Diseño de un Procesador para el Alineamiento Global de Secuencias de DNA

Martin A. Lozano, Jaime Velasco-Medina

Grupo de Bio-nanoelectrónica
EIEE, Universidad del Valle, A.A. 25360, Cali, Colombia
E-mail: martloza@univalle.edu.co, jvelasco@univalle.edu.co

ABSTRACT

Este artículo presenta el diseño de un procesador para el alineamiento global de secuencias de DNA. El algoritmo implementado en hardware es el algoritmo de Needleman y Wunsch, el cual es basado en usar programación dinámica. El procesador fue diseñado usando captura esquemática y descripción estructural en VHDL, la síntesis y simulación se realizó usando Quartus II versión 5 de Altera, y el diseño fue sintetizado en la Stratix II EP2S15F484C3. Los resultados de las simulaciones muestran que el procesador de DNA presenta un buen desempeño usando poca área. En este caso, 43,044 ns para el alineamiento optimo de dos secuencias de DNA, cada una de 7 bases, y 1674 ALUTs y 54 Registros.

***Palabras claves*:** Alineamiento de secuencias de DNA, programación dinámica, FPGA y VHDL

1. INTRODUCION

Debido al proyecto del Genoma Humano en la actualidad existe una gran cantidad de información sobre la composición de los aminoácidos, por lo tanto se debe recurrir al uso de herramientas bioinformáticas que permitan procesar rápidamente esta información.

En este contexto, el alineamiento de dos secuencias de DNA permite conocer la relación evolutiva, es decir conocer si las dos secuencias de DNA tienen un ancestro común del cual pudieron haber evolucionado por bifurcación de especies o por duplicación de genes, sin embargo el alineamiento de secuencias de DNA se debe formalizar rigurosamente desde el punto de vista matemático. En este caso, el alineamiento óptimo entre dos secuencias de DNA es aquel que hace máxima la suma de las puntuaciones de los residuos alineados de acuerdo con una matriz de substitución dada. En principio, es posible determinar el alineamiento óptimo entre dos secuencias elaborando todos los alineamientos posibles y calculando para cada uno de ellos la suma de las puntuaciones de los residuos alineados. Sin embargo, el problema es muy difícil de resolver cuando existen "gaps", en este caso el número de alineamientos posibles entre las dos secuencias de DNA es muy grande y el cálculo exhaustivo es muy difícil de alcanzar usando los actuales recursos computacionales [1].

Determinar el grado de semejanza entre dos secuencias de DNA requiere elaborar el alineamiento y contar (directa o indirectamente) el número de posiciones equivalentes conservadas. Cuando la relación entre dos secuencias es lejana y exhiben una pobre conservación global elaborar el alineamiento no es trivial, puesto que es difícil (si no imposible) identificar los aminoácidos equivalentes en las dos secuencias, en particular porque residuos individuales e incluso regiones enteras de la proteína pueden haber desaparecido (o haberse insertado) a lo largo de la evolución en una de las dos secuencias. Entonces, varios algoritmos progresivos son utilizados e implementados en software, los cuales son simples, utilizan pocos recursos computacionales, emplean poco tiempo y los resultados presentan un nivel de acierto muy aceptable. Sin embargo, la desventaja es que los resultados generados son de índole heurístico, es decir, los resultados son aproximados.

Teniendo en cuenta las consideraciones anteriores, este trabajo presenta la implementación en hardware de un algoritmo para el Alineamiento Global de Secuencias de DNA, el cual presenta ventajas tales como: tomar y procesar toda la secuencia, mayor velocidad de procesamiento, capacidad de alineamiento basado en el algoritmo de Needleman y Wunsch generando resultados muy precisos, mayor portabilidad y mayor eficiencia.

Este trabajo está organizado de la siguiente forma: En la sección 2 se describen los algoritmos básicos para el alineamiento de secuencias de DNA, en la sección 3 se presenta una descripción funcional del procesador, en la sección 4 se presenta el diseño del procesador para el alineamiento de secuencias de DNA, en la sección 5 se presentan los resultados de simulación. Finalmente, en la sección 6 se presentan las conclusiones y el trabajo futuro.

2. TRABAJOS PREVIOS

En la literatura revisada se encuentran muy pocas implementaciones en hardware de algoritmos para alineamiento de secuencias de DNA. Bin Wang en [2] presentan la implementación de un procesador basado en una arquitectura Cell Matriz usando programación dinámica. En este caso, este trabajo solo presenta la implementación de la matriz de puntuación y la recuperación de las secuencias alineadas no es considerada. Por lo tanto, en [2] no se implementa completamente el algoritmo para el alineamiento de secuencias de DNA y los resultados de simulación presentados son para secuencias de máximo 11 bases de longitud.

En este trabajo presentamos la implementación total de un algoritmo para el alineamiento de secuencias de DNA, es decir, la implementación de la matriz de puntuación y la recuperación de las secuencias para alcanzar el alineamiento optimo.

3. ALGORITMOS DE ALINEAMIENTO DE SECUENCIAS DE DNA

Existe una gran variedad de algoritmos de comparación y parámetros que pueden ser usados para evaluar similitudes en secuencias de proteínas y DNA [3-8]. En este caso, la selección del mejor algoritmo depende del problema a resolver. Dos algoritmos óptimos para calcular los puntajes de similitud han sido descritos en [9] y [10]. El primero es el algoritmo de Needleman-Wunsch (1970), el cual calcula los puntajes de similitud global entre dos secuencias. El segundo es el algoritmo de Smith-Waterman (1981), el cual calcula los puntajes de similitud local.

Los algoritmos que calculan comparaciones locales (encontrar similitudes entre dos secuencias, ignorando diferencias fuera de las regiones más parecidas) son los más apropiados para hacer búsquedas de proteínas y ADN en bases de datos, mientras los algoritmos de comparación global son los más apropiados para construir árboles evolutivos, es decir cuando las homologías se han establecido previamente [1].

4. DISEÑO DE LA ARQUITECTURA DEL PROCESADOR DE DNA

La arquitectura del procesador se basa en la implementación directa del algoritmo propuesto por Needleman y Wunsch para el alineamiento global de secuencias de DNA, donde la entrada de los datos a la matriz de puntuación se realiza en forma paralela. En este caso, el procesador diseñado selecciona de manera autónoma la ruta óptima de todos los posibles valores que genera la técnica de programación dinámica para obtener las secuencias alineadas.

Con el propósito de facilitar el diseño del procesador de DNA, el algoritmo de Needleman y Wunsch es divido en tres niveles, los cuales son mostrados en la Figura 1. En el primer nivel el propósito es calcular los valores de la matriz de puntuación, en el segundo nivel el objetivo es generar las señales que permiten realizar el seguimiento, las cuales son las entradas de la unidad de control y el tercer nivel es el encargado de realizar la inserción o no inserción de los espacios en las secuencias para obtener la alineación optima.

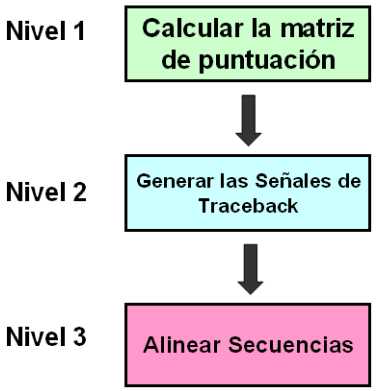

Figura 1. Niveles del algoritmo de Needleman Wunsch para el alineamiento de secuencias de DNA

En la Figura 2 se muestra la arquitectura del procesador de DNA, en la cual se observa la interconexión de los bloques funcionales y la unidad de control. En este caso, el procesador realiza la alineación de las secuencias a partir de la información del bloque funcional que implementa la matriz de puntuación.

El procesamiento de las secuencias se realiza en forma paralela, es decir, una vez obtenida la matriz de puntuación, la cual es implementada usando un circuito combinacional, se generan las señales de control para realizar la alineación de las secuencias.

La implementación de la matriz de puntuación se basa en usar una celda básica, la cual permite calcular los valores para cada celda de la matriz. Las entradas de la matriz son los valores de las secuencias a comparar y las salidas son los valores de puntuación calculados por las celdas básicas.

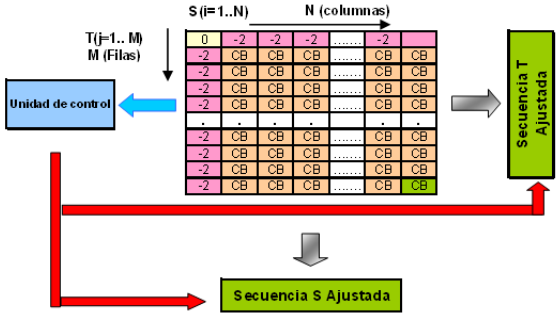

Figura 2. Arquitectura del procesador de DNA

4.1 Diseño de la celda básica

La celda básica (CB) es la encargada de calcular cada uno de los valores de la matriz de puntuación. En la Figura 3 se muestra el diagrama de bloques de la celda básica.

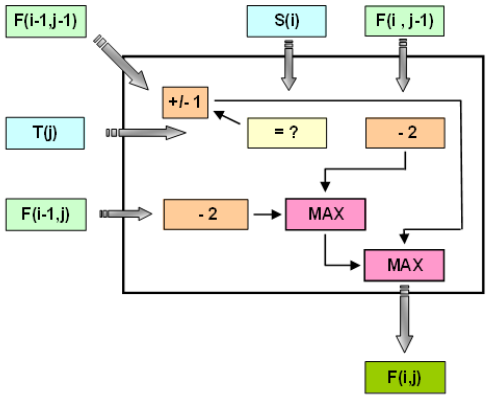

Figura 3. Diagrama de la celda básica

Cada celda básica recibe como entradas los valores de las celdas vecinas (diagonal anterior, fila anterior, columna anterior), y los valores correspondientes a las secuencias a comparar S(*i*) y T(*j*), los cuales están representados por un índice. La salida es el valor óptimo de puntuación según sus entradas, el cual sirve para calcular los valores óptimos de puntuación de las celdas siguientes, hasta completar la matriz (*i*, *j*) que representa las secuencias a comparar.

4.2 Diseño de la matriz de puntuación

Las celdas básicas con los valores de salida "*0*" y "*-2*" de la Figura 2 son constantes y representan las condiciones iniciales para el procesamiento de la matriz de puntuación.
Por ejemplo, la celda básica con valor de salida "*-2*" tiene como entrada el valor de puntuación de "*0*" de la celda anterior (fila), entonces la salida será el resultado de la resta entre 0 y 2 (0-2=-2), el cual será el valor de puntuación de la celda, es decir "*-2*".

En la Figura 4 se muestra el diagrama de bloques de la matriz de puntuación y las celdas básicas son interconectas de tal manera que permiten la comunicación entre las respectivas entradas y salidas de las celdas vecinas.

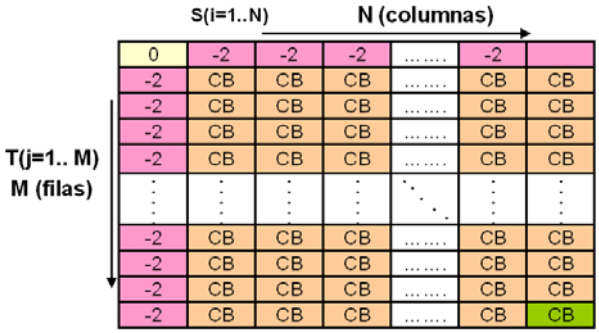

Figura 4. Diagrama general de la matriz de puntuación

4.3 Diseño de la unidad de control

La unidad control es la encargada de generar las señales que controlan los dos bloques encargados de realizar el alineamiento de las secuencias de DNA. Esta unidad tiene como señales de entrada, las señales de seguimiento (calculadas a partir de los valores de puntuación) generadas por la matriz de puntuación. En la Figura 5 se muestra un diagrama de la unidad de control.

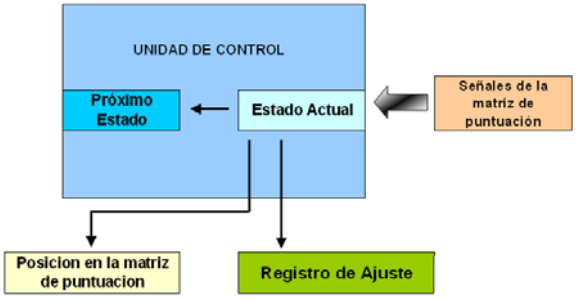

Figura 5. Unidad de control

4.4 Diseño de la unidad de ajuste

Esta unidad es la encargada de realizar las operaciones de inserción o no inserción de espacios en la secuencia de salida de acuerdo a las señales generadas por la unidad control. La salida es la respectiva secuencia con su alineación óptima.

5. RESULTADOS DE SIMULACIÓN

Con el propósito de verificar el funcionamiento del procesador de DNA, varias simulaciones fueron llevadas a cabo. Inicialmente se realizo la verificación de la celda

básica usando algunos valores de la matriz de puntuación mostrados en la Tabla 1.

		G	A	A	C	C
	0	-2	-4	-6	-8	-10
G	-2	1	-1	-3	-5	-7
G	-4	-1	0	-2	-4	-6
A	-6	-3	0	1	-1	-3
C	-8	-5	-2	-1	2	0

Tabla1. Valores de la matriz de puntuación

En la Figura 6 se muestra un ejemplo para describir el funcionamiento de la celda básica, en este caso se calcula el valor de la posición (i=2, j=2) y los resultados de simulación para este ejemplo se muestran en la Figura 7.

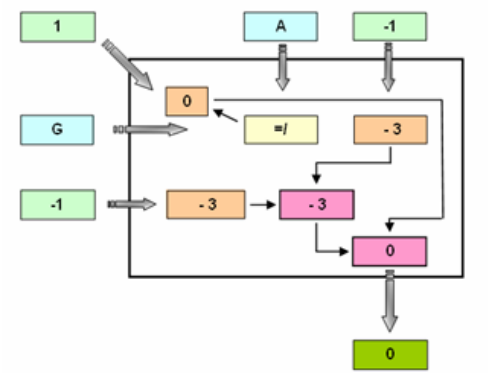

Figura 6. Valor de puntuación de la celda básica para la posición (i=2, j=2)

Posteriormente se realizo la verificación de la matriz de puntuación considerando todos los valores presentados en la Tabla 1 y los resultados de simulación se muestran en la Figura 8. Finalmente se realizaron simulaciones para alinear dos secuencias (cada una de 7 bases de longitud), en el caso de secuencias de 7 bases, el procesador calcula el alineamiento óptimo de las secuencias de DNA en 43,044 ns, el cual cumple con el patrón de seguimiento (traceback) mostrado en la Figura 9, las celdas de color verde indican el recorrido que se debe realizar para extraer las secuencias alineadas, también se muestra el seguimiento con sus inserciones y las secuencias alineadas.

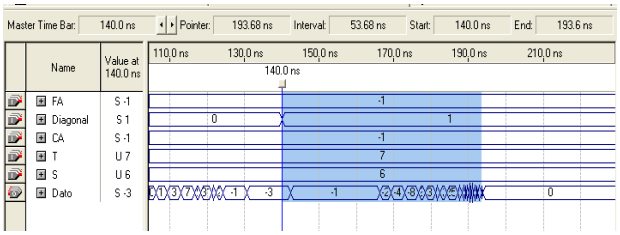

Figura 7. Resultados de simulación de la CB

Desde la Figura 7 se observan los siguientes resultados:

- **FA**: Valor de fila anterior para la posición (i=1, j=2) = "*–1*"
- **CA**: Valor de la columna anterior para la posición (i=2, j=1) = "*-1*"
- **Diagonal**: Valor de puntuación de la celda diagonal anterior para la posición (i=1, j=1) = "*1*"
- **T**: Valor de la secuencia T(j) para la posición (i=2, j=2) = "*7*"
- **S**: Valor de la secuencia S(i) para la posición (i=2, j=2) = "*6*"
- **Dato**: valor de puntuación óptimo de celda para la posición (i=2, j=2) = "*0*"

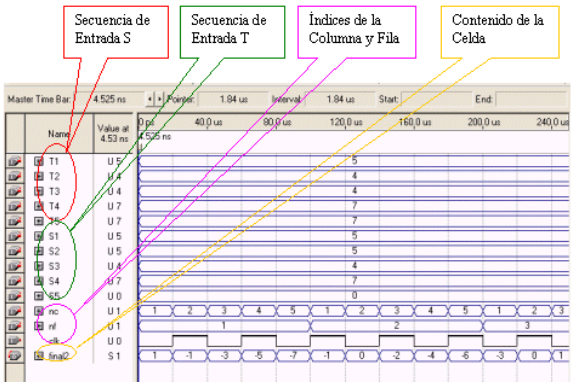

Figura 8a. Simulación de la matriz de puntuación

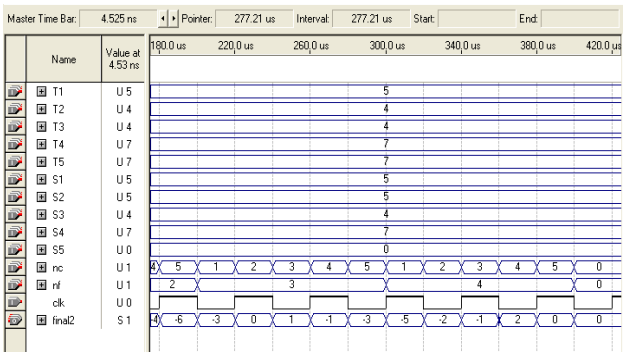

Figura 8b. Simulación de la matriz de puntuación

		G	A	A	C	C
	0	-2	-4	-6	-8	-10
G	-2	1	-1	-3	-5	-7
G	-4	-1	0	-2	-4	-6
A	-6	-3	0	1	-1	-3
C	-8	-5	-2	-1	2	0

Figura 9a. Traceback (Seguimiento)

		G	A	A	C	C
	0	-2	-4	-6	-8	-10
G	-2	1	-1	-3	-5	-7
G	-4	-1	0	-2	-4	-6
A	-6	-3	0	1	-1	-3
-					-	
C	-8	-5	-2	-1	2	0

Figura 9b. Alineamiento óptimo

Aplicando las reglas para la obtención de las secuencias alineadas, la Figura 9b muestra donde se realiza la inserción de espacios. Entonces, el alineamiento óptimo para las secuencias es:

S(i): G A A C C
T(j): G G A - C

Los resultados de simulación para las secuencias alineadas se muestran en la Figura 10.

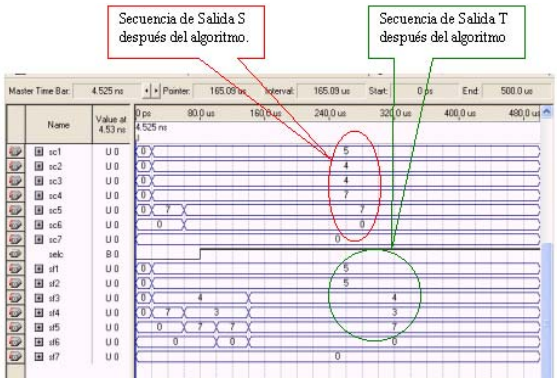

Figura 10. Simulación de las secuencias alineadas

6. CONCLUSIONES Y TRABAJO FUTURO

Este artículo presenta el diseño de un procesador para el alineamiento global de secuencias de DNA. El algoritmo implementado en hardware es el algoritmo de Needleman y Wunsch, el cual es basado en usar programación dinámica. El procesador fue diseñado usando captura esquemática y VHDL estructural, la síntesis y simulación se realizó usando la herramienta Quartus II versión 5 de Altera, y el diseño fue sintetizado en el FPGA Stratix II EP2S15F484C3.

Los resultados de simulación permiten verificar el correcto funcionamiento del procesador de DNA, es decir, el procesador permite realizar el alineamiento optimo de dos secuencias de DNA, en este caso el alineamiento se realizo para secuencias con longitud 7, cada una. Teniendo en cuenta los resultados, el procesador diseñado puede ser usado para estudiar los árboles genealógicos y procesar información para otras aplicaciones en genética.

El trabajo futuro será orientado inicialmente a realizar el alineamiento de secuencias de DNA con una longitud de 64 bases (*versión final del articulo*) y a realizar una aplicación software, la cual sirva de interfaz entre el procesador de DNA y el usuario. Posteriormente, la idea es realizar la implementación en hardware de algoritmos eficientes para el alineamiento de secuencias de DNA y proteínas, y finalmente el trabajo será orientado a implementar en hardware procesadores genómicos y proteómicos de alto desempeño.

7. REFERENCES

[1] Roderic Guigó I Serra, "Bioinformática: La creciente interconexión entre la biología y la informática", Boletín Electrónico de la Sociedad Española de Genética, pp. 4, Julio 2003.

[2] Bin Wang, "Implementation of a Dynamic Programming Algorithm for DNA Sequence Alignment on the Cell Matrix Architecture" http://www.cellmatrix.com/entryway/products/pub/wang2002.pdf , 2002.

[3] "Alineamiento de Secuencias: Programación Dinámica" http://www.ccpg.fq.edu.uy/Cursos/BIOINF101/Slides/Clase_04/Clase04_6.pdf

[4] Z. Luthey-Schulten, "Sequence and Structure Alignment", http://www.ks.uiuc.edu/Training/SumSchool/2004/materials/lectures/Day6/Mon22a.pdf

[5] Marina Alexandersson, "Sequence Analysis – Pairwise Sequence alignment", http://www.fcc.chalmers.se/~**marina**/files/BioI_PairAlign_2005.pdf.

[6] Oswaldo Trelles, "Comparación de Secuencias Biológicas Algoritmia", http://ub.cbm.uam.es/support/courses/Leon2005_arrays/alignment/CompBioS-Alg.pdf

[7] Vladimir Likic, "The Needleman–Wunsch Algorithm for Sequence Alignment", http://www.ludwig.edu.au/course/lectures2005/Likic.pdf

[8] David J. Lipman, Stephen F. Altschul, and John D. Kececioglu, "A tool for Multiple Sequence Alignment", http://www.pnas.org/content/vol86/issue12.

[9] Saul B. Needleman, Christian D. Wunsch, "A General Method Applicable to Search for Similarities in the Amino Acid Sequence of Two Proteins", *J. Mol. Biol.*, 48, pp. 443-453, 1970, http://www.cs.umd.edu/class/spring2003/cmsc838t/papers/needlemanandwunsch1970.pdf

[10] http://www.cecalc.ula.ve/bioinformatica/BIOTUTOR/tutoriales.htm