada, a menudo con otros tipos de información, para establecer los promotores de genes o las regiones codificadoras [1]-[7] o detectar lugares que son característicos de regiones reguladoras de los genes [8] [9]. Ello puede ser utilizado para establecer firmas genómicas especie-específicas. Por lo tanto, un procedimiento rápido para detectar todas las palabras de un cierto tamaño puede ser de interés muy general, especialmente si se puede aplicar a secuencias muy largas, como cromosomas completos de cualquier genoma. En este estudio, mostramos la factibilidad de un estudio rápido de palabras de hasta 12 nucleótidos con un ordenador personal. Como modelo para probar nuestro procedimiento, mostramos los resultados de la comparación de los cromosomas humanos 21 y 22. Significativamente, la comparación de estos dos cromosomas, cada uno de ellos con aproximadamente 33 Megabases (33 Mb, 33 millones de nucleotidos) de tamaño, puede ser realizada en unos pocos minutos sobre un ordenador personal estándar.

En la siguiente sección, detallaremos el nuevo algoritmo, mostrando sus propiedades generales. En la sección 3, describiremos y validaremos los resultados de cuando el método es aplicado a un caso real: la búsqueda en cromosomas humanos completos de singularidades, es decir, de secuencias cromosoma-específicas . En la sección 4 mostraremos algunas conclusiones sobre el potencial de este método.

II. El algoritmo rápido de búsqueda exhaustiva de secuencias de 12 nucleotidos de longitud a escala genómica.

Vamos a explicar a continuación el funcionamiento básico de nuestro algoritmo. Como secuencia modelo, usaremos los cromosomas humanos 21 y 22. Estos dos cromosomas han sido elegidos porque están

Departamento de Informática y *Departamento de Genética, Universidad de Valencia, Campus de Burjassot. C/Doctor Moliner 50. 46100 Burjassot (Valencia). Correo electrónico: Vicente.Arnau@uv.es.

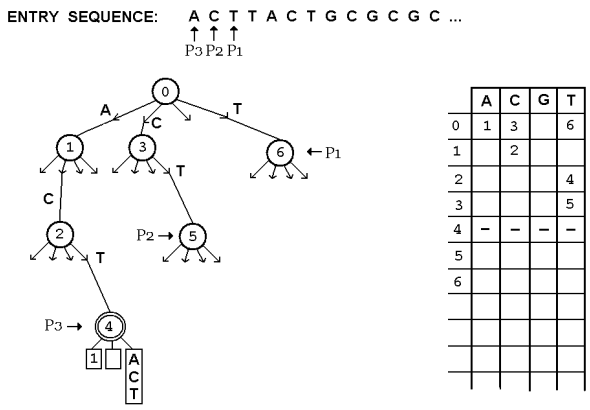
completamente secuenciados y son los cromosomas humanos mejor estudiados en términos de estructura, numero y localización de los genes, ADN repetitivo, y otras interesantes características (ver [11]-[14]).

Hagamos primero unas consideraciones sobre la complejidad del problema. Por una parte, aunque los cromosomas 21 y 22 son los cromosomas humanos más pequeños, su tamaño es considerable, pues cada uno de ellos posee 33 millones de nucleótidos (33 Mb). Son mucho más grandes que el tamaño de genomas completos de otros eucariotas, como la secuencia completa y extensamente analizada de la levadura *Saccharomyces cerevisiae* (alrededor de 12 Mb) y bastantes veces mayor que el tamaño de algunos procariotas como la *Escherichia coli*, la bacteria mejor analizada, con un tamaño de genoma de 5 Mb. Por lo tanto, la comparación entre estos dos cromosomas no deja de ser un buen test a escala genómica. Por otra parte, es sabido que hay 4 nucleótidos diferentes y por lo tanto el número total de secuencias diferentes de 12 nucleótidos es 4^{12} , es decir alrededor de 16.8 millones. Consideremos este número, es evidente que cualquier algoritmo de búsqueda exhaustivo basado en la lectura secuencial y adscripción de cada una de las palabras encontradas a una de estas 16.8 millones de posibles alternativas, sería demasiado lento para ser utilizado.

Algunas estrategias diferentes de análisis de cadenas podemos encontrarlas en [15][16], pero nosotros vamos a utilizar un algoritmo distinto que pasamos a detallar. Nuestro algoritmo lee los nucleótidos almacenados en un fichero de forma secuencial, una a uno, y para cada lectura realiza una serie de acciones sobre una estructura de datos de tipo árbol. Este árbol posee 12 niveles, y en su nivel inferior se encuentran los nodos correspondientes a las cadenas de 12 nucleótidos reconocidas. Su construcción es dinámica, pues en el instante inicial solo tenemos el nodo raíz en el nivel superior. Este nodo posee 4 punteros, que corresponden a los 4 posibles nucleótidos que pueden ser leídos, a saber: Adenina, Citosina, Guanina y Timina, que se abrevian por sus iniciales (A, C, G y T). Son las 4 bases nitrogenadas que forman el ADN. Esta estructura de nodo se repite para los 11 primeros niveles del árbol, pero los nodos del nivel inferior son distintos, deben almacenar la cadena reconocida, así como la frecuencia de aparición de dicha cadena en cada una de las secuencias estudiada. Para nuestro caso, bastará con dos índices pues el estudio lo hemos centrado en la comparación entre dos secuencias de ADN correspondientes a los dos cromosomas seleccionados.

Haremos uso de una serie de punteros, uno por nivel, para tener siempre actualizada la información de cual es la última cadena de 12, 11, 10, ..., 1 nucleótidos leídos. Esto nos permitirá que cada vez que leamos un nuevo nucleótido de una secuencia de entrada, tengamos de forma muy rápida reconocida cual es la cadena de 12 nucleótidos que acabamos de reconocer y además tener actualizada la información de las subcadenas de longitud inferior.

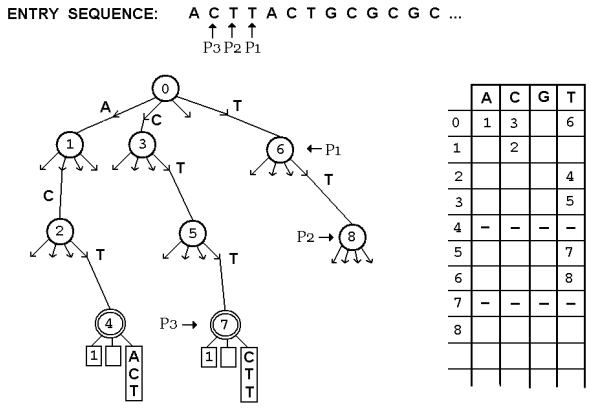
Para ilustrar mejor todo el proceso, mostraremos un ejemplo de funcionamiento de un reconocedor de cadenas de longitud 3 nucleótidos. Sea pues una cadena de entrada como la que aparece en la figura 1 en su parte superior. Partiendo de un nodo raíz, tras la lectura de los tres primeros nucleótidos tendremos ya una estructura de árbol creada como la que aparece en esta figura. Tendremos ahora 3 punteros, P1 es el puntero de primer nivel y nos informará de cual es el último nucleótido leído, P2 es el puntero de segundo nivel y apuntará a un nodo que nos informará de cuales han sido los dos últimos nucleótidos leídos, en este caso CT. Y finalmente P3, apunta a un nodo final o de nivel 3, que es un nodo distinto, pues de él ya no cuelgan más nodos, sino información acerca de cual es la cadena de 3 nucleótidos leída y la frecuencia de aparición de esta cadena en la secuencia de entrada.



	A	C	G	T
0	1	3		6
1		2		
2				4
3				5
4	–	–	–	–
5				
6				

Fig. 1. Reconocimiento de cadenas de 3 nucleótidos. La estructura de árbol mostrada es implementada haciendo uso de una tabla.

Cuando leemos un nuevo nucleótido, como puede observarse en la figura 2, toda la información de los punteros es actualizada, pero con un orden. Los punteros de cada nivel cambian de nodo a partir de la información del puntero de nivel inmediatamente inferior y del nuevo nucleótido leído. Por lo tanto la actualización es secuencial y se realiza de abajo a arriba en el árbol. La comparación de la figura 1 con la figura 2 mostrará este funcionamiento.



	A	C	G	T
0	1	3		6
1		2		
2				4
3				5
4	–	–	–	–
5				7
6				8
7	–	–	–	–
8				

Fig. 2. Actualización de los punteros de nivel con la lectura de un nuevo nucleótido.

En el árbol de nodos, sólo han sido creados aquellos nodos que son utilizados, aunque inicialmente se reverve toda la memoria necesaria para el árbol completo.

Ahora extenderemos este ejemplo a la longitud de cadena que nos ocupa que es de 12 nucleótidos. Para ello el árbol deberá tener 12 niveles y usaremos 12 punteros, aunque en realidad sólo usamos 11 punteros, pues no es necesario en esta estructura de árbol tener un puntero al último nivel formado por nodos finales. En la figura 3 puede observarse como los 11 punteros usados permiten tener actualizada la información de todas las subcadenas de longitud inferior a 12 reconocidas.

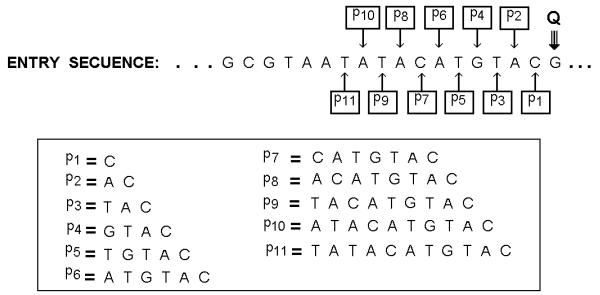


Fig. 3. Uso de los punteros de nivel para el reconocimiento de cadenas de longitud 12.

Cuando el puntero Q que se muestra en la figura 3 nos proporciona un nuevo nucleótido en una secuencia de entrada dada, a partir del puntero P11 y este nuevo nucleótido, obtenemos una nueva cadena de 12 nucleótidos. Además, la frecuencia de aparición de esta cadena en su nodo final correspondiente será actualizada. Los 11 punteros actualizan su posición en el árbol, P11 a partir de P10, luego P10 a partir de P9, y así sucesivamente hasta P1, que apuntará al último nucleótido leído. Estas actualizaciones se realizan con cada nuevo nucleótido leído, con lo cual, al acabar de leer la secuencia de entrada, ya tenemos construido el árbol con todas las cadenas de 12 nucleótidos existentes y con su frecuencia de aparición.

El árbol se construye de forma dinámica, si en un momento dado un nodo no existe se crea, y si ya existe solo es referenciado. Es importante resaltar los dos tipos de nodos. En la figura 4 puede apreciarse este detalle.

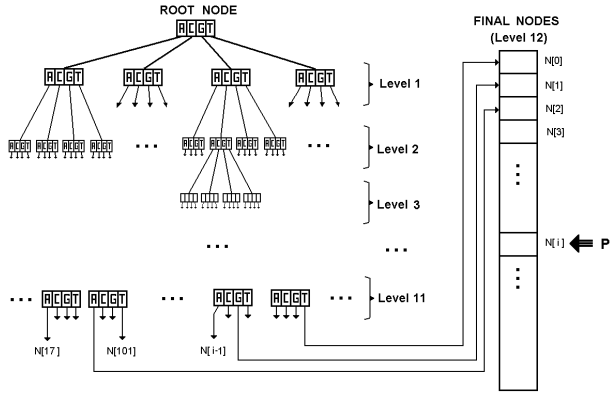


Fig. 4. Árbol dinámico y niveles. El puntero P referenciando la primera posición libre en la tabla de nodos finales.

Los nodos finales son diferentes a los nodos de los 11 primeros niveles. Se crean de forma consecutiva y son accedidos desde los nodos del nivel 11, cuyos punteros apuntan hacia 4 posibles nodos finales. Un puntero P nos informa de cual es la siguiente posición libre en el array de nodos finales. Cuando las primeras 12 bases de la primera secuencia de entrada son leídas, se crea el nodo final N[0] con la información de la palabra leída y su frecuencia de aparición, en este caso toma el valor 1. Al leer el siguiente nucleótido, se crea una nueva secuencia que se almacenará en N[1]. Siguiendo este proceso, si un nuevo nucleótido leído nos conduce a una palabra no existente, un nuevo nodo final será creado a continuación del último, y si esta palabra de 12 nucleótidos ya existe, solo se incrementará su contador de frecuencia. En la figura 5 puede observarse este funcionamiento con una secuencia dada de entrada.

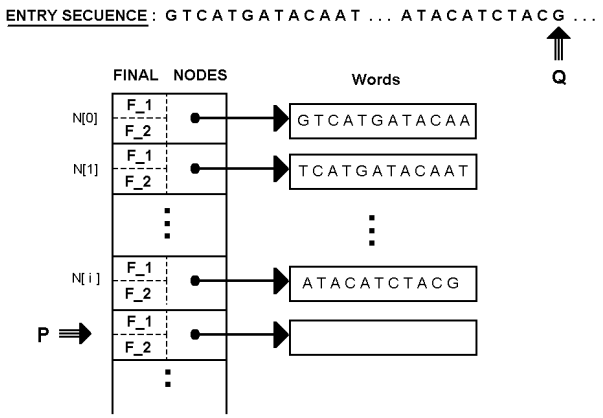


Fig. 5. La figura muestra como los nodos finales son creados de forma consecutiva según aparecen nuevas palabras de 12 nucleótidos.

Tras la lectura del primer cromosoma el árbol no está completo. En él solo existen los nodos necesarios para informarnos de que palabras de 12 nucleótidos han sido encontradas. Además, aunque no lo usemos para este estudio, también nos informa de cuantas palabras distintas de 11, 10, 9, ...,1 nucleótido han aparecido. Así pues, como solo nos interesan las palabras de 12 nucleótidos, solo los nodos finales poseen un contador que llamamos F_1 con la frecuencia de aparición de cada una de las palabras de 12 nucleótidos encontradas.

Cuando leamos a continuación el segundo cromosoma, utilizaremos el mismo árbol de soluciones generado por el primer cromosoma y el puntero P nos seguirá informando de cual es la posición del primer nodo final vacío. De esta forma, cada vez que reconocemos una palabra de 12 nucleótidos en la secuencia del segundo cromosoma, si esta palabra ya existe en el primer cromosoma, en su nodo final correspondiente incrementaremos ahora el contador F_2, si no, pueden ocurrir dos cosas: que sea la primera vez que aparece en el segundo cromosoma y entonces crearemos en la posición apuntado por P un nuevo nodo final, inicializando F_2 a 1, o que exista sólo en el segundo cromosoma, con la cual ya ha sido creada y solo tenemos que incrementar su contador F_2.

Al final del proceso, tras la lectura de las secuencias de los dos cromosomas, el puntero P nos indicará del número de cadenas distintas de 12 nucleótidos

encontradas entre los dos cromosomas. Y en cada nodo final tendremos dos valores, F_1 y F_2 que nos informarán de las veces que cada cadena ha aparecido en uno y otro cromosoma. En este momento podemos prescindir de la estructura de datos en forma de árbol, liberar la memoria, y quedarnos sólo con la tabla de nodos finales. Utilizando esta tabla de nodos finales será fácil encontrar secuencias únicas para uno de los dos cromosomas, sin más que analizar que palabras de 12 nucleótidos aparecen muchas veces en uno de ellos y ninguna en el otro.

Para nosotros ha sido de vital importancia el hecho de que el algoritmo almacene las palabras de forma consecutiva y en el mismo orden en que aparecen por primera vez. Ha tenido dos consecuencias que pasamos a describir. La primera de ellas es que permite compactar la información de las palabras encontradas en la memoria del computador. Y la segunda es que nos ha permitido detectar secuencias singulares de un mayor tamaño. Por ejemplo, una palabra de 20 nucleótidos que aparece con frecuencia en un cromosoma pero raramente en el otro, puede ser fácilmente detectable. La primera vez que aparece en un cromosoma, el programa fragmentará esta palabra de 20 nucleótidos en un conjunto de 9 palabras de 12 nucleótidos que serán almacenadas consecutivamente en la tabla de nodos finales. Estas 9 palabra son muy difícil que aparezcan de forma espontánea en el resto del cromosoma; en un cromosoma de 33 Mb una cadena de 12 nucleótidos es de esperar que aparezca unas 2 veces de forma casual. Todo esto nos permite encontrar cadenas cromosoma-específicas de tamaño mayor que 12 nucleótidos. Hemos utilizado esta propiedad para realizar el análisis que se muestra en la siguiente sección.

El programa que realiza el algoritmo ha sido desarrollado en lenguaje C. Utiliza múltiples punteros definidos como variables enteras que referencian posiciones en una tabla cartesiana de 4 entradas, realizada sobre números enteros, donde cada entrada corresponde con un de los 4 nucleótidos posibles.

III. Búsqueda de secuencias únicas en los cromosomas humanos 21 y 22.

Como demostración de las características de nuestro programa, mostraremos en esta sección datos finales de la comparación exhaustiva de los cromosomas humanos 21 y 22. Las secuencias de estos cromosomas fueron obtenidas de las páginas web del *National Center for Biothecnology Information* (NCBI) http://www.ncbi.nlm.nih.gov/. Encontramos las palabras de longitud 12 nucleótidos que son cromosoma-específicas, realizando la comparación de los dos cromosomas y teniendo en cuenta las dos posibles orientaciones de lectura de las dobles hélices de los mismos. Es decir, comparamos 1) el cromosoma 21 con el cromosoma 22, 2) cromosoma 21 con cromosoma 22 invertido-complementado, y 3) cromosoma 22 con cromosoma 21 invertido-complementado. Como ejemplo de los resultados de este análisis, la comparación del cromosoma 21 con el 22 genera un total de 11457580 palabras presentes en al menos uno de los dos cromosomas. Esto es un 68.3% de las palabras de longitud 12 nucleótidos posibles. El programa lee y analiza los cromosomas con un ratio de 45 Mb/minuto. Por este motivo, los ficheros de análisis se generan en apenas 2 minutos.

Usamos dos valores de corte, que nosotros llamamos F_SUP y F_INF, para encontrar las secuencias sobre representadas en uno de los dos cromosomas. Las palabras que están presentes en un cromosoma con una frecuencia superior a F_SUP y en el otro cromosoma con una frecuencia inferior a F_INF son seleccionadas y se anotan en un fichero junto con su posición en la tabla de nodos finales. Como ejemplo, podemos ver en la Tabla 1 el resultado de compara los cromosomas 21 y el 22 con F_INF=0 en el cromosoma 22 y F_SUP=50 en el 21. Y también podemos ver en la Tabla 2 el resultado de la comparación de los cromosomas con F_INF=0 en el cromosoma 21 y F_SUP=50 en el 22. Como hemos comentado con anterioridad pueden observarse en esta tablas que hay muchas secuencias de 12 nucleótidos encontradas de forma consecutiva.

TABLA I

PALABRAS ENCONTRADAS CON UNA FRECUENCIA SUPERIOR A 50 EN EL CROMOSOMA 21 Y CON FRECUENCIA IGUAL A 0 EN EL CROMOSOMA 22.

No. de nodo	Palabra	Frec. Chr_21	Frec. Chr_22
2811439	AAGCGCATTCAC	58	0
7124108	GCAGGCGTTTCC	57	0
8389206	AGGCGTTTCCCC	57	0
8389207	GGCGTTTCCCCT	56	0
8824281	GGAAGCGCATTC	51	0
8824282	GAAGCGCATTCA	62	0
9033764	GCGTTTCCCCTT	57	0
9033766	TTACCTGCACCG	56	0
9033767	TACCTGCACCGA	54	0
9033768	CTGCACCGAGCC	54	0
9033787	TCCACGCAGGCG	55	0
9033791	CGCAGGCGTTTC	54	0

TABLA II

PALABRAS ENCONTRADAS CON UNA FRECUENCIA SUPERIOR A 50 EN EL CROMOSOMA 22 Y CON FRECUENCIA IGUAL A 0 EN EL CROMOSOMA 21.

No. de nodo	Palabra	Frec. Chr_21	Frec. Chr_22
9139033	CATCATCGAATG	0	81
9139034	ATCATCGAATGG	0	126
9139045	CGAATGGAATCA	0	160
9139053	TCGAATGGAATC	0	196
9139054	GAATCATCATCG	0	73
9139055	AATCATCATCGA	0	80
9139063	AATCGAATGGAA	0	105
9139076	GGAATCATCGAA	0	54
9139103	GAATCATCGAAT	0	55
9139104	AATCATCGAATG	0	62
9139105	CATCGAATGGAA	0	99
9139106	ATCGAATGGAAT	0	197
9139108	GAATGGAATCGA	0	71
9139109	ATGGAATCGAAT	0	94
9139110	TGGAATCGAATG	0	92
9139111	GGAATCGAATGG	0	85
9153783	CAAGCCAGCCAA	0	172
9167410	CAGATACATTGT	0	60
9314281	CTAACGAGGACG	0	71
9314282	TAACGAGGACGC	0	73
9314295	GGCATCGCTAAC	0	56

9314296	GCATCGCTAACG	0	56
9314297	CATCGCTAACGA	0	66
9314298	ATCGCTAACGAG	0	65
9314299	TCGCTAACGAGG	0	139
9314308	CGCCCAGGGCAT	0	59
9314309	CCCAGGGCATCG	0	66
9314310	CCAGGGCATCGC	0	97
9314322	AACGAGGACGCC	0	109
9314323	ACGAGGACGCCG	0	121
9314324	CGAGGACGCCGC	0	82
9314325	AGGACGCCGCCC	0	99
9314326	GGACGCCGCCCA	0	103
9314327	GACGCCGCCCAG	0	66
9314328	ACGCCGCCCAGG	0	64
9314356	GAGGACGCCGTC	0	54
9314357	AGGACGCCGTCC	0	55
9314358	GGACGCCGTCCA	0	54
9314434	CGCTAACGAGGA	0	91
9314557	GCTAACGAGGAC	0	79
9314566	TGAGGACGCTGT	0	90
9415879	GAGGACGCTGTG	0	65
9415900	CGGTGAGGACGC	0	54
9494059	GGCGTCGCTAAC	0	70
9494060	GCGTCGCTAACG	0	69
9494061	CGTCGCTAACGA	0	73
9494062	GTCGCTAACGAG	0	73
9836715	CCTCCATCTGAC	0	68
10137604	CCAACACAGATA	0	72
10245373	CAAAGGATTCCA	0	72
10513310	CAGTCATACTGA	0	56
10783478	AGTCATACTGAC	0	53
11205601	GTAGGTTCCCCT	0	59
11351944	GTCATACTGACT	0	52
11440776	GAACACTGCTAC	0	85

Para ver si el programa ha funcionado correctamente determinando las palabras que son cromosoma-específicas, realizamos la búsqueda descrita con valores de *F_INF* = 0 y *F_SUP* = 40, y utilizamos el programa de búsqueda BLAST usando la página de búsqueda para similitudes casi exactas, disponible en la página web del NCBI http://www.ncbi.nlm.hih.gov/BLAST/; para encontrar todas las palabras detectadas en el Genoma Humano. Vemos como nuestros resultados están completamente validados, las secuencias que encontramos con nuestro método presentes en uno solo de los dos cromosomas, son también detectadas con el programa BLAST como presentes en el mismo cromosoma y ausentes en el otro. Estos resultados no sólo muestran que nuestros análisis son correctos, además muestran que el método permite encontrar palabras únicas de una mayor longitud. Además, estas palabras pueden ser encontradas e interpretadas biológicamente utilizando el programa BLAST. La Tabla 3, mostrada al final de este trabajo, muestra el resumen de los resultados encontrados, incluyendo numero de acceso en la base de datos del NCBI, la posición en el cromosoma y el significado biológico. En esta tabla, han sido fusionas múltiples palabras encontradas de forma consecutiva.

Las secuencias que aparecen más de 40 veces en un cromosoma y ausentes en el otro son muy raras y muy características. De hecho, podemos ver en la Tabla 3 que todas ellas pueden ser clasificadas en 2 tipos diferentes. O bien, poseen características de repeticiones tandem, que por algún motivo están ausentes en uno de los cromosomas (aunque ellas han sido todas encontradas por BLASTN en otros lugares del Genoma Humano además de en el cromosoma 21 o en el 22). O por otra parte, detectamos secuencias que pertenecen a algunos genes con estructuras altamente repetitivas (como por ejemplo las pertenecientes a los genes que codifican unas proteínas conocidas como mucinas, ver [17]).

IV. Conclusiones

El método descrito en este trabajo, nos permite la determinación exhaustiva de las palabras de 12 nucleótidos en secuencias largas de ADN, haciendo uso de un ordenador personal y en un tiempo muy reducido. Estas palabras son de suficiente longitud como para ser encontradas fácilmente con el programa estándar de búsqueda BLASTN, y en alguna de las bases públicas disponibles.

Entre las aplicaciones más generales de este programa están la búsqueda de singularidades o secuencias únicas en cromosomas o genomas (como se muestra aquí), el análisis preciso del número de veces que algunas secuencias características están presentes en dos moléculas o la caracterización del número y tipos de secuencias repetidas (por ejemplo, hemos realizado el análisis de las secuencias ALU detectadas en estos cromosomas humanos, encontrando resultados que son compatibles con los descritos en [11] y [12]). Tambien lo estamos utilizando para ensamblar secuencias de ADN con buenos resultados. La única limitación de este algoritmo es que no es extensible sobre un PC a cadenas de más de 14 nucleótidos por problemas de memoria.

V. Agradecimientos

Este trabajo ha sido financiado por la CICYT (grant no. TIC2000-1151-C07-04) y por la *Fundació La Caixa* (grant no. 01/080-00):

VI. Referencias

[1] S. Karlin, A. M. Campbell, J. Mrázek. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* 32:185-225 (1998).

[2] A. J. Gentles, S. Karlin. Genome-scale compositional comparisons in eukaryotes. *Genome Res.* 11:540-546 (2001).

[3] E. Uberbacher, R. Mural. Locating protein coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl. Acad. Sci. USA* 388:11261-11265 (1991).

[4] E. Zinder, G. Stormo. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural netwoks. *Nucleic Acids Res.* 21:607-613 (1993).

[5] V. Solovyev, A. Salamov, C. Lawrence. Predicting internal exons by oligonucleotide composition and discriminant analysis of sliceable open reading frames. *Nucleic Acids Res.* 22:5156-5163. (1994).

[6] D. Prestridge. Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* 249:923-932. (1995).

[7] Q. Chen, G. Hertz, G. Stormo. PromFD 1.0: a computer program that predicts eukaryotic pol II promoters using strings and IMD matrices. *Comput. Applic. Biosci.* 13:29-35 (1997).
[8] A. Brazma, I. Jonassen, J. Vilo, E. Ukkonen. Predicting gene regulatory elemetns in silico on a genomic scale. *Genome Res.* 8:1202-1215 (1998).
[9] J. van Helden, B. André, J. Collado-Vives. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281:827-842 (1998).
[10] S. Karlin, J. Mrázek. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* 94:10227-10232 (1997).
[11] I. Dunham, N. Shimizu, B. A. Roe, S. Chissoe, *et al.* The DNA sequence of human chromosome 22. *Nature* 402:489-495 (1999).
[12] M. Hattori, A. Fujiyama, T. D. Taylor, H. Watanabe *et al.* The DNA sequence of human chromosome 21. *Nature* 405:311-319 (2000).
[13] P. Kapranov, S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. A. Fodor, T. R. Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296:916-919 (2002)
[14] Ch. Chen, A. J. Gentles, J. Jurka, S. Karlin. Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* 99:2930-2935 (2002).
[15] P. Baldi, S. Brunak. "Bioinformatics. The Machine Learning Approach", Second Edition. A Bradford Book. The MIT Press. (2001).
[16] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison "Biological sequence analysis. Probabilistic models of proteins and nucleic acids". Cambridge University Press, Cambridge (1998).
[17] J. L. Desseyn, J. P. Aubert, N. Porchet, A. Laine. Evolution of the large secreted gel-forming mucins. *Mol. Biol. Evol.*17:1175-1184 (2000).
[18] V. Arnau, I. Marín. A Fast Algorithm For The Exhaustive Analysis of 12-Nucleotide-Long DNA Sequences. Applications to Human Genomics Evolution.. *HiCOMB 2003, Niza (Francia).* (2003).

Secuencia resumen	Números de acceso	Localización cromosómica	Descripción de las secuencias	Notas adicionales
GCGGAAGCGCATTC	AP000335	21q22	Repeticiones en tándem	CHROM. 21-specific
TCCACGCAGGCGTTTCCCCTT TTACCTGCACCGA	XM_066238	21q22	LOC128934 Gen similar a zinc finger 347	CHROM. 21-specific
CGAATGGAATCGATGG ATGGAATCGAATGGAA GAATCATCATCGAATGGAAT GAATCATCGAAT	AP000543	22q11	Secuencias satélite relacionadas, semejantes a $(CGAAT)_n$ $(AATAG)_n$	CHROM. 22-specific
CATCGCTAACGAGGACGCCGCCCAGGG CATCGCTAACGAGGACGCCGTCCA GAGGTCGCCGCCC CCACGGCGTCGCTAACGAGGTCGC CAGGGCATCGCTA CCAGGGCGTCGCTAA	XM_092883	22q12	Secuencias que pertenecen al gen relacionado con las mucinas LOC164854	CHROM. 22-specific
TGGGCGGCGTCCT	XM_092877 XM_092883	22q11 22q12	Genes relacionados con las mucinas LOC164573 y LOC164854	CHROM. 22-specific Hallado en dos genes relacionados (LOC164854, LOC164573), cercanos en el cromosoma, pero en orientaciones opuestas
TTCCCCTGTGCGT	AL021392	22q13	Repeticiones en tándem	CHROM. 22-specific
GGTTGAAGTCTC	AL078613	Desconocido	Repeticiones en tándem	CHROM. 22-specific
GCGGTGAGGACGCTGTG	XN_066267	22q11	Genes relacionados con las mucinas LOC128983	CHROM. 22-specific