

Fecha de presentación del Informe: Día Mes Año **1. Datos generales del Proyecto**

Código del proyecto: CI 2766			
Título del proyecto: Aceleración en Hardware del Análisis Multifractal del Genoma Humano			
Facultad o Instituto Académico: Facultad de Ingeniería			
Departamento o Escuela: Escuela de Ingeniería de Sistemas y Computación (EISC)			
Grupo (s) de investigación: Grupo de Bioinformática (EISC) y Grupo de Bionanoelectrónica (Escuela de Ingeniería Eléctrica y Electrónica)			
Entidades: Universidad del Valle			
Palabras claves: Análisis Multifractal, DNA., genoma humano, hardware, FPGA.			
Investigadores	Nombre	Tiempo asignado	Tiempo dedicado
Investigador Principal	Pedro Moreno Tovar ✓	2 h/s	2 h/2
Coinvestigadores	Jaime Velasco Medina ¹ Evaluador ✓	-	2 h/s
Asistente de Investigación	Jorge Duarte Sánchez ✓ E2	40 h/s	40 h/s
Otros participantes	Cris Loaiza ²	-	4 h/s

¹ El profesor Jaime Velasco Medina estuvo vinculado al proyecto de manera indirecta como co-director de tesis de maestría del estudiante Jorge Duarte.

² El estudiante Cris Loaiza estuvo vinculado al proyecto indirectamente con la realización de su tesis de pregrado

2. Resumen ejecutivo:

El análisis multifractal ha permitido estudiar el genoma humano desde una perspectiva diferente proporcionando una manera de medir la variabilidad y estabilidad genética a lo largo del genoma. Estos resultados han sido usados para explicar algunas anomalías genéticas y otras particularidades genéticas con importantes implicaciones en estudios de enfermedades genéticas, diversidad humana, entre otros. El análisis multifractal es usado para caracterizar sistemas dinámicos, procesos y construcciones geométricas asignándoles una función llamada espectro multifractal o espectro de singularidad, el cual es un conjunto de dimensiones fractales obtenidas a diferentes escalas. El análisis multifractal de un genoma es llevado a cabo dividiendo el genoma en secuencias de ADN más pequeñas llamadas fragmentos y obteniendo el espectro multifractal de cada uno de estos usando la representación del juego del caos y el método del conteo de cajas. Este proceso requiere mucho tiempo de cómputo porque implica el procesamiento de una gran cantidad de datos utilizando representación en punto flotante. Para reducir el tiempo de cómputo, en este proyecto nosotros optimizamos el algoritmo que calcula las dimensiones fractales y diseñamos una arquitectura hardware para implementar los algoritmos mejorados utilizando un SoC-FPGA de bajo costo. Estos dispositivos híbridos combinan la flexibilidad y el paralelismo de los FPGAs con la funcionalidad de los sistemas embebidos basados en procesadores ARM de alto desempeño.

La validez, precisión y rendimiento de la arquitectura hardware fueron probadas con un genoma humano completo comparando los resultados con datos obtenidos usando un clúster de 43 núcleos físicos y un PC. Nuestra arquitectura logró una aceleración de 2.34x en comparación al clúster, y de 193.5x en comparación al PC, con errores promedio del orden de 0.01%. Estos resultados demuestran que nuestro diseño puede ser usado para acelerar el análisis multifractal del genoma humano y facilitar estudios en diversos campos de la genómica obteniendo resultados confiables en mucho menor tiempo en comparación a un PC, y a mucho menor costo y en menor tiempo en comparación a un clúster.



Executive Summary

The Multifractal analysis has allowed to study the human genome from a different approach providing a way to measure the genetic variability and stability along the genome. This results have been used to explain some chromosome abnormalities and other genetic particularities with important implications in studies of genetic diseases, human diversity, among others. The multifractal analysis is used to characterize dynamic systems, processes and geometric constructions assigning them a function called multifractal spectrum or singularity spectrum, which is a set of fractal dimensions obtained for different scales. The multifractal analysis of a genome is carried out by dividing the genome in smaller DNA sequences called fragments and obtaining the multifractal spectrum of each fragment using the chaos game representation (CGR) and the box-counting method. This is a time consuming process because it involves the processing of large data sets using floating-point representation. In order to reduce the computation time, in this project we optimized the algorithm that calculates the fractal dimensions and designed a hardware architecture to implement the improved algorithms using a low-cost SoC-FPGA. These hybrid devices combine the flexibility and parallelism of FPGAs with the functionality of embedded systems based on high performance ARM processors.

The validity, accuracy and performance of the hardware architecture was tested with a complete human genome comparing the results with data obtained with a cluster of 43 physical cores and a PC. Our architecture achieved a speed up of 2.34x compared to the cluster, and 193.5x compared to the PC, with average errors in the order of 0.01%. These results demonstrate that our design can be used to accelerate the multifractal analysis of the human genome and facilitate studies in diverse fields of genomics obtaining reliable results in much less time compared to a PC, and at a much lower cost compared to a cluster.

3. Síntesis del proyecto:

En este proyecto se desarrolló un sistema que permite acelerar el análisis multifractal de secuencias de ADN lo cual permite estimar la cantidad de información genética a lo largo de la secuencia analizada e identificar su variabilidad y estabilidad genética. De esta manera es posible encontrar fragmentos del genoma humano susceptibles a perturbaciones genéticas que pueden dar lugar a enfermedades, entre otras aplicaciones.

Objetivo General

Acelerar el análisis multifractal de secuencias de ADN mediante el procesamiento computacional basado en aceleradores de hardware implementados en FPGAs de última generación.

Objetivos Específicos

1. Identificar las rutinas o procesos del algoritmo de análisis multifractal de secuencias de ADN que pueden ser acelerados en hardware.
2. Implementar las rutinas del algoritmo de análisis multifractal de secuencias de ADN en hardware.
3. Diseñar el sistema embebido usando codiseño (software y hardware) para ejecutar el algoritmo de análisis multifractal de secuencias de ADN en una tarjeta de desarrollo basada en FPGAs de última de última generación.
4. Realizar el análisis de una secuencia del genoma humano utilizando el sistema diseñado.
5. Poner a disposición de la comunidad científica nacional e internacional una herramienta eficiente de análisis multifractal de secuencias de ADN para acelerar la investigación en las áreas de la genómica.

Descripción del Diseño

La aceleración del análisis multifractal del genoma humano se logró con la implementación de una arquitectura hardware, denominada Procesador Multifractal, en un SoC-FPGA, el cual

Elaborado por: Vicerrectoría de Investigaciones

combina la flexibilidad y paralelismo de los FPGA convencionales con un sistema de procesamiento hardware (HPS) basado en un procesador ARM. En la Fig. 1 se muestra un diagrama de bloques del sistema completo diseñado. El procesador ARM ejecuta una aplicación en un sistema operativo Linux que controla la transmisión de las secuencias de ADN del genoma humano almacenadas en una tarjeta SD hacia el procesador multifractal. El genoma está codificado en varios archivos de texto en formato FASTA en donde cada cromosoma corresponde a un archivo diferente. La aplicación lee todos los símbolos de los archivos FASTA y los envía a un buffer de datos (RAM), luego los símbolos son enviados al procesador multifractal por DMA (direct memory access) para ser procesados.

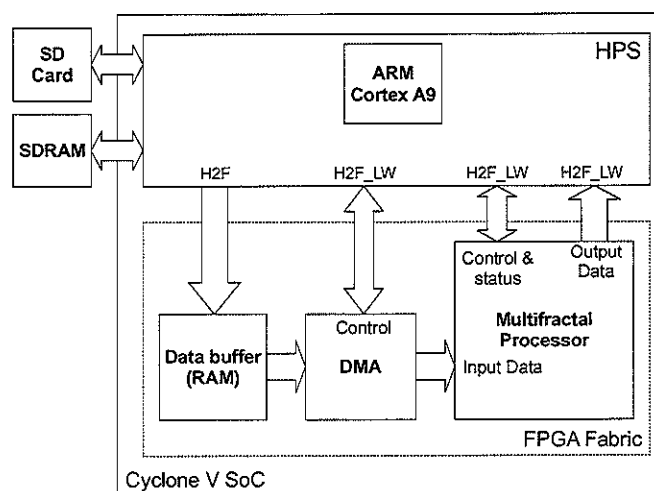


Fig. 1. Diagrama de bloques del sistema diseñado para calcular las dimensiones fractales

En la Fig. 2 se muestra el diagrama de bloques del procesador multifractal diseñado. Este tiene cuatro bloques principales: CGR, box-counting, funciones de partición y regresión lineal. Cada uno de estos bloques ejecuta uno de los sub-procesos del algoritmo que calcula las dimensiones fractales.

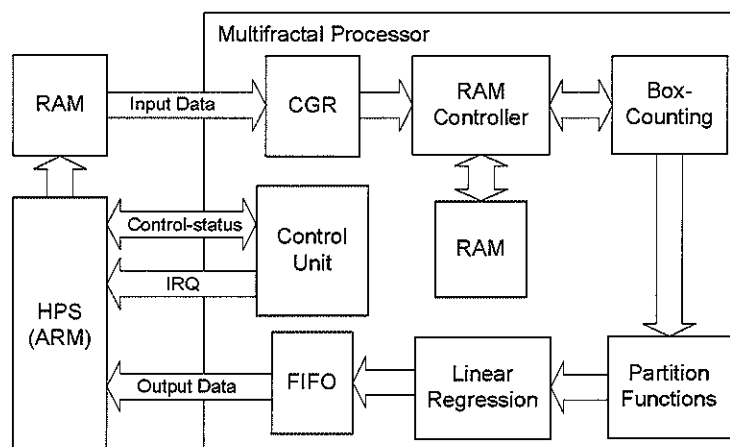


Fig. 2. Diagrama de bloques del procesador multifractal.

El diseño de cada bloque combina procesamiento paralelo y secuencial aprovechando la capacidad de implementar hardware específico y eficiente en la FPGA. En algunos de los bloques se utiliza punto flotante para la representación de los datos mientras que en otros se utiliza punto fijo dependiendo del rango de los valores manejados.

Resultados

El procesador fue verificado obteniendo las dimensiones fractales de los 24 cromosomas de un genoma humano de referencia. El tamaño total de los 24 archivos FASTA es aproximadamente 2.7GB y fueron divididos en fragmentos de 300,000 bases, obteniendo 9,754 fragmentos completos. Las dimensiones fractales de cada fragmento fueron calculadas en un rango de exponentes entre -20 y 20, usando un tamaño de grilla máximo de 256x256 cajas. En total 399,914 dimensiones fractales fueron calculadas.

Los resultados se compararon con datos de software obtenidos con un clúster de tres servidores, cada uno con dos procesadores Intel Xeon E5-2650 V2 de ocho núcleos físicos para un total de 48 núcleos y 558GB de RAM. Los resultados también se compararon con un computador (PC) con un procesador Intel Core i7-3632QM de cuatro núcleos y 12GB de RAM.

El genoma humano completo fue procesado en 12.5 minutos usando el Procesador Multifractal, mientras que el clúster requirió 29.3 minutos y el PC, 40 horas y 19 minutos.

La Fig. 3 muestra las gráficas de los espectros multifractales de los fragmentos 1, 97 y 125 del cromosoma 20 obtenidos con el procesador multifractal y el clúster. Como se puede ver, los resultados obtenidos con ambas implementaciones son muy cercanos entre sí. La Fig. 4 muestra el mapa multifractal del cromosoma 18. En este caso también se observa que los datos obtenidos con el procesador multifractal siguen con bastante precisión los resultados obtenidos en software.

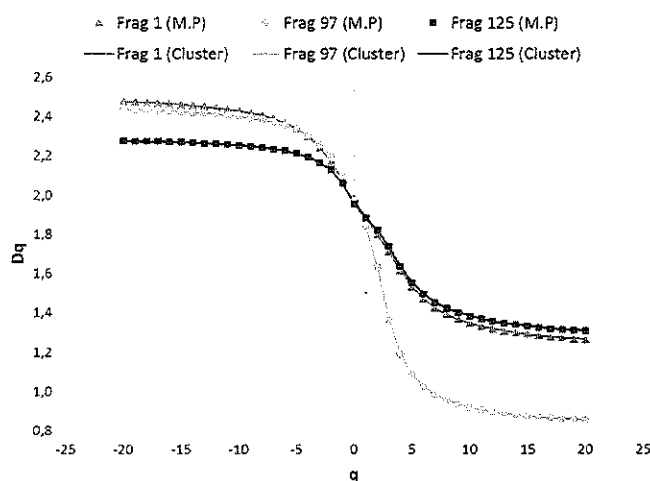


Fig. 3. Espectros Multifractales de los fragmentos 1, 97 y 125 obtenidos con el Procesador Multifractal (M.P) y el clúster.

Coefficientes de correlación y medidas del error de los grados de multifractalidad de cuatro cromosomas obtenidos con el Procesador Multifractal en comparación a los obtenidos con el clúster

cromosoma	Coefficiente de Correlación	Max abs(Error relativo)	Promedio de abs(Error relativo)	Desv. Estd. Error Relativo
10	0,999994	0.70%	0.0056%	0.041%
13	0,999980	1.16%	0.0093%	0.073%
18	0,999987	0.54%	0.0105%	0.056%
20	0,999983	0.85%	0.0139%	0.074%

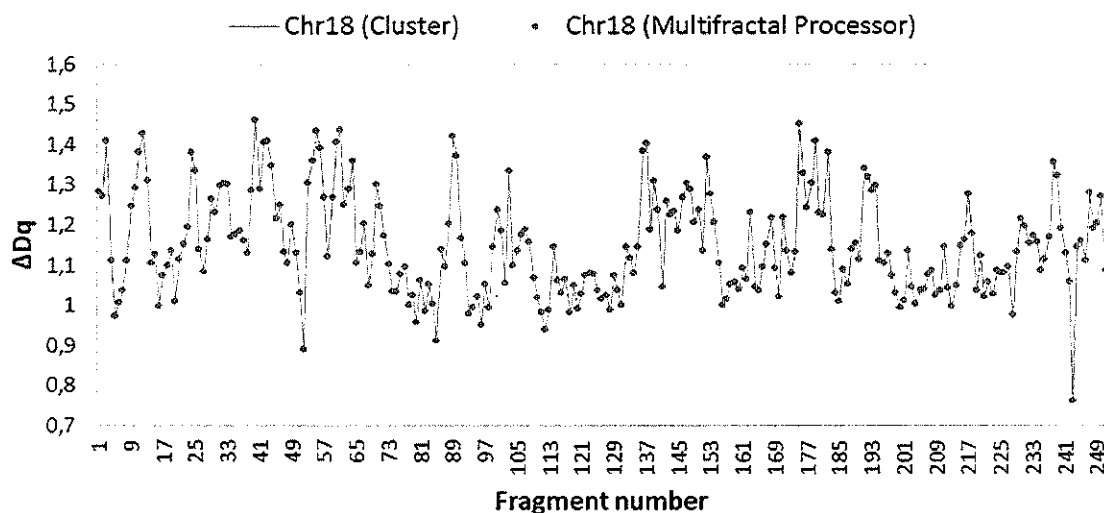


Fig. 4. Mapa Multifractal del cromosoma 18 del genoma humano

Conclusiones Principales

A partir de la optimización del algoritmo usado para el análisis multifractal de secuencias de ADN, se implementó una arquitectura hardware, denominada Procesador Multifractal, que permite acelerar 2.34x y 193.5x el cálculo de las dimensiones fractales de un genoma humano completo con respecto a un clúster de 48 núcleos y a un PC, respectivamente, con errores promedios relativos del orden de 0.01%.

En este trabajo se logró aprovechar las funcionalidades de los dispositivos híbridos SoC-FGPA integrando el procesador multifractal diseñado con un procesador ARM embebido para crear un sistema completamente funcional de bajo costo que permite leer secuencias de ADN en formato FASTA desde una tarjeta SD, enviar los datos al procesador, y guardar las dimensiones fractales calculadas en un archivo de resultados.

4. Impactos actual o potencial:

Elaborado por: Vicerrectoría de Investigaciones



Durante la ejecución de este proyecto se aportó en la formación de un estudiante de maestría en ingeniería electrónica quien con este trabajo aplicó sus conocimientos de diseño hardware en un problema de bioinformática, además le permitió apropiarse nuevas tecnologías, específicamente el manejo de los dispositivos SoC-FPGA, lo cual facilitará su apropiación a otros estudiantes a corto plazo.

Con la publicación de nuestro trabajo en una revista de categoría A1, no solo daremos a conocer nuestra metodología utilizada para llevar a cabo proyectos de integración de desarrollo hardware y bioingeniería, sino que también servirá como plataforma para mostrar a la comunidad científica las investigaciones que se están llevando a cabo en la Universidad del Valle en estudios de genómica utilizando técnicas novedosas como el análisis multifractal, y dispositivos de última tecnología como los SoC-FPGA.

El sistema desarrollado en este proyecto permitirá llevar a cabo el análisis multifractal de muchas secuencias de ADN en un tiempo práctico reduciendo costos operativos, lo cual favorecerá estudios de cáncer que nuestro grupo de trabajo está realizando actualmente. No se descarta también la posibilidad de transferir este conocimiento a otros grupos o centros de investigación nacionales o internacionales que también realicen análisis multifractal de secuencias de ADN.

5. Productos:

Elaborado por: Vicerrectoría de Investigaciones

Tabla No. 1. Cantidad y tipo de productos pactados en el *Acta de Trabajo y Compromiso* y productos finalmente presentados

TIPO DE PRODUCTOS	No. de PRODUCTOS PACTADOS				No. de PRODUCTOS PRESENTADOS			
Productos de nuevos conocimientos								
Artículo en revista ISI-SCOPUS:	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Artículo completo publicado en revistas indexadas	A1	A2	B	C	A1	A2	B	C
			1		1			
Libros de autor que publiquen resultados de investigación	-				-			
Capítulos en libros que publican resultados de investigación	-				-			
Productos o procesos tecnológicos patentados o registrados	-				-			
• Prototipos y patentes	-				-			
• Software	-				-			
Productos o procesos tecnológicos usualmente no patentables o protegidos por secreto industrial	-				-			
Normas basadas en resultados de investigación	-				-			
Formación de recursos humanos	No. de estudiantes vinculados	No. de tesis			No. De estudiantes Vinculados	No. De tesis		
Estudiantes de pregrado	2	-			-	1		
Semillero de Investigación	1	-			-	-		
Estudiantes de maestría	1	-			1	1		

Elaborado por: Vicerrectoría de Investigaciones



TIPO DE PRODUCTOS	No. de PRODUCTOS PACTADOS		No. de PRODUCTOS PRESENTADOS	
Estudiantes de doctorado	-	-	-	-
Joven investigador	-	-	-	-
Productos de divulgación				
Publicaciones en revistas no indexadas	-		-	-
Ponencias presentadas en eventos (congresos, seminarios, coloquios, foros)	No. de ponencias nacionales	No. de ponencias internacionales	No. de ponencias nacionales	No. de ponencias internacionales
	1	-	1	-
Propuesta de investigación				
Propuestas presentadas en convocatorias externas para búsqueda de financiación.	1		1	

En el proyecto no se vincularon estudiantes de pregrado en la modalidad de monitorías o asistencias de investigación debido a que los recursos del rubro de personal aprobado para el proyecto solo fueron suficientes para la vinculación del estudiante de maestría Jorge Duarte como asistente de investigación por 12 meses. Sin embargo, el estudiante de ingeniería de sistemas Cris Loaiza participó en el proyecto con aportes que permitieron evaluar diferentes metodologías para mejorar el algoritmo que calcula las dimensiones fractales y llevó a cabo su trabajo de grado titulado "Desarrollo de una Aplicación para el Análisis Multifractal "detrended", a partir de la temática desarrollada en este proyecto. La constancia de la dirección de tesis se adjunta como soporte.

El estudiante de maestría vinculado como asistente de investigación desarrolló su trabajo de grado a partir de los resultados parciales obtenidos durante su participación en el proyecto. Esto permitió profundizar más en el desarrollo de la arquitectura diseñada e implementarla utilizando el dispositivo SoC-FPGA con el cual se consiguieron los resultados finales del proyecto.



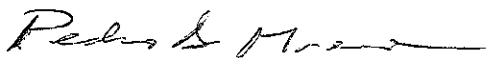
No se vincularon semilleros de investigación debido a que durante la ejecución del proyecto no se presentaron estudiantes que cumplieran con los requisitos establecidos por la universidad para este programa.

Presentación a convocatoria externa: Actualmente el grupo de investigación está ejecutando un proyecto financiado con recursos de Sistema General de Regalías del Dpto. del Valle, llamado abreviadamente: Proyecto “Ómicas” en cáncer mamario, el cual fue presentado a finales del 2013 y aprobado en noviembre de 2014 y comienzo de ejecución en febrero de 2015. Uno de las actividades de este proyecto es analizar los exomas de pacientes con cáncer mamario mediante análisis multifractal. En éste proyecto, nosotros venimos trabajando bajo el enfoque multifractal fundamentado en el algoritmo de “box counting”, el cual ha sido muy exitoso. Con el fin de determinar si un enfoque alternativo llamado análisis multifractal “detrended” pudiera mejorar las correlaciones arrojadas con el método del “box counting” se planteó una tesis de pregrado desarrollada por el estudiante de ingeniería de sistemas Cris Loaiza. Sus resultados permitieron determinar que ambos métodos dan correlaciones muy parecidas, no obstante, el método “detrended” consume más recurso de procesamiento que el que venimos trabajando. En el proyecto referido de regalías se ha continuado con el desarrollo del sistema presentado en este informe con el propósito de realizar algunas optimizaciones que permitan disminuir aún más el tiempo de procesamiento. Esto se justifica ya que este sistema será utilizado en el análisis multifractal de las muestras que se están obteniendo actualmente para el proyecto de regalías, en el estudio de cáncer mamario en la región del suroccidente colombiano pero a nivel de PCs. En efecto, los resultados generados por el proyecto en FPGA permitirá en un futuro cercano ejecutar de manera acelerada el análisis multifractal de los exomas (ya sea, por “box counting” o “detrended”), mediante el uso de chips FPGA. Esto para mitigar las necesidades impuestas por el uso de potentes y costosos servidores y que se puedan efectuar en un simple PC en la praxis médica diaria.

Tabla No. 2. Detalle de productos

Tipo de producto:	Artículo. Se adjunta borrador del artículo para sumisión a la revista a más tardar este fin de semana.
Nombre General:	BMC Bioinformatics, categoría A1 según Publindex Colciencias. Q1 internacional
Nombre Particular:	Accelerating the Multifractal Analysis of DNA Sequences using a SoC-FPGA
Ciudad y fechas:	Inglaterra, Octubre de 2016.
Participantes:	Jorge Duarte, Jaime Velasco, Pedro A. Moreno
Sitio de información:	https://bmcbioinformatics.biomedcentral.com
Formas organizativas:	Grupo de Bioinformática y Grupo de Bionanoelectrónica, Facultad de Ingeniería, Universidad del Valle.

La presente versión del informe contiene las observaciones de los evaluadores:



Firma del investigador principal

VoBo. Vicedecano de Investigaciones

Por favor presente su informe impreso y en formato digital en hoja tamaño carta, letra arial 11, con espacios de 1 1/2