MODELO DE PREDICCIÓN DE EXONES QUE INTEGRA CARACTERÍSTICAS FRACTALES DEL ADN EN UNA TÉCNICA DE APRENDIZAJE DE MÁQUINA DE MINERÍA DE DATOS

CARLOS E. TÉLLEZ V.

EDWIN F. CALDÓN P.

Monografía para optar al titulo de

Ingeniero de Sistemas

Director

NESTOR M. DIAZ M.

Ingeniero De Sistemas

Asesora

MARTHA I. ALMANZA P.

Ph.D.(c) en Mejoramiento Genético

UNIVERSIDAD DEL CAUCA

FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES

DEPARTAMENTO DE SISTEMAS

GRUPO DE BIOLOGÍA MOLECULAR AMBIENTAL Y CÁNCER – BIMAC

GRUPO DE TECNOLOGÍAS DE LA INFORMACIÓN - GTI

POPAYÁN, CAUCA

2008

Nota de Aceptación:

_____

_____

_____

Director del proyecto

_____

NESTOR M. DIAZ M.

Jurado

_____

NOMBRE

Jurado

_____

NOMBRE

Popayán, _____ de _____ de 2008

Resumen

El presente trabajo analiza la búsqueda de patrones en secuencias ADN que evidencien la presencia de regiones codificantes, con el fin de construir un modelo de predicción de exones, aprovechando las características estadísticas y fractales presentes en el ADN y medir la capacidad de dichas características en la clasificación de secuencias codificantes.

Es de vital importancia la identificación de componentes en las secuencias de ADN tales como, regiones codificantes y no codificantes, regiones promotoras, regiones dadoras, entre otras señales, con un alto nivel de confianza, ya que dichos componentes son la base de la expresión genética, la cual permite por técnicas de comparación de genomas eucariotes completar la información de las secuencias en las Base de Datos genómicas que aun no están secuenciadas experimentalmente o inferir información de organismos no conocidos con base en las predicciones realizadas. Pero el problema de predicción de secuencias codificantes aún no se resuelve a un nivel satisfactorio y cabe aplicar otra clase de enfoques, que ayuden a construir nuevos caminos de exploración en este tema.

Mediante Minería de Datos se aplica medidas estándar usadas en la predicción de genes y medidas fractales utilizadas en teoría de la información y sistemas complejos, para analizar patrones presentes en las secuencias de ADN con el fin discriminar secuencias codificantes de no codificantes. Con los patrones encontrados se crea un modelo aplicando técnicas de Aprendizaje de Máquina usadas en Minería de Datos para medir la efectividad de clasificación de los patrones encontrados entre exones e intrones.

Los patrones encontrados por las medidas estadísticas tienen una efectividad promedio del 70% de clasificación, las medidas fractales usadas en sistemas complejos tienen una efectividad del 80% y las medidas aplicadas en la teoría de la información ofrecen información importante acerca de los genomas estudiados pero no es lo suficientemente especifica para obtener información de una secuencia de forma independiente. Empleando sólo las medidas estadísticas como atributos del modelo de clasificación de exones se obtuvo en promedio 79.8% de efectividad con árboles de decisión (AD), 78% con Redes Bayesianas (BN) y 77.8% con Redes Neuronales (PM); usando sólo atributos fractales se obtuvo 88% con AD, 86.3% con BN y 85% con PM; y la combinación de los dos tipos de atributos se obtuvo un 91.8% con AD, 89% con BN y 89.4% de efectividad con PM.

Los resultados de esta investigación sugieren que las medidas fractales ofrecen un alto porcentaje de efectividad de clasificación frente a las medidas estadísticas y en combinación se obtiene un alto nivel de clasificación.

Este trabajo muestra que se deben seguir invirtiendo esfuerzos en la aplicación de medidas fractales para la búsqueda de patrones o señales en las secuencias de ADN. Extrapolar estas medidas en la búsqueda de otras señales en genes como regiones promotoras, regiones terminadoras, entre otras y desarrollar herramientas bioinformáticas para investigadores científicos.

Índice de contenido

Lista de Cuadros

Índice de Gráficas

Agradecimientos

Esta Tesis es el resultado de la contribución de muchas personas que, de alguna u otra manera nos apoyaron, haciendo que el camino recorrido se haya hecho menos difícil. A todos ustedes expresamos nuestra gratitud y compartimos este logro alcanzado.

Como parte de una fe que compartimos, queremos agradecer al creador por habernos colocado en este camino y en compañía de personas tan especiales con las que compartimos y aprendimos, por habernos dado la fortaleza y sabiduría para afrontar las dificultades y culminar este objetivo.

A nuestros respectivos padres que con sacrificios económicos, con gran amor y confianza depositada en cada uno de nosotros, hicieron posible alcanzar el primer escalón hacia nuestros sueños y que seguirán acompañándonos y apoyándonos por siempre.

A los profes Nestor Díaz, Luis Garreta, Ember Martinez, Carlos Cobos, Patricia Vélez, Pedro Moreno y Lorena Vidal que nos regalaron muchos minutos de sus tiempos para darnos sus aportes, criticas y consejos.

Un agradecimiento especial a Martha Almanza quien nos guió en el proceso investigativo dándonos una visión más amplia de la bioinformática y sus aplicaciones, además de los consejos que nos ayudaron a crecer personal y profesionalmente.

A los compañeros y amigos del grupo BIMAC: Adrian Ararat, Fabian Tobar, Yhon Acosta y Miguel Guevara con los cuales compartimos la mayor parte del tiempo en el laboratorio y quienes nos apoyaron en temas relacionados con la Biología y hacer de nuestra estadía en el grupo BIMAC más amena.

A todos los compañeros y compañeras de estudio quisiéramos darles las gracias por los buenos momentos que hemos compartido tanto profesionales como personales, de los cuales hemos aprendido y apreciamos bastante, en especial un cariñoso reconocimiento a los que nos han demostrado su apoyo y brindado sus ánimos y consejos durante este periodo universitario: Claudia Montealegre, Ana Maria Chimunja, Marly Quiñones, James Erazo, Diego Garcia, Gabriel Muñoz, Julio Palechor.

Al software libre por ofrecer la oportunidad de desarrollar y utilizar herramientas al alcance de todos.

Siempre se cometen injusticias en los agradecimientos, pues la memoria es a menudo traicionera, por ello agradecemos a todos aquellas personas que no se alcanzaron a mencionar antes y esperamos que reciban nuestros mas sinceros agradecimientos.

1. Introducción

La predicción bioinformática de genes es un tema central en la era genómica y postgenómica. A la fecha se han secuenciado casi completamente mas de 23 genomas eucariotes, mas de 230 genomas procariotes y más de 500.000 genes descubiertos mediante la aplicación de herramientas computacionales y técnicas matemáticas (NCBI, 2008). En efecto, a la fecha existe una colección de más de 20 aplicaciones, cada una con sus ventajas, desventajas y niveles de predicción del gen. Pese a este espectacular desarrollo, muchos interrogantes y limitaciones emergen día a día, dada la compleja tarea que significa entender la organización del gen en los genomas y dada las muchas alternativas teóricas por entender y aplicar en estos problemas biológicos.

La predicción de genes consiste en detectar dentro de una secuencia de ADN, las regiones codificantes e inferir la estructura del gen. Desde el punto de vista computacional, el problema se puede ver como un problema de identificación de patrones, sólo que los patrones que componen el gen son bastante irregulares, lo que lo convierte en una tarea no trivial y requiere un largo camino para resolverlo de manera no ambigua. Las regiones codificantes son las que constituyen en últimas lo que se conoce como exones, se podría ver de forma simple al gen como un conjunto de exones.

La identificación o predicción de exones hace parte del problema de predicción de genes (Wang y Li, 2004; Mathe et al., 2002; Zhang, 2002), el cual consiste en identificar dentro de una secuencia de ADN los distintos componentes del gen, de los cuales los exones son los más importantes. Este problema a pesar de ser ampliamente estudiado, aún no está resuelto y es complejo a nivel computacional, ya que los genes, a pesar de presentar unos componentes determinados (promotores, exones, intrones, etc.), presentan una estructura bastante irregular.

Sobre este problema y específicamente sobre la identificación de exones se han creado varios modelos (Majoros et al., 2004; Burge y Karlin, 1996; Salzberg et al., 1999; Xu et al., 1994), los cuales se han implementado en distintos programas predictores de genes tales como GenScan, GlimmerM, Genezilla, Morgan, Veil, Grail, entre otros. Estos programas utilizan distintos algoritmos de Aprendizaje de Máquina (AM) para identificar exones, entre ellos: los árboles de decisión (Morgan), las redes neuronales (Grail), programación dinámica (Morgan) y los modelos ocultos de Markov (GenScan, Genezilla, Veil, GlimmerM).

Pero a pesar de toda esta variedad de programas y algoritmos, el problema de identificación de exones aún no se resuelve al ciento por ciento, ya que los predictores actuales ofrecen para una misma secuencia de ADN diferentes probabilidades de acierto entre lo que es un exón y no-exón, y por lo tanto se necesitan nuevos modelos que intenten otros abordajes. En este sentido, últimamente han tomado gran importancia las propiedades fractales del ADN como un elemento que puede ayudar a caracterizar diferentes elementos de los genes, entre estos los exones (Vélez et al., 2004).

Sin embargo, para involucrar estas características del ADN en un nuevo modelo se requiere de técnicas adecuadas para este tipo de problemas, de las cuales las de Aprendizaje de Máquina (Machine Learning) han mostrado gran efectividad, y de estas las que usan algoritmos de aprendizaje supervisado, debido a la gran cantidad de bases de datos de secuencias de ADN ya caracterizadas (exones ya identificados) que pueden usarse como conjuntos de entrenamiento.

Por lo tanto, el problema objeto de estudio que aquí se analiza, consiste en la construcción de un modelo de identificación de exones que tenga en cuenta medidas que evidencien la presencia de exones, tales como las características fractales del ADN, con la aplicación de técnicas de AM usadas en Minería de Datos, con el fin de probar la influencia de las características fractales en el modelo de identificación de exones.

El impacto de este trabajo se ve reflejado como un aporte al enfoque investigativo del grupo naciente de bioinformática de la Universidad del Cauca (Grupo de Biología Molecular Ambiental y Cáncer - BIMAC), el cual se encuentra en el marco de un proyecto más amplio como es el análisis de secuencias de genes y genomas completos (Vélez et al., 2004).

Lo anterior lleva a formular las siguientes preguntas de investigación:

¿Qué algoritmos de AM usados en Minería de Datos son adecuados al problema de Identificación de exones?

¿Cómo construir un modelo de identificación de exones que integre las características fractales del ADN?

¿En qué medida las características fractales del ADN influencian la clasificación de una secuencia de ADN como exón o no-exón?

Para generar acercamientos y discusiones en relación con estas inquietudes, es necesario lograr un entendimiento de importantes áreas del conocimiento, que se presentan en el capítulo 2, como Biología Molecular (Sección 2.1), Medidas estándar o clásicas para la predicción de genes (Sección 2.2), Fractales (Sección 2.3), Bases de Datos Genómicas (Sección 2.4), Minería de Datos (Sección 2.5) especialmente la Minería de Texto. Finalmente se presentan las técnicas usadas en Aprendizaje de Máquina (Sección 2.6) y Modelos de predicción de genes/exones (Sección 2.7).

En el capitulo 3 se presentan todos los aspectos de la solución propuesta, como la aplicación de la metodología de Minería de Datos (Sección 3.1) para la búsqueda de patrones en las secuencias de ADN se define el conjunto de datos, se presenta una propuesta del Modelo de predicción de exones con los patrones encontrados y las evaluaciones hechas al modelo. Terminando este capítulo con la metodología de Desarrollo de Software usada (Sección 3.2).

En el capitulo 4 se presentan el análisis de los resultados obtenidos en cada una de las medidas que se desarrollaron.

En el capitulo 5 se crea y evalúa el modelo de predicción de exones con las diferentes técnicas de clasificación.

En el capítulo 6 se hacen acercamientos y discusiones a las soluciones dadas a las preguntas de investigación, se presentan las recomendaciones, se plantean mejoras a las dimensiones fractales usadas y temas a tener en cuenta para continuar con el desarrollo de la Predicción de exones.

2. Marco teórico

En esta sección se presentan algunos conceptos de Biología Molecular tales como, células, ácidos nucleicos, genes, cromosomas, genomas y el Dogma Central de la Biología Molecular donde se explica brevemente la Expresión de Genes. Continuando con Fractales, Análisis Unidimensional y Bidimensional y Análisis Zipf-Mandelbrot. Terminando esta sección con Aprendizaje de Máquina, Minería de Datos y Modelo de predicción de genes/exones.

2.1 Conceptos básicos de Biología Molecular

El trabajo desarrollado se encuentra enmarcado en el área de Biología Molecular, una ciencia que estudia los procesos celulares que ayudan a que la información genética se transmita de forma eficiente de unos seres a otros y se exprese en los nuevos individuos. Algunos conceptos específicos sobre esta área, se describen a continuación.

2.1.1 Organismos

Un organismo es un sistema individual vivo tal como un animal, una planta, un hongo o un micro-organismo. Todo organismo es capaz de reaccionar a estímulos, reproducirse, crecer y morir, manteniéndose relativamente estable durante esta etapa. Los organismos se pueden dividir en dos grupos: procariotes y eucariotes. Los procariotes representan las bacterias y arqueas. Todos los hongos, animales y plantas son eucariotes. Un organismo puede ser unicelular o en su defecto, como los humanos, de billones de células agrupados en tejidos y órganos, estos últimos se denominan organismos complejos (Celis, 2001).

2.1.2 Células

Son elementos fundamentales de los organismos, se pueden ver como los ladrillos de una construcción. En éstas se encuentra el material genético, elemento fundamental de la descendencia genética.

Según la complejidad estructural, las células se dividen en dos grandes grupos: procariotes y eucariotes. Las células eucariotes tienen compartimentos bien definidos como el núcleo y las organelos (mitocondrias, cloroplastos). Las células procariotes no tienen compartimentos internos y en particular no tienen núcleo por lo tanto su material genético permanece en el citoplasma. En las células de organismos eucariotes, el material e información genética se encuentra en el núcleo (Celis, 2001).

2.1.3 Cromosomas

Los cromosomas son estructuras organizadas de ADN que se encuentran en el núcleo de las células. Es un pequeño cuerpo que durante la división celular tiene forma de equis en el cual se organizan los genes, elementos reguladores y otras secuencias de nucleótidos. Regularmente las células eucariotes tienen grandes cromosomas lineales (alargados), mientras que las células procariotes tienen pequeños cromosomas circulares. Su número es constante para un organismo determinado; en *H. sapiens* se tienen 46: 44 son autosómicos y 2 son sexuales (Katsumi, 2002).

2.1.4 Genes

El gen se considera como la unidad de almacenamiento de información y unidad básica de herencia. Los genes se encuentran en cada uno de los cromosomas que se duplican durante la división celular, permitiendo que la herencia se lleve a cabo. Los genes pueden ser considerados como segmentos separados y discretos de una molécula de ADN, donde esta almacenada la información genética (Celis, 2001).

Teniendo en cuenta que los genes no están situados siempre de forma contigua en la secuencia de ADN, sino que están separados por una gran cantidad de sub-secuencias que se conocen como regiones intergénicas (esto varia entre distintos genomas). No esta claro el papel que juegan algunas de estas regiones por eso se han denominado “residuo ADN” (junk DNA). En los procariotes los genes comprenden casi todo el cromosoma, presentándose pocas regiones intergénicas, es decir, una proporción del 90% son genes y el 10% son regiones intergénicas. Este no es el caso de los eucariotes, donde el volumen de genes es mucho menor en relación con las regiones intergénicas. En el caso del genoma humano, menos del 5% se compone de genes (Katsumi, 2002).

2.1.5 Genoma

La información genética contenida en las células de un organismo comprenden su genoma. Cuando se habla del genoma eucariote se refiere sólo al ADN contenido en el núcleo de la célula, organizado en cromosomas.

2.1.6 Ácidos Nucleicos

Los ácidos nucleicos son moléculas que almacenan las informaciones relativas a los desenvolvimientos y divisiones de las células, las cuales forman los organismos vivos. Existen dos tipos de ácidos nucleicos: ADN o ácido desoxirribonucleico y ARN ácido ribonucleico (Almeida, 2002).

En su estructura, los ácidos nucleicos se pueden ver como cadenas lineales o biopolímeros compuestas de unidades químicas llamadas nucleótidos o biomonómeros (par base: pb). Biológicamente existen cinco bases nitrogenadas principales, que se clasifican en

dos grupos: bases púricas (derivadas de la purina) y las bases pirimidínicas (derivadas de la pirimidina). La adenina (A) y la guanina (G) son púricas, mientras que la timina (T), la citosina (C) y el uracilo (U) son pirimidínicas. En el ADN se encuentran las bases A, G, C y T. En el ARN se encuentra la base U en lugar de la base T.

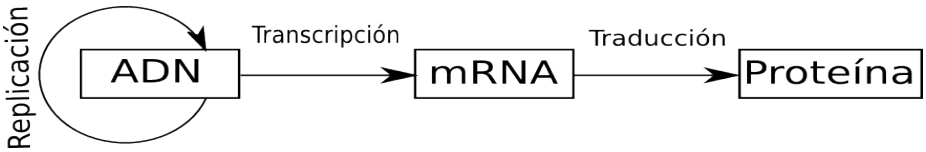
Las moléculas de ADN se componen de dos hebras, que se unen entre si formando una estructura helicoidal, conocida como doble hélice, las dos hebras se acoplan por la unión regular sus nucleótidos. Una base A siempre se une a una base T y una base G a una base C (Gráfica 1). Esa disposición exclusiva y única entre las bases (bases complementarias) se debe generalmente a su tamaño, a su forma y a su composición química. Las dos hebras son anti-paralelas, es decir, las hebras poseen orientación 5' -> 3' opuesta una en relación a otra (Almeida, 2002).

```
5'      ...  A A T C G G C A T T C G G T C  ...      3'
             | | | | | | | | | | | | | | |
3'      ...  T T A G C C G T A A G C C A G  ...      5'
```

Gráfica 1: Ejemplo de la unión de las hebras de ADN
Fuente: (Almeida, 2002)

2.1.7 Dogma central y expresión de Genes

Un gen es un segmento de ADN que lleva información genética, la cual esta disponible para la célula por medio de la expresión genética. Cuando eso sucede, una copia del gen se sintetiza (transcripción) en una molécula de mARN (ARN mensajero), la cual se usa para la fabricación de una proteína (traducción). Dicho flujo de información que involucra ADN, ARN y proteína fue descrito por Francis Crick como el Dogma Central (Katsumi, 2002) (Gráfica 2):

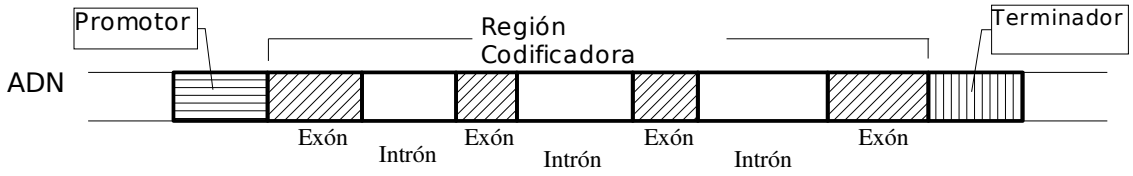


Gráfica 2: Dogma Central de la Biología Molecular

Como un resultado de este dogma se tiene que una proteína está asociada a un gen determinado.

En la estructura básica, los genes poseen una región denominada **promotor**, que es el responsable de su activación, en seguida de este se encuentra la **región codificadora** y una región terminadora (Gráfica 3). Una región codificadora es un segmento de gen que contiene la información usada para sintetizar una proteína, y se compone principalmente de

regiones codificantes (**exones**) y, regiones no-codificantes (**intrones**), las cuales se encuentran intercaladas entre si. Los exones son los segmentos que realmente se transcriben y traducen en proteínas. En cambio los intrones se remueven del mARN antes de realizar el proceso de traducción, lo que se conoce como splicing. Además entre el exón y el intrón se encuentran dos estructuras conocidas como sitios de splicing y corresponden al sitio dador (entre exón e intrón) y sitio aceptor (entre intrón y exón). El terminador es un segmento de ADN que señala el final de la síntesis de la molécula de mARN.



Gráfica 3: Estructura de un Gen eucariote

Fuente: (Katsumi, 2002)

Para efectos computacionales, el ADN se representa como una secuencia de 4 letras correspondientes a los cuatro nucleótidos: Adenina (A), Citosina (C), Guanina (G), Timina (T) (Gráfica 1). El tamaño de la secuencia varía de acuerdo a la especie, desde millones de letras, en el caso de las bacterias, hasta billones de letras, en el caso de los mamíferos. Así, las proteínas son secuencias cortas de otro alfabeto, el alfabeto de 20 aminoácidos. Cada tres nucleótidos codifican un aminoácido. Cada una de las 64 (4^3) posibilidades se denomina codón. De los 64 codones posibles, algunos codifican un mismo aminoácido; por eso se tienen 20 aminoácidos. Para determinar el inicio y final de la traducción de una secuencia se usan los codones de inicio y parada respectivamente, esta sub-secuencia limitada por estos dos tipos de codones se denomina marco de lectura abierto (ORF - Open Reading Frame) (Katsumi, 2002).

2.1.8 Organismos Modelo en Biología

Los organismos modelos son especies ampliamente estudiados para entender fenómenos biológicos particulares con la expectativa de descubrir hechos en los organismos modelos que puedan proveer información que pueda ser aplicada en otros organismos (Deutsch y Long, 1999).

Esto es posible porque los principios biológicos fundamentales tales como el metabolismo, la regulación y las rutas metabólicas, y los genes que codifican para estos se conservan a través de la evolución (Moreno et al., 2006).

Las principales características para que un organismo sea modelo son: ciclo de vida corto, tamaño adulto pequeño, económicos y eficientes para mantener y estudiar, genes

fácilmente manipulables, fácilmente disponible y son focos de proyectos de investigación (Moreno et al., 2006).

Los principales organismos modelo en biología molecular son:

Saccharomyces cerevisiae (Goffeau et al., 1996), *Caenorhabditis elegans* (C. elegans, 1998), *Drosophila melanogaster* (Adams et al., 2000), *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), *Oriza sativa* (Yu et al., 2002), *Mus musculus* (Waterston et al., 2002), *Homo sapiens* (Venter et al., 2001), *Gallus gallus* (International Chicken Genome Sequencing Consortium, 2004) (Wallis et al., 2004).

2.2 Medidas estándar para la predicción de genes

Una medida estadística puede definirse como una función que calcula, dada una secuencia de ADN, un número real relacionado con la posibilidad que una secuencia codifique o no para proteína. Muchas de estas medidas se basan en la posición de los codones (tripleta de nucleótidos o tripletas de par base) o en la periodicidad de la aparición de las bases, o una combinación de ambas (Guigó, 1998).

2.2.1 Uso de Codón

Medida que calcula frecuencias de uso de tripletas, esta entre las más usadas para definir frecuencias de uso para cada uno de los 64 posibles codones (Fickett y Tung, 1992). Básicamente se hace un conteo de codones en la secuencia de ADN, en cada uno de las tres fases de lectura, y luego se calcula las frecuencias de ocurrencia de los codones encontrados, para finalmente calcular un índice de probabilidad entre las frecuencias halladas (probabilidades de encontrar codones codificantes) y la frecuencia de hallar un codón no codificante; esto con el fin de establecer un valor que permita discriminar secuencias codificantes de no-codificantes (Guigó, 1998).

Las frecuencias relativas de cada uno de los codones y que pertenecen a un aminoácido, son diferentes debido; primero, a que en el proceso de la traducción tiende a usar abundante RNA de transcripción (y por supuesto los codones correspondientes a este RNA de transcripción). El uso de codón es el mismo para los genes que se expresan de forma continua y abundantes en cada especie. Segundo, a las diferentes tasas de mutación que son especificas para cada especie principalmente los cambios que se presentan entre pares de bases: GC cambia a AT y AT cambia a GC (Medrano-Soto, 2004).

2.2.2 Hexámeros

Un hexámero es la unión de 6 par base o dos codones, el uso de hexámeros puede ser aplicado de manera análoga que el uso de codón. De acuerdo a las 6 combinaciones de los nucleótidos Adenina ,Timina ,Guanina y Citosina se tienen 4096 hexámeros posibles.

La frecuencia de aparición de los hexámeros en una secuencia puede ser utilizado para discriminar regiones codificantes de no codificantes. El análisis de hexámeros tiene mayor poder discriminatorio entre estas secuencias, dadas las relaciones químicas y físicas de los aminoácidos que constituyen una proteína. Con las diferentes frecuencias de ocurrencia de todos los hexámeros en las secuencias de exones e intrones se puede encontrar una frecuencia diferente para cada secuencia (Claverie y Bougueleret, 1986).

2.2.3 Ficket

Ficket (Fickett y Tung, 1992) evalúa el azar posicional en una secuencia: en secuencias codificantes la tercera base de un codón tiende a ser la misma con más frecuencia que la esperada por azar; esto debido al uso preferencial de ciertos codones, es una propiedad universal. Toma ocho mediciones en una ventana. Cuatro de ellas son simplemente las frecuencias de las bases. Las otras cuatro miden la asimetría de la composición de la base en las tres posiciones de codón. Es decir, con f(b,i), define como f(b)=max(f(b,1),f(b,2),f(b,3)) / [1+min (f(b,1), f(b,2),f(b,3))]. Es usado para estimar similitud en regiones codificantes y las estimaciones separadas son todas combinadas usando una suma lineal de pesos.

2.2.4 Contenido G+C

El contenido de nucleótidos G y C en las secuencias es importante ya que dependiendo de su cantidad, en determinadas regiones con mayor o menor contenido de C y G, se manifiesta ciertas propiedades tales como la densidad de genes o composición de secuencias repetitivas. El contenido GC es expresado usualmente como porcentaje, aunque algunas veces como una razón (llamada razón G+C o razón GC). El contenido GC en porcentaje se calcula: [(G+C)/(A+T+G+C)]*100, mientras que la razón GC se mide: (A+T)/(G+C) (Koski, 2001).

2.3 Fractales

El padre de la geometría fractal Benoít Mandelbrot (1924 -) fue quien utilizó y acercó el concepto fractal a la ciencia, por medio de la sistematización de sus ideas logró la creación de ilustraciones para sus ensayos, mostrando que la relación presente entre la forma y el contenido esta estrechamente ligada, para él es importante alcanzar una representación gráfica de las matemáticas, en este caso los fractales (Mandelbrot y Hudson, 2004).

En el ensayo “Les objets fractals: Forme, hasard et dimension”, publicado en 1975, Mandelbrot introduce el término fractal, en donde expresa mas o menos lo siguiente: tomé la palabra fractal que proviene del adjetivo latín fractus. El verbo correspondiente en latín, fragere, que significa “quebrar”: para crear fragmentos irregulares. Así en adición a “fragmentado”, fractus debería también significar “irregular”.

La geometría Fractal consiste en identificar, analizar, cuantificar y manipular patrones repetitivos; es una herramienta a la vez analítica y sintética. Un nuevo instrumento de medida, que no dice que tan largo, pesado, caliente o sonoro es algo, sino de cuán intrincado e irregular es (Mandelbrot y Hudson, 2004).

La geometría Fractal no utiliza formas perfectas como cuadros, triángulos, rectángulos entre otras, como lo hace la geometría convencional, la euclídea, si no que hace uso de formas con una complejidad irregular y entre mas fina sea la escala, mas detalles se revelan de los objetos estudiados.

La geometría euclidiana estudia formas que se pueden abstraer directamente por medio de puntos, rectas, círculos, triángulos, entre otros, y cuya dimensión fractal es entera, pero esta no es suficiente para explicar las formas de la naturaleza por ejemplo, las nubes, montañas, hortalizas, sistema circulatorio, y demás, los cuales tienen un comportamiento que no puede ser explicado por la geometría euclídea y cuyas dimensiones fractales son fraccionarias.

2.3.1 Definiciones para fractal

No se ha logrado dar forma a una definición precisa y formal para el término fractal, a continuación se presentan algunas aproximaciones que están relacionadas con el problema de buscar patrones en las secuencias de ADN:

- Un fractal es una clase especial de invarianza o simetría que relaciona un todo con sus partes: el todo puede descomponerse en partes que evocan el todo. (Mandelbrot y Hudson, 2004).
- Fractal es todo objeto que posee autosimilaridad. Una primera imagen intuitiva corresponde a un objeto infinitamente doblado sobre sí mismo, con infinitos pliegues, con infinita estructura (Sole y Manrubia, 2001).
- Un fractal es un objeto geométrico que al ampliarlo muestra una serie repetitiva de detalles, de tal forma que, sin importar a que escala se examine, la estructura parece ser la misma. Un fractal conserva el mismo aspecto si se observa a la escala de kilómetros, metros o milímetros (González, 1996).

2.3.2 Características de los fractales

Las principales características que definen un objeto o sistema fractal son (Burgos et al., 1996):

Autosimilaridad: muestra la relación de las partes con el todo, con la cual cada una de las partes tiene las mismas características del objeto completo.

Invarianza de escala: La forma de un objeto fractal permanece constante independiente de la escala de medida.

Dimensión fractal: Es una medida del grado de irregularidad, interrupción o fragmentación de un objeto fractal, en general se caracteriza por ser un dato fraccionario.

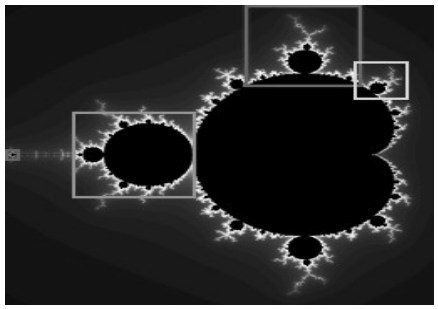
2.3.3 Tipos de fractales

Se presentan principalmente dos tipos: los lineales y los no lineales. Los lineales son idénticos a cualquier escala a la que son expuestos, como por ejemplo el Triangulo de Sierpinski (Gráfica 4), que muestra 4 iteraciones quitando fragamentos cada vez más pequeños. En estos fractales no importa a que escala se vea el objeto siempre conservará una forma similar a la original, se vera lo mismo sin importar cuantas veces se aumente o disminuya la escala.

Gráfica 4: Triángulo de Sierpinski

Fuente: (Ivars, 1992)

Los fractales no lineales conservan una estructura similar; pero dentro de las diferentes escalas a la que es observado no guardan exactamente su forma original, por ejemplo el Conjunto de Mandelbrot (Gráfica 5), en donde al acercarse a ciertas partes de la imagen se aprecia de nuevo la imagen en miniatura de la imagen real. Si se aumenta o disminuye la escala en este tipo de fractal se observan unas pequeñas variaciones con respecto al original, no son idénticas pero si semejantes.



Gráfica 5: Conjunto de Mandelbrot

Fuente: (Ivars, 1992)

Las secuencias de ADN por ser parte del complejo sistema de la naturaleza tienen características de fractales no lineales así como las nubes, las montañas, los árboles, etc.

2.3.4 Dimensión fractal

La dimensión fractal (D) es una medida que indica el grado de irregularidad y fragmentación de un sistema. Los valores son enteros en la geometría euclidiana, pero para la geometría fractal la mayoría de estos valores son fraccionarios.

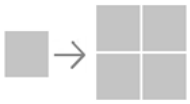
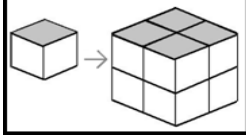
La dimensión fractal indica en cierta forma los grados de libertad que tiene un objeto, así un punto tendrá dimensión cero y cero grados de libertad, una recta posee dimensión 1 con un grado de libertad, un plano se podrá mover con 2 grados de libertad lo que indica una dimensión de 2 y una figura en el espacio tiene 3 grados de libertad lo que significa que su dimensión será 3. A este tipo de dimensiones se les conoce como dimensiones topológicas.

En la naturaleza existen curvas que no pueden ser medidas correctamente por una dimensión topológica, por ejemplo el interior de un pulmón o el follaje de un árbol, la rugosidad presente en este tipo de curvas representa un aumento en la dimensión, por lo tanto si se tiene una curva rugosa su dimensión podrá estar entre 1 y 2, y para una superficie rugosa estará entre 2 y 3.

2.3.4.1 Cálculo de la Dimensión Fractal

Para encontrar la dimensión ocupada por algún objeto se debe encontrar el factor de escala matemático por el cual se puede reproducir (Félix Hausdorff 1868-1942). Una definición sencilla de dimensión será $S=L^D$, En donde S es el número de figuras o copias iguales a la original, L es el factor de reproducción por el cual se desea ampliar la figura y D la dimensión de la figura.

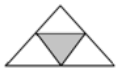
El cálculo de la dimensión en estructuras geométricas da como resultado números enteros y los valores de dimensionalidad son iguales a la dimensión topológica a la que pertenecen (Gráfica 6). Así por ejemplo, para duplicar el cubo (L), se necesitan 8 copias del cubo original (S) y su dimensión entera es 3.

Gráfico	Figura	L	$S=L^D$	D = log S / log L
	Línea	2	$2=2^1$	1
	Cuadrado	2	$4=2^2$	2
	Cubo	2	$8=2^3$	3

Gráfica 6: Dimensión de algunas figuras geométricas

S: número de copias; L: factor de reproducción. Adaptado de (Argote, 2004)

La estimación de la dimensión fractal de los fractales clásicos es fraccionaria y fue desarrollada por Mandelbrot en trabajos realizados entre 1977 y 1987. Para conocer la dimensión fraccionaria de un fractal se debe despejar la dimensión de la formula $S=L^D$, obteniendo D= logS/logL, con este procedimiento se puede ver que sus dimensiones son fraccionarias o fractales, aún con una dimensión topológica entera (Gráfica 7). En la gráfica, el triangulo de Sierpinski es un triangulo al cual se le quita la región que comprende la unión de los puntos medios de cada lado del triangulo grande, dando como resultado 3 triángulos con dimensión fraccionada (D) igual a 1,585, menor a la dimensión del triangulo original que era 2 (pues estaba sobre un plano).

Gráfico	Figura	L	$S=L^D$	D = log S / log L
	Conjunto de Cantor	2	$2=3^D$	0,631
	Curva de Koch	2	$4=3^D$	1,262
	Triángulo de Sierpinski	2	$3=2^D$	1,585

Gráfica 7: Dimensión de algunas figuras fractales

S: número de copias; L: factor de reproducción. Adaptado de (Argote, 2004)

2.3.5 Aplicabilidad de los fractales

Los fractales tienen una diversa gama de aplicaciones, gracias a la facilidad que tienen para simular de forma más precisa fenómenos de la naturaleza, lo que no sucede con la geometría euclidiana. Los fractales son utilizados en biología, física, economía, música, cine, entre otros. A continuación se hace una pequeña muestra de algunos usos.

- **Biología**: las aplicaciones de la geometría fractal en biología ha tenido un amplio uso, por ejemplo la existencia de autosimilaridad en las secuencias de ADN (Stanley et al., 1992), el estudio de la complejidad en estructura y funcionalidad de los pulmones (Bennett et al., 2000) y corazón (Ivanov et al., 1999), solo por mencionar algunos.

- **Física**: con el auge de las telecomunicaciones se busca mejorar cada vez más la capacidad de las antenas de comunicación. El uso de fractales ha permitido nuevos diseños y mejoras en el servicio prestado. Actualmente, este tipo de antenas se usan en sistemas celulares, dispositivos microondas, aeronáutica, entre otros.

- **Economía**: se han realizado varias aplicaciones de los fractales en la economía para determinar el comportamiento de los mercados bursátiles, muchos fenómenos

que ocurren en las finanzas no pueden ser explicados con los análisis económicos clásicos, pero con la ayuda de los fractales en las finanzas ha permitido transformar procesos complejos como la variación de precios de una moneda en una idea sencilla (Mandelbrot y Hudson, 2004).

- **Música**: se puede componer música a partir de la geometría fractal, existen programas que permiten componer música fractal, por ejemplo Musinum (www.musinumworld.com) que lo hace utilizando las series numéricas autosemejantes, Gingerbread por medio del conjunto de Mandelbrot, entre otros (Pérez, 2000).

- **Cine**: los efectos visuales generados mediante fractales se han utilizado para sets de grabación, reduciendo los costos que involucra la creación y mantenimiento de efectos como la lluvia, tormenta, nubes, cuerpos celestes, entre otros, para una escena en especial, una de esas películas es nada menos que la guerra de las galaxias (Aranda, 1998).

2.3.6 Fractales y ADN

Los primeros estudios que aplicaron geometría fractal al estudio del ADN se inician con la propuesta de Jeffrey (Jeffrey, 1990), con la representación del Juego del Caos (en inglés: Chaos Game Representation, CGR) que permite encontrar patrones autosimilares. Posteriormente se encuentran, Peng (Peng et al., 1992) y su modelo de caminata por el ADN, Hao (Hao et al., 2000) propuso un método de visualización mediante el conteo de la frecuencia de la aparición de subsecuencias de longitud. Jianbo Gao (Gao et al., 2004; 2005) planteó un índice denominado desviación fractal de la periodicidad 3 (period-3 fractal deviation – PFD), para determinar característica fractales tanto en regiones codificantes como en las no-codificantes. Los anteriores son algunos proyectos que han analizado, desde diferentes puntos de vista, la fractalidad de las secuencias de ADN; a continuación se presentan la fundamentación teórica de otros enfoques, para el análisis fractal de las secuencias de ADN, los cuales se usan en este proyecto para hallar patrones que puedan determinar si una región es codificante o no.

2.3.6.1 Análisis de Rango Reescalado R/S

El análisis de rango reescalado R/S fue introducido por Hurst (Hurst, 1951) y posteriormente desarrollado por Mandelbrot y colaboradores en artículos publicados entre 1969 a 1983. Es una de las pruebas más ampliamente usadas para determinar la dimensión fractal de una serie temporal. El estadístico R/S es el cociente entre el rango de las sumas parciales de las desviaciones de las medias de una serie de tiempo y la desviación estándar.

El exponente de Hurst se presenta en muchas áreas de las matemáticas aplicadas. Incluyendo los fractales y la teoría del caos, procesos de memoria a largo plazo y análisis

espectral. Su aplicación va desde la biofísica a la simulación del tráfico de redes (networking).

El exponente de Hurst es uno de los resultados de la aplicación del análisis de rango reescalado. Este análisis muestra la dinámica compleja presente en un sistema al cuantificar el flujo de sucesos que se encuentran interconectados en él.

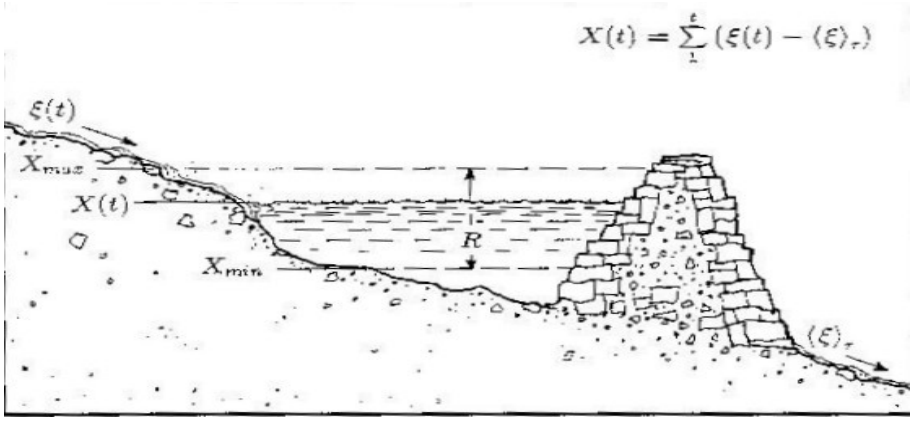
El rango reescalado es utilizado para analizar el comportamiento de un sistema a lo largo del tiempo. Al caracterizar porciones de la curva generada por medio de una pendiente, el exponente de Hurst, describe la irregularidad presente en el sistema a lo largo de las distintas ventanas en el tiempo. La mayor o menor fractalidad se debe a cuantas de estas ventanas presentan menor o mayor irregularidad, respectivamente (Suárez, 2004).

El rango reescalado se utiliza para analizar la fractalidad de las secuencias de ADN. Para lograr esto se mapea la secuencia de ADN a una secuencia de números enteros y se trata como un registro fractal en el tiempo. El objetivo es lograr distinguir regiones codificantes de las no codificantes por medio de las correlaciones que se encuentran entre estos dos tipo de secuencias (Yu y Chen, 1999).

Historia. Harold E. Hurst era un ingeniero hidrólogo británico (1880-1978). En el momento del diseño de la represa debió determinar la capacidad de almacenamiento, la cual es dependiente del flujo de entrada al río proveniente de diferentes fuentes como lluvias y riachuelos y un flujo controlado de salida utilizado primordialmente en el riego. Muchos hidrólogos habían supuesto que el flujo del río tenia un comportamiento aleatorio, por ser un sistema complejo; pero Hurst opinaba diferente.

Estudió los registros de río Nilo, y tal era la importancia del comportamiento del Nilo para los egipcios que Hurst tuvo a su disposición una base de datos de 800 años con los registros de estos fenómenos naturales. Al realizar el trabajoso análisis descubrió que a flujos más grandes del promedio normal eran seguidos por flujos todavía más grandes. Inesperadamente el proceso cambiaba a flujos menores que el promedio y eran seguidos por flujos aun menores que los anteriores. Parecían ciclos pero no eran periódicos.

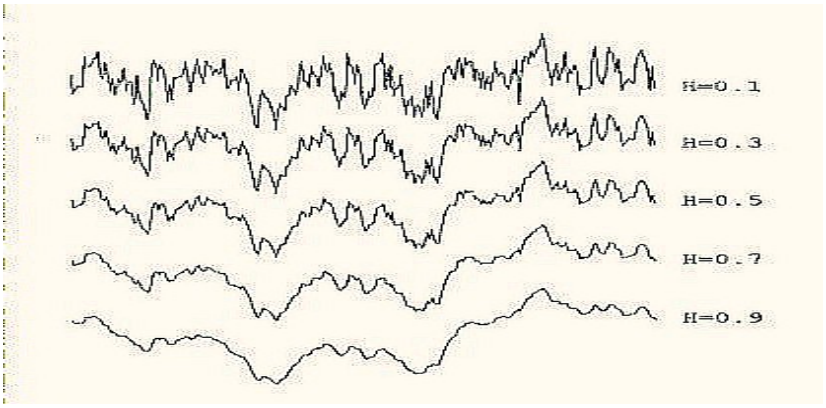
La información de la diferencia entre el valor del flujo máximo (Xmax) acumulado del río y el mínimo (Xmin), a lo largo de un periodo de tiempo X(t), le sirvió para entender la tendencia del flujo del río Nilo entre los periodos de sequía y lluvia (Gráfica 8), y así manejar el caudal y las crecidas del río en la construcción de la represa. Logrando desarrollar un método estadístico que muestra la dependencia a largo plazo de una serie de tiempo, conocido como el análisis del rango reescalado (análisis R/S).



Gráfica 8: Xmax y Xmin de la suma de desviación estándar respecto a la media

Fuente: (Feder, 1988)

Interpretación del resultado del exponente de Hurst. El exponente Hurst es usado generalmente como medida de complejidad en series de tiempo de largo plazo. Al aplicarse a un conjunto de datos proporciona una medida que indica si se guarda o no una relación entre ellos. Si los datos no guardan ninguna relación se dice que su complejidad es alta debido a la ausencia de correlaciones, lo que se presenta en comportamientos aleatorios; pero si los datos guardan una relación entonces su complejidad es baja, indicando persistencia, lo que muestra un comportamiento correlacionado (Gráfica 9).



Gráfica 9: Movimiento Fraccional Browniano y el exponente de Hurst

Fuente: (Saupe y Heinz, 1988)

El exponente de Hurst arroja resultados que están entre 0 y 1, se pueden interpretar de la siguiente manera (Mandelbrot y Hudson, 2004):

0 >= H <= 0.5: Series de tiempo con un "comportamiento antipersistente" o no correlacionado. Se presenta en series temporales en las que un periodo de crecimiento es seguido por uno de decrecimiento, consiste en la tendencia a regresar continuamente al lugar de procedencia. La antipersistencia aumenta a medida que H tiende a cero.

H = 0.5: Indica ausencia de correlaciones en series de tiempo completamente aleatorias. Los incrementos son independientes y su correlación es cero ya que para algún elemento actual tiene una correlación nula (cero) con un elemento futuro o con cualquiera de los instantes anteriores o posteriores.

0.5 >= H <= 1: "comportamiento persistente" o correlacionado, indica que la serie temporal tiene una tendencia. Se presenta en series temporales en las que un periodo de crecimiento es seguido de otro análogo. Cuanto mayor sea el valor de H, más fuerte es la tendencia.

El exponente de Hurst esta relacionado directamente con la "dimensión fractal euclidiana", dando una medida de la rugosidad (frecuencia de los valores de una variable) de una superficie dada (Gráfica 9). La dimensión fractal se ha utilizado para medir la rugosidad de las costas (Mandelbrot, 1993), por ejemplo. La relación entre la dimensión fractal (DF) y el exponente de Hurst (H) es: DF=2–H. La serie de tiempo será más irregular (aleatoria) si su dimensión fractal (DF) se acerca a 2.

2.3.6.2 Análisis fractal bidimensional

Havlin y Ben-Avraham (Havlin y Ben-Avraham, 1982) propusieron una Dimensión Fractal (DF) para describir y cuantificar la configuración de un típico polímero regular simple. Analizaron un polímero como una caminata vista a diferentes escalas de longitud. Un polímero formado por N_o monómeros podría ser recorrido en N pasos, es decir, monómero por monómero, cada dos monómeros etc. Estos investigadores desarrollaron el concepto de DF para un polímero en concordancia con las ideas desarrolladas por Mandelbrot (1977). En 1995, Xiao y colaboradores presentaron el concepto de DF para secuencias de ADN en espacios métricos bi y tridimensionales a partir de las metodologías propuestas por Hamori (1985) y Gates (1986) y retomando el análisis conceptual y procedimiento matemático desarrollado por Havlin y Ben-Avraham en 1982.

En este proyecto de investigación, se adaptaron las metodologías de los estudios anteriormente mencionados para elaborar un predictor bidimensional de exones en secuencias de ADN de un genoma eucariote. Cabe resaltar que las secuencias de ADN son biopolímeros formados por cuatro diferentes tipos de monómeros, los nucleótidos: Adenina, Guanina, Timina y Citosina. A continuación, se realiza una breve exposición de las ideas centrales de estos autores en relación con los conceptos y cálculos de la DF y la Dimensión Fractal Local (DFL).

Havlin y Ben-Avraham (1982) definen la DF como una medida que representa la magnitud de rugosidad de un polímero. Encuentran que la configuración simple de una cadena de monómeros posee propiedades autosimilares estadísticas, lo que actualmente se conoce como autoafinidad. La dimensión fractal se comporta como una cantidad macroscópica en el sentido que su incertidumbre desaparece a medida que el sistema se incrementa. En contraste, la distancia de extremo a extremo se comporta como una cantidad microscópica ya que las fluctuaciones relativas permanecen finitas incluso cuando $N_o \rightarrow \infty$.

El procedimiento para el análisis bidimensional consiste en identificar un fragmento de longitud a_o en un polímero y contar el número b_o de detalles en este fragmento; posteriormente, amplificar el mismo fragmento por un factor de amplificación a_1/a_o y contar los b_1 detalles. La ecuación sería:

$$(b_1/b_o) = (a_1/a_o)^D \qquad (2.1)$$

En donde, a_o es la longitud del fragmento a estudiar; b_0 son los detalles del fragmento a_o; b_1 es el detalle de la sección amplificada por el factor a_1/a_0.

D es independiente del factor de amplificación (a_1/a_o) para la mayoría de las longitudes de ese rango de escala (por ejemplo, el rango a_o), entonces el valor de D es denominado DF del polímero en ese rango. La suposición que D es independiente de a_o expresa la propiedad de autosimilaridad del polímero a diferentes escalas de longitud. Sí el polímero es una línea recta, entonces D=1, mientras que para una cadena aleatoria ideal, es fácilmente demostrable que D=2 (Havlin y Ben-Avraham, 1982).

La ecuación (2.1) es útil para definir la noción de dimensión fractal local (DFL, $D(a_o)$), de acuerdo a la relación:

$$(b_1/b_o) = (a_1/a_o)^{D(ao)}, \qquad (a_1-a_o)/a_o << 1 \qquad (2.2)$$

La $D(a_o)$ difiere de la D de la ecuación (2.1) en que esta puede ser definida por cualquier valor de a_o (la palabra local en DFL se refiere a la escala local de longitud). En contraste, la DF esta definida y coincide con $D(a_o)$ solamente cuando $D(a_o)$ no depende de (a_o) para un rango infinito de valores de (a_o). El promedio de las distancias al cuadrado de cada N pasos esta definida por la siguiente ecuación (Havlin y Ben-Avraham, 1982):

$$\langle R_N^2 \rangle_{No} = \frac{1}{No - N + 1} \sum_{i=1}^{No-N+1} \langle R_{i,i+N}^2 \rangle_{No} \qquad (2.3)$$

Donde $\langle R^2_{i,i+N} \rangle_{No}$ son las distancias al cuadrado entre el i-esimo y los (i+N)-esimos elementos que constituyen un polímero. De la ecuación (2.2) se puede obtener el valor de la DFL:

$$D_{No}(N) = \ln\left(\frac{N+1}{N}\right) / \ln\left(\frac{\langle R^2_{N+1} \rangle_{No}}{\langle R^2_N \rangle_{No}}\right)^{1/2} \qquad (2.4)$$

En donde, N_o esta considerando polímeros finitos. $D_{No}(N)$ es la DFL asociada con una escala local de longitud correspondiente a N monómeros (la cadena completa consiste de N_o monómeros). $\langle R^2_N \rangle_{No}$ son las distancias medias de un fragmento de N_o monómeros separados por N pasos. La cantidad $\left(\langle R^2_{N+1} \rangle_{No} / \langle R^2_N \rangle_{No}\right)^{1/2}$ juega el papel de factor de amplificación a_1/a_o presentado en la ecuación (2.2). Se espera que b_o sea proporcional a N y que b_1 sea proporcional a N+1.

La DFL es una medida de la rugosidad de un polímero a una cierta escala N.

Ahora, sí $D_{No}(N)=D$ es una cantidad independiente de N, entonces la ecuación (2.4) es equivalente a (Havlin y Ben-Avraham, 1982):

$$\left[\left(\langle R^2_N \rangle_{N_o}\right)^{1/2}\right]^D = AN \quad 1 \ll N \ll N_o \qquad (2.5)$$

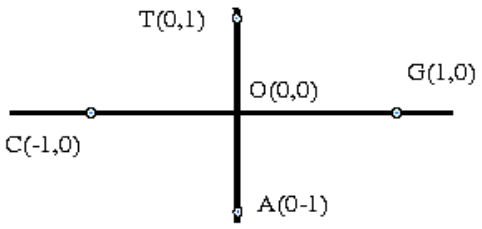
Esta expresión significa que si tenemos un sistema con un tamaño $L = (\langle R^2_N \rangle_{No})^{1/2}$, entonces el número de segmentos necesarios para cubrir el sistema es proporcional a la D-ésima potencia de L. Luego, de acuerdo con la definición de dimensión fractal de Haursdorff, D es la dimensión fractal y existe auto-similaridad en esta escala de longitud de la secuencia de nucleótidos (Xiao et al., 1995). La DF se calcula, según la ecuación (2.5), graficando $ln(N)$ *vs* $ln(\langle R^2_N \rangle_{No})^{1/2}$; obteniendo la pendiente de la distribución de puntos R^2_N hasta longitud de la cadena completa.

Los modelos de representación de las secuencias de ADN tienen como fin facilitar el análisis reduciendo significativamente la cantidad de registros o datos pero sin perder legibilidad en la información. El modelo 3D de Hamori (1985) permite un análisis tanto global como local de las características de secuencias muy largas.

La representación de Hamori (1985) es muy costosa computacionalmente y los resultados dependen de dicha capacidad, por eso Gates (1986) propuso una representación simple en dos dimensiones. Esta representación en dos dimensiones provee una información de patrones globales relacionados con las secuencias, como homologías,

estructuras repetitivas, abundancias de bases relacionadas, caminos probables de evolución y divergencia evolutiva. Un análisis en estructuras (de grano más fino), es decir, en cuanto a su composición y distribución de bases, tiene una diferencia en las características de intrones y exones en organismos eucariotes.

La dimensión fractal propuesta por Gates (Gates, 1986) describe de forma cuantitativa las características de la secuencia de nucleótidos al ser representadas en un plano cartesiano. En la Gráfica 10 se muestra que a cada nucleótido se le asigna un vector fijo en el espacio métrico y se construye una representación uniendo estos vectores de inicio a fin. Cada nucleótido esta representado por coordenadas en el plano:



Gráfica 10: Disposición de los nucleótidos en el espacio métrico

Según Gates (1986) la dimensión teórica fractal (Dtf) de una secuencia de ADN se define como el cociente del logaritmo de la longitud total del camino (número de pares bases en la secuencia) y el logaritmo de la distancia desde el origen al punto final:

$$Dtf = \ln[n(A)+n(C)+n(G)+n(T)] / \ln[\,|\, n(G)-n(C) \,|\, + \,|\, n(T)-n(A) \,|\,] \qquad (2.6)$$

donde n(X) es el número de bases X en la secuencia de ADN.

La anterior dimensión teórica no es muy informativa, pues sólo tiene en cuenta una distancia que va desde el inicio al punto final en el espacio métrico, y no refleja los detalles de conformación y autosimilaridad de toda la secuencia (Xiao et al., 1995).

2.3.6.3 Ley de Zipf-Mandelbrot

El código genético podría describirse como un código del lenguaje escrito, en donde podrían emplearse como abecedario: los cuatro diferentes nucleótidos o los 64 diferentes codones que constituyen las cadenas proteicas o las diferentes longitudes de exones e intrones que conforman un gen (Mandelbrot, 1993).

El estudio de las secuencias del ADN genómico es de fundamental importancia dado que toda la información de la herencia esta contenida en estas macromoléculas. Pero ¿cómo esta almacenada la información? Esta pregunta esta aún por resolver. Un punto clave son las relaciones de las bases nitrogenadas en las secuencias codificantes y no codificantes del ADN.

Las leyes empíricas de Zipf (Zipf, 1932; 1949) y Zipf-Mandelbrot (Mandelbrot, 1954) son medidas lingüísticas y estadísticas, de la teoría de la Información y de la geometría fractal, que miden la estructura, la organización y la riqueza de los lenguajes naturales.

a) La ley de Zipf (LZ)

El profesor George K. Zipf (1902-1950) realizó estudios sobre el comportamiento estadístico de la distribución de las palabras en un texto (Zipf, 1932; 1949). Zipf formuló la existencia de un principio de mínimo esfuerzo en los lenguajes naturales en el sentido que las palabras compuestas con un menor número de letras son usadas más a menudo por los hablantes de un lenguaje. El consideró este principio aplicable no sólo a los sonidos del habla sino también a otros elementos del lenguaje, especialmente a las palabras.

Zipf (1949), desarrolló una organización jerárquica para las frecuencias de aparición de las palabras en un texto, asignando rangos a estas frecuencias y haciendo corresponder a la mayor frecuencia el menor rango, lo que muestra un comportamiento de ley de potencia del fenómeno. Identificaba una palabra en particular y le asignaba un índice S igual al lugar o rango de la palabra en el listado y un P(s), que es la frecuencia de la repetición de esa palabra, es decir, el número de veces que aparece la palabra en el texto. La Ley de Zipf, sostiene la siguiente relación matemática:

$$P(\delta) = A(\delta)^{-\alpha},\ A=P(1);\ \alpha = 1;\ \delta = 1,2,3\ldots \qquad (2.7)$$

La relación matemática señala que existe una relación lineal e inversa entre los valores de una variable estudiada y la frecuencia de su aparición, relación denominada rango-tamaño (rank-size). El rango (δ) multiplicado por la frecuencia $P(\delta)$ es una constante (A) y α es un valor cercano a 1. Esta relación, es cierta, como un aspecto general de la lingüística, es independiente de los usuarios de la lengua y los tipos de textos y los idiomas. La Ley de Zipf puede explicar el aparente equilibrio entre uniformidad y diversidad en nuestro uso de las palabras.

La ley de Zipf proporciona una descripción sencilla de la organización jerárquica de un sistema, cuyos parámetros (δ, α y A) se pueden utilizar como indicadores para describir su evolución.

b) La ley de Zipf-Mandelbrot (LZM)

En 1983, Mandelbrot generalizó la Ley de Zipf y definió la dimensión fractal estadística a partir de una linealización logarítmica de las frecuencias y de los rangos calculando el inverso de la pendiente de esta linealización. Usando ideas de la teoría de la información explicó la ley de rango-tamaño, considerando el costo de comunicación de las palabras en términos de las letras que contienen y de los espacios que las separan. Este costo aumenta con el número de letras en una palabra y por extensión en un mensaje.

Mandelbrot (1983), propuso la siguiente relación entre el rango y la frecuencia (Ley de Zipf-Mandelbrot (LZM):

$$P(\delta+V) = A(\delta+V)^{-1/\alpha}, \qquad A, V, \alpha > 0 \tag{2.8}$$

donde:

- P es la probabilidad o frecuencia de una palabra.
- $(\delta+V)$ es el rango de una palabra con probabilidad P.
- $1/\alpha$: dimensión fractal, es un parámetro de distribución e indica que tan rápido decae la frecuencia.
- $0<\alpha<<1$: es una medida de la riqueza del vocabulario, el texto es fractal.
- $\alpha \geq 1$: número finito de palabras en un diccionario
- $V = 1/(N-1)$, $0<V<10$; número de letras distintas en el vocabulario.
- A es una constante de escala que se fija experimentalmente.

La ley de ZM es una generalización, es decir, a partir de la observación empírica se buscan generalizaciones, patrones o leyes matemáticas que permitan resumir la información estudiada.

Las leyes, de Zipf y Zipf-Mandelbrot tienen un carácter universal debido a la cantidad de aplicaciones en todos los campos de estudio concebibles: procesos biológicos(Burgos y Moreno, 1996) ; distribución del tamaño de ciudades (Gabaix, 1999), distribución en Redes de Internet (Breslau et al., 1999), entre otros.

2.4 Bases de datos genómicas

En las bases de datos bioinformáticas se encuentran las secuencias de nucleótidos, aminoácidos y anotaciones de los genomas que los investigadores buscan utilizar con el fin de contrastar información de manera practica eficiente y rápida. Estas bases de datos han permitido acercar a investigadores de diferentes áreas del conocimiento: biólogos moleculares, bioquímicos, matemáticos, físicos e ingenieros, particularmente ingenieros de sistemas cuyo trabajo transversal ha contribuido al análisis y desarrollo de la era post-genómica (Biotech Magazine, 2007).

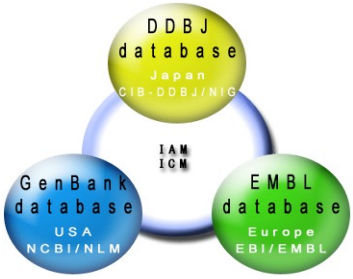
Según Cuello et al. (2005) las principales bases de datos oficiales son:

- GenBank: base de datos de Estados Unidos, mantenida por el Centro Nacional de Información de Biotecnología (NCBI), y por la Biblioteca Nacional de Medicina (NLM).

- EMBL (European Molecular Biology Laboratory): base de datos europea, mantenida por el Laboratorio Europeo de Biología Molecular (EMBL) y el Instituto Europeo de Bioinformática (EBI).

- DDBJ (DNA Databank of Japan): base de datos de Japón, mantenida por el Centro de Información de Biología y Banco de Datos de ADN del Japón (CIB-DDBJ) y el Instituto Nacional de Genética (NIG).

En 1980 las bases de datos llegaron a un acuerdo para utilizar el mismo formato y compartir automáticamente datos e información; lo que indica que las tres bases de datos comparten prácticamente la misma información en un momento determinado. La integración es coordinada por la Junta Internacional de Consulta (IAM) y la Junta Internacional de Colaboración (ICM), como se muestra en la Gráfica 11 (Cuello et al., 2005).



Gráfica 11: Bases de datos oficiales.

Fuente: http://www.ddbj.nig.ac.jp/intro-e.html

Algunos de los inconvenientes de las bases de datos mencionadas son las redundancias, los errores y la falta de información sobre algunas secuencias, debido a la información proporcionada por los investigadores y que el tratamiento automático no puede eliminar o corregir, por lo tanto surgen nuevas bases de datos "curadas" de los errores mencionados, una de ellas es ENSEMBL que cuenta con múltiples referencias a otras bases de datos y una serie de herramientas que facilitan análisis mas elaborados sobre los datos, entre otras ventajas.

ENSEMBL, es un proyecto conjunto entre EMBL y Wellcome Trust Sanger Institute (WTSI) para desarrollar sistemas software que produzcan y mantengan automáticamente la anotación de determinados genomas eucariotes. Es principalmente financiada por la Wellcome Trust.

La información y software generados del proyecto son de libre uso y acceso, se pueden restringir las consultas usando criterios tales como: secuencias codificantes, regiones promotoras, genes, proteínas, entre otros; adicionalmente, comparar esta información entre genomas de varias especies.

2.4.1 Formatos de las bases de datos oficiales

Los archivos más usados en las Bases de Datos Oficiales son:

- .faa = archivo Fasta de secuencias de aminoácidos.
- .ffn = archivo Fasta de secuencias de nucleótidos para regiones codificantes.
- .fna = archivo Fasta de secuencias de ácidos nucleidos completos (codificantes y no codificantes).
- .gbk = archivo plano de secuencias y su información en formato GenBank.
- .gbs = archivo resumen del formato GenBank.
- .ptt = archivo de la Tabla de proteínas.
- .tar.Z = archivo Unix empaquetado y comprimido (no todos los archivos son comprimidos), donde todos los archivos anteriormente mencionados están presentes, entre otros.

La información relevante para este proyecto esta en las secuencias de nucleótidos (codificantes y no codificantes) y en la información general respecto al genoma como posiciones de genes, exones e intrones (anotaciones del genoma) presentes en los archivos gbk. Se utilizaron en este proyecto los formatos GenBank y Fasta, a continuación se presenta una breve descripción.

GenBank presenta una visión centralizada de los registros de las secuencias. Una secuencia en formato gbk o GenBank inicia con información de la secuencia de ADN y luego la secuencia como tal (Gráfica 12). Los identificadores más utilizados de este formato son:

- LOCUS: Inicio de archivo gbk
- DEFINITION: Indica el organismo y cromosoma que contiene el archivo.
- /gene=: Símbolo para representar los genes, el valor a continuación será el identificador del gen.
- gene: indica la posición del gen dentro de la secuencia. Ejemplo: gene 11721..14685
- tRNA: Un gen que codifica una RNA de transferencia.
- rRNA: Un gen que codifica una ARN ribosomal.
- misc_RNA. Un gen que codifica otras clases de ARN.
- CDS: Secuencia que codifica para una proteína.

- ORIGIN: Indica el inicio de la secuencia.
- //: final de la secuencia.

La importancia del archivo gbk radica en que describe las posiciones y longitudes de secciones especiales respecto al genoma como genes, exones e intrones tal como se encontró en los laboratorios.

Se analizaron porciones de nucleótidos del genoma en archivos independientes con formato Fasta para facilitar la manipulación de las secuencias, mayor rapidez de procesamiento y organizar el análisis.

```
LOCUS       NT_039169           19423349 bp    DNA     linear   CON 20-JUN-200
DEFINITION  Mus musculus chromosome 1 genomic contig, strain C57BL/6J.
ACCESSION   NT_039169 REGION: 1..19423349
VERSION     NT_039169.7  GI:149233633
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
FEATURES             Location/Qualifiers
     source          1..19423349
                     /organism="Mus musculus"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:10090"
                     /chromosome="1"
     gene            complement(204563..661579)
                     /gene="Xkr4"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BestRefseq. Supporting evidence
                     /db_xref="GeneID:497097"
     mRNA            complement(join(204563..207049,411783..411982,
                     660633..661579))
     CDS             complement(join(206103..207049,411783..411982,
                     660633..661429))
     gene            1511779..1516818
                     /gene="LOC664830"
                     /note="Derived by automated computational analysis using
                     gene prediction method: GNOMON. Supporting evidence
                     includes similarity to: 2 ESTs, 4 Proteins"
                     /pseudo
                     /db_xref="GeneID:664830"
     misc_RNA        join(1511779..1511789,1512909..1513662,1514495..1516818)
ORIGIN
        1 gaattctttt ctatgattta gtttaatatg ttttctgggt gtttcagctg aaacttttgt
       61 ccttctttta ttcctatttt tcttaggttt ggtcttttca tagtgtccca gatttcctgg
      121 ctgtttcttg ttaggatttt tttagattta acatttctgt catagattaa tctattttgc
      181 agatgtaatt gtatgtatta taattgtaat agtatatact tgtatgtact taaaatattt
//
```

Gráfica 12: Formato gbk de Ensembl, Cromosoma 1 de *M. musculus*

Un archivo en formato fasta inicia su definición con el símbolo “>” seguido de un código de identificación y el resto de la línea, hasta fin de linea, tiene la descripción de la secuencia, esta puede ser opcional y su información varia de acuerdo a las necesidades del investigador, las otras líneas tendrán los caracteres de la secuencia. La secuencia termina cuando la siguiente línea comienza con un ">", lo que indica el inicio de un nuevo registro (Gráfica 13).

```
> ENSMUSG00000058248 | 1 | Exon | lon: 303 1001 1303 protein_coding
GGCGCACCGAGCCCGGGGCACTCGAGCGCACACCCGGGCGGGCCGCAGCGGGGAGGTGCGGAGCGCAGCG
AGGGCGGCAGCGGGAGCCAGCCGCCGCCCCTGCGCTCCGGGGCTGCGGAGCGCATGGGGCGCGTGGAGCC
AGGAACGCTTGCGGGCTGCGCACCCGGCCGCCTGCTGCGGTGAGACACGGCGCCGGACGCCCCCAGAGTC
CGGCAGTCGGGAGGATGACCATGGCTGGCGGCCGGCGGGGACTAGTGGCCCCGCAGAACACGTTTCTGGA
GAACATCGTGCGGCGGTCCAACG
> ENSMUSG00000058248 | 1 | Exon | lon: 124 31550 31673 protein_coding
ACACTAATTTTGTGTTGGGGAATGCCCAGATAGTCGACTGGCCCATCGTGTACAGCAATGATGGATTCTG
CAAGCTGTCTGGCTACCACCGGGCAGAAGTGATGCAGAAAAGCAGCGCCTGCAG
> ENSMUSG00000058248 | 1 | Exon | lon: 107 35062 35168 protein_coding
TTTTATGTATGGAGAGCTGACTGACAAGGACACAGTTGAAAAGGTTCGCCAGACCTTTGAGAACTACGAG
ATGAATTCCTTCGAAATTCTGATGTACAAGAAGAACA
> ENSMUSG00000058248 | 1 | Exon | lon: 129 48942 49070 protein_coding
GGACACCTGTGTGGTTTTTTGTGAAGATCGCTCCGATCAGGAACGAACAGGATAAAGTGGTCCTGTTCCT
TTGCACTTTCAGCGACATAACTGCATTCAAGCAGCCCATCGAGGACGACTCCTGCAAAG
> ENSMUSG00000058248 | 1 | Exon | lon: 119 52136 52254 protein_coding
GTTGGGGGAAGTTTGCCCGACTGACCAGAGCTCTGACAAGCAGCAGGGGAGTCCTGCAGCAGCTGGCCCC
CAGTGTACAGAAGGGTGAGAATGTTCACAAGCACTCGCGCCTGGCAGAG
```

Gráfica 13: Formato Fasta del cromosoma 1 de *M. musculus*

2.5 Minería de Datos

La Minería de Datos (Data Mining) es la búsqueda de correlaciones, patrones y tendencias en grandes volúmenes de datos. Apoya la toma de decisiones ya que reúne varias ventajas de la Estadística, la Inteligencia Artificial mediante el Aprendizaje de Máquina y el procesamiento masivo de las bases de datos. La Minería de Texto (Text Mining), es un caso particular de la Minería de Datos, que se enfoca en el descubrimiento de patrones y nuevos conocimientos no explícitos en conjuntos de datos planos, que surgen de relacionar el contenido de varios de ellos, es decir, su objetivo es descubrir aspectos tales como tendencias, desviaciones y asociaciones entre volúmenes de información textual. Por lo tanto este trabajo se desarrolló siguiendo la metodología de la minería de datos pero con datos en texto plano. La Minería de Datos comprende la siguientes etapas (Gráfica 14) (Chapman et al., 2000):

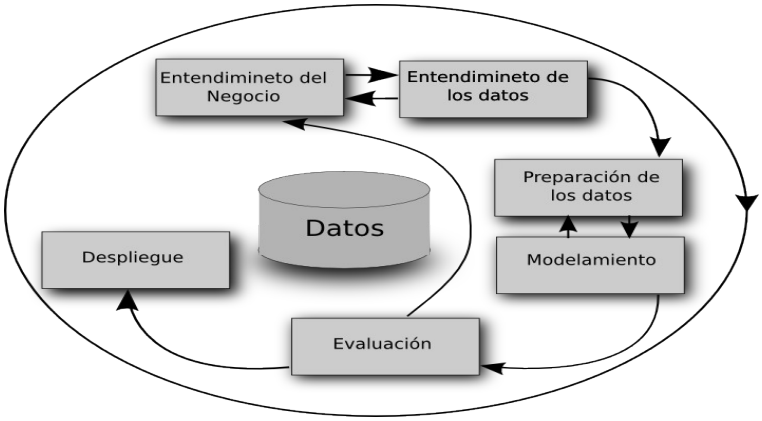
Entendimiento del negocio: conocer los objetivos y requerimientos del proyecto desde la óptica del negocio y definir el problema de Minería de Datos, además de generar un plan para alcanzar dichos objetivos.

Entendimiento de los datos: Conocer como esta almacenada la información en la colección o Base de Datos y generar actividades que permitan mejorar su calidad, obtener nuevos subconjuntos aptos para los análisis posteriores.

Preparación de los datos: actividades necesarias para transformar los datos y generar el conjunto de datos que van a servir de entrada al modelo final. En esta etapa se hace un análisis de rendimiento, se seleccionan los atributos o campos necesarios para el modelo e información que será usada para generar el modelo. “Se ha estimado que esta etapa toma el 60% de tiempo y esfuerzo invertido en todo el proceso de Minería de Datos” (Pyle, 1999).

Modelamiento: Selección y aplicación de diferentes técnicas de modelamiento aplicadas a los parámetros con el fin de obtener valores óptimos para el modelo. Si es necesario transformar los datos, se vuelve a la etapa anterior de preparación de datos.

Evaluación: Hacer una revisión de los diferentes modelos construidos, además de los pasos ejecutados en su construcción, y ver si cumple con el objetivo del modelo desarrollado o resuelve en parte el problema para el que fue implementado de lo contrario se podría rechazar el modelo o hacerle mejoras.



Gráfica 14: Etapas de la Metodología de Minería de Datos. Adaptado de: (Chapman et al., 2000)

2.6 Aprendizaje de Máquina

El Aprendizaje de Máquina se define como "un programa de computador que puede aprender desde la experiencia con respecto a alguna clase de tarea y ejecutar una medición" (Mitchell, 1997).

Existen tres categorías principales de aprendizaje:

- Supervisado o aprendizaje a partir de ejemplos: el instructor o experto define clases y provee ejemplos de cada una. El sistema debe obtener una descripción para cada clase. Cuando el experto define una única clase, provee ejemplos positivos (pertenecen a la clase) y negativos (no pertenecen a la clase). En este caso, los ejemplos más relevantes son los cercanos al limite, ya que proveen información útil sobre los limites de la clase.

- Por refuerzo o aprendizaje observando el mundo que lo rodea: la información de entrada al sistema es el feedback o retro-alimentación del mundo exterior en respuesta a la acción ejercida; algo parecido a prueba y error.

- No supervisado o aprendizaje a partir de observaciones y descubrimientos: el sistema debe agrupar los conceptos sin ayuda de un experto, sólo recibe ejemplos;

pero no hay información de las salidas correctas (clases). Por lo tanto, debe observar los ejemplos y buscar características en común que permitan formar descripciones tanto en conjunto como individuales.

Algunas técnicas de Aprendizaje de Máquina son:

- Descubrimiento de conocimiento en Bases de Datos (KDD): área que trata de extraer información útil y conocimiento oculto en las bases de datos. La Minería de Datos como un etapa del KDD aplica varios métodos para extraer patrones útiles en las bases o conjunto de datos planos. Toda esta información se interpreta y se evalúa con el fin de descubrir nuevo conocimiento (Tan y Gilbert, 2001).

- Redes Neuronales Artificiales (RNA): de manera semejante a las neuronas biológicas del cerebro humano; las RNA son un conjunto de unidades interconectadas, donde cada una de ellas toma valores finitos de entradas, que puede ser las salidas de otras unidades, y produce un valor de salida, que pueden servir de entrada a otras unidades (Mitchell, 1997). La variedad de modelos esta regida por la topología de red, las características de los nodos y las reglas de actualización de estado. La topología se refiere a la estructura de interconexiones entre los nodos (neuronas) en términos de capas y/o enlaces hacia atrás (redes recurrentes o cíclicas) o hacia adelante (redes con propagación hacia adelante o acíclicas). Las características de los nodos, se refiere a las operaciones que pueden ejecutar, tales como suma de pesos (mecanismo de cada nodo para ajustarse, en la fase de aprendizaje) de las entradas y luego amplificándolas. Las reglas de actualización pueden ser para los pesos y/o estados de los elementos procesados (neuronas). Una RNA puede ser vista como un grafo dirigido en el cual las neuronas son nodos y las aristas dirigidas (con pesos) son conexiones, entre neuronas, salientes y entrantes (Nilsson, 1997).

  Una de las implementaciones mas típicas es Perceptrón Multicapa (Multilayer Peceptron) que usa como algoritmo de aprendizaje supervisado backpropagation (propagación inversa) donde se ajustan los pesos de acuerdo a un porcentaje de error resultante de la evaluación entre la salida predefinida y la que arroja la red. Los pesos que se ajustan están en las capas ocultas de la red, esto dificulta tener control sobre los ajustes hechos ya que son realizados de forma automática.

- Árboles de decisión (AD): son árboles cuyos nodos internos son pruebas (patrones de entrada) y cuyos nodos hoja son categorías o clases. Un AD asigna una clase (o salida) a un patrón de entrada filtrando dichos patrones a través del recorrido de las pruebas en el árbol. Un AD puede ser representado como un conjunto de reglas de la forma: si cumple con cierto valor dentro de un rango; entonces ejecutar un proceso determinado; sino ejecutar proceso alterno (***if-then-else***); demostrando su capacidad para dividir un proceso complejo en una colección

de procesos más simples, y por lo tanto haciendo la solución mas interpretable. Un AD ayuda en la exploración de datos de la siguiente manera:

- Reduce el volumen de datos por la transformación más compacta, que preserva las características esenciales y provee una síntesis más precisa.
- Descubre si los datos contienen clases bien separadas, tal que puedan ser interpretadas en el contexto del análisis realizado.
- Mapear los datos en forma de árbol permite que los valores predichos puedan ser generados a la inversa, desde las hojas hasta la raíz. Esto puede ser usado para predecir las salidas de un nuevo dato o consulta. También permite que el algoritmo sea más legible y transparente.

El algoritmo, ID3, permite construir árboles de arriba hacia abajo, iniciando con la pregunta: ¿qué atributo debe ser probado en la raíz del árbol? Para responder esto, cada atributo se evalúa usando una prueba estadística para determinar que tan bien, por si solo, clasifica los ejemplos de entrenamiento. El mejor atributo se selecciona y se usa como prueba en el nodo raíz. Luego se crea un descendiente del nodo raíz para cada posible valor de este atributo, y los ejemplos de entrenamiento se ordenan en el respectivo nodo descendiente. Este proceso se repite usando ejemplos de entrenamiento asociados con cada nodo descendiente para seleccionar el mejor atributo a probar en dicho punto del árbol (para más detalles del algoritmo ver Mitchell, 1997, pp. 55, 56).

El algoritmo C4.5, son mejoras hechas al ID3; principalmente poda el árbol de decisión, no dando tanta importancia a demasiadas variables (para más detalle ver Mitchell, 1997, pp. 66-72).

- Modelos Ocultos de Markov (HMM): un HMM de primer orden es un modelo estocástico (aleatorio) para series de tiempo definidas por un conjunto finito de estados, un alfabeto discreto de símbolos, una matriz probabilística de transición $T=(t_{ji})$, una matriz probabilística de emisión $E=(e_{iX})$. Los sistemas aleatorios evolucionan de un estado a otro mientras emiten símbolos del alfabeto. Cuando un sistema esta en un estado dado *i*, tiene una probabilidad t_{ji} de moverse al estado *j* y una probabilidad e_{iX} de emitir el símbolo X. Así, un HMM puede ser visto como dos dados diferentes asociados con cada estado: un dado de emisión y un dado de transición. El supuesto esencial de los HMM de primer orden es que las emisiones y transiciones dependen sólo del estado actual, y no del anterior. Sólo los símbolos emitidos por el sistema son observados, no el camino recorrido entre los estados; de ahí su apelativo de oculto. Los caminos aleatorios ocultos pueden ser vistos como ocultos o variables latentes bajo observación (Nilsson, 1997).
- Red Bayesiana (Bayes Network - BN): es un grafo de probabilidades donde cada nodo representa un suceso conocido con cierta posibilidad de ocurrir y las aristas

que unen los nodos representan la dependencia condicional entre un suceso y otro, así las BN pueden inferir sucesos futuros con base en las probabilidades de los sucesos anteriores (Tan y Gilbert, 2001).

2.7 Modelos de predicción de genes/exones

Entendiendo el modelo como una abstracción de las principales características del mundo real o problema, se puede tratar de definir un modelo de exones: abstracción de la realidad, en este caso exones e introns de una molécula de ADN, para obtener mediciones matemáticas y estadísticas sobre los fenómenos que en ellos ocurren, que de otra manera no se pudieran observar directamente.

Para probar los modelos compuestos por medidas estadísticas, fractales y su combinación, se acude a muchos métodos y técnicas para su implementación y así observar y medir, de acuerdo a ciertos estímulos, los resultados obtenidos. Algunos de estos métodos y técnicas los proporciona el área de Aprendizaje de Máquina como: bayes, redes neuronales, árboles de decisión, algoritmos genéticos, cadenas de Markov, programación dinámica, SVM (Máquinas de Soporte Vectorial) entre otras (Baldi y Brunak, 2001).

Los diferentes programas de predicción de genes, como GRAIL, GenScan o Morgan, usan combinaciones de métodos para lograr una mayor precisión en sus inferencias. A continuación se presentan ciertas características de los modelos implementados por algunas de las herramientas más usadas.

Modelo implementado por GRAIL, basado en Redes Neuronales Artificiales

La aplicación de las redes neuronales se ha extendido en muchos campos, entre ellos la Bioinformática. Entre variedad de herramientas de identificación de genes/exones para organismos eucariotes podemos encontrar GRAIL (Uberbacher y Mural, 1991), una de las primeras en usar redes neuronales para el reconocimiento de genes/exones en secuencias de ADN. Esta usa una red neuronal que combina una serie de algoritmos predictivos para reconocer zonas potencialmente codificantes con ventanas de longitud fija sin buscar características adicionales. Luego GRAIL II (Uberbacher et al., 1996) presenta algunas mejoras: el sistema considera zonas discretas como regiones codificantes, usa una ventana deslizante de tamaño fijo para evaluar zonas potencialmente codificantes, así como en la mejora del tiempo de procesamiento.

Modelo implementado por GenScan, basado en modelos ocultos de Markov

Las secuencias biológicas, así como el habla, puede ser modelados como la salida de un proceso que pasa a través de una serie de estados discretos, algunos de los cuales son "ocultos" al observador, y por lo tanto puede ser una solución al problema de búsqueda de nuevas secuencias codificantes en secuencias de ADN. En este modelo los introns y

exones internos se dividen de acuerdo a la fase, es decir, si un intrón coincide exactamente con un codón se considera fase 0, si coincide después de la primera base de un codón es fase 1, y después de la segunda base de un codón esta en fase 2 (Henderson et al., 1998).

Modelo implementado por Genezilla, basado en modelos ocultos de Markov generalizados

Similar a GenScan; pero adicionalmente tiene la característica que es bastante configurable e incluye software de entrenamiento para el usuario final, gracias a que es Open Source, el usuario puede disponer de él según sus necesidades. El programa permite predecir genes de distintos organismos eucariotes, ya que se puede entrenar con las secuencias del organismo específico y realizar los ajustes a la configuración de los parámetros del software para mejorar el proceso de predicción (Majoros et al., 2004; 2005).

Modelo implementado por Morgan, basado en árboles de decisión

Morgan (Multi-frame Optimal Rule based Gene Analyzer) es un sistema integrado para la búsqueda de genes en vertebrados, entre las técnicas que usa se encuentran los árboles de decisión. Los nodos internos del árbol de decisión poseen valores de características que son testeados para cada subsecuencia pasada por el árbol. Las características pueden ser varias medidas de las regiones codificantes o señales fuertes. Los nodos hoja del árbol contienen las etiquetas de la clase que se asociarán a una secuencia secundaria. Una vez clasificado, varios de los componentes son ensamblados dentro de un modelo óptimo de genes usando un método de programación dinámica (Claverie, 1997).

2.7.1 Evaluación del modelo de predicción de exones

Los programas de predicción de genes/exones usan varias técnicas computacionales, conjuntos de datos de entrenamiento, prueba o validación, y para medir y comparar la eficiencia con que lo hacen, se han definido unas medidas de evaluación, con base en probabilidades, para exones identificados correctamente en las secuencias de ADN (a nivel de secuencias codificantes) y los pares de bases como parte de las regiones codificantes (a nivel de nucleótidos) (Burset y Guigó, 1996) (Pita y Pértegas, 2003).

A nivel de modelos de predicción de genes/exones se cuentan los aciertos y desaciertos de una predicción sobre una secuencia de prueba, comparando el valor predicho (codificante o no codificante) con el valor real de cada secuencia, a través de todas las secuencias de prueba. El Cuadro 1 muestra la relación de aciertos y desaciertos de predicción arrojados por el predictor versus los reales:

VP (Verdaderos Positivos): número de secuencias codificantes que han sido predichas correctamente como codificantes. El predictor cataloga una secuencia de muestra como codificante cuando realmente es codificante.

VN (Verdaderos Negativos): número de secuencias no-codificantes que han sido predichas correctamente como no-codificantes. El predictor cataloga una secuencia de muestra como no-codificante cuando realmente es no-codificante.

FN (Falsos Negativos): número de secuencias codificantes que han sido predichas como no-codificantes. El predictor cataloga una secuencia de muestra como no-codificante cuando realmente es codificante (el predictor se equivoca).

FP (Falsos Positivos): número de secuencias no-codificantes que han sido predichas como codificantes. El predictor cataloga una secuencia de muestra como codificante cuando realmente es no-codificante (el predictor se equivoca).

Ejemplo: una secuencia del conjunto de datos de prueba es un exón (secuencia codificante) y en el momento de pasar por el predictor da como resultado que es un intrón (secuencia no codificante), esto es un desacierto y se cuenta como falso negativo (FN).

Cuadro 1: Variables de medición

		Realidad	
Predicción		codificante	no codificante
	codificante	VP	FP
	no codificante	FN	VN

La calidad de los modelos de genes/exones se puede calcular por medio de varias medidas. Sin embargo, las dos medidas ampliamente usadas en modelos como Grail, GenScan, Genezilla y Morgan, para medir la precisión en las predicciones son la **Sensibilidad** (**Sn**) y la **Especificidad** (**Sp**):

$$Sn = \frac{VP}{(VP + FN)} \qquad Sp = \frac{VN}{(VN + FP)} \tag{2.9}$$

La **sensibilidad** es la capacidad que tiene el predictor para clasificar correctamente secuencias codificantes y la **especificidad** es la capacidad del predictor para clasificar correctamente secuencias no codificantes. En términos de probabilidad, esto es, **Sn** es la probabilidad que una secuencia codificante sea catalogada en la predicción como codificante, y **Sp** es la probabilidad que una secuencia no codificante sea clasificada por el predictor como no codificante (Burset y Guigó, 1996) (Pita y Pértegas, 2003).

3. Metodología de investigación

Las Bases de Datos de este estudio están en los formatos Fasta y Gbk, que contienen la secuencia y anotación, respectivamente, de los genomas estudiados: *M. musculus, H. sapiens, C. elegans, S. cerevisiae, G. gallus, A. thaliana, O. sativa y D. melanogaster*. Los archivos se encuentran disponibles en las bases de datos del NCBI (http://www.ncbi.nlm.nih.gov) y ENSEMBLE (http://www.ensembl.org). La extracción, análisis de datos y aplicación de medidas estadísticas estándar, medidas fractales, se llevó a cabo en el lenguaje de programación Python.

3.1 Minería de Datos

3.1.1 Etapa de entendimiento del negocio

Uno de los objetivos importantes para los investigadores científicos es poder identificar regiones codificantes en el ADN, por tal motivo el objetivo de este proceso de Minería de Datos es buscar patrones en las secuencias del ADN que discriminen secuencias codificantes de las no codificantes. Para esto se necesitaron conocimientos fundamentales en biología molecular, estadística, fractales, aprendizaje de máquina y minería de datos, todos estos temas resumidos en el capitulo del marco teórico (**Capitulo 2**).

El problema de la efectividad de predicción de secuencias codificantes, se puede resolver intentando dar respuesta a las siguientes preguntas:

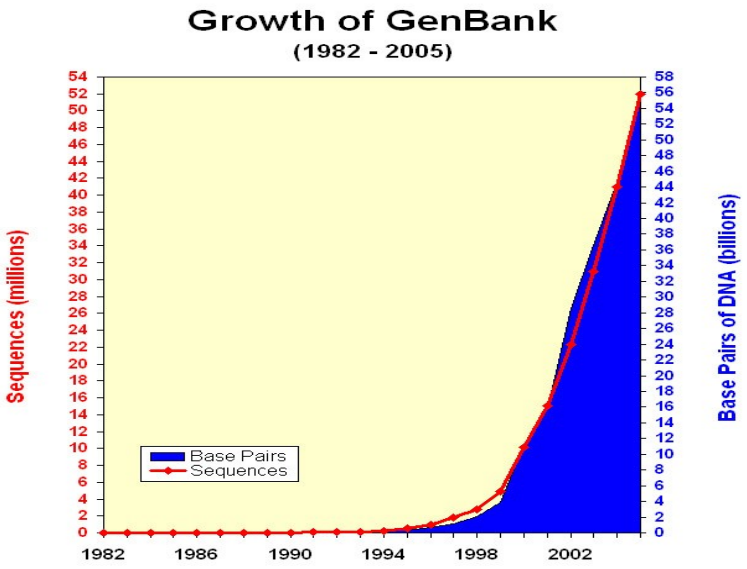
- ¿ Las medidas estadísticas son suficientes para la predicción de secuencias codificantes ?
- ¿ Las medidas fractales mejoran el promedio alcanzado por las medidas estadísticas ?
- ¿ Las medidas estadísticas y fractales son complementarias ?

En el mercado hay varios programas para crear modelos de clasificación como Clementine, Oracle Data Mining, SAS, Neuro Shell y Weka, este último se usó para crear el modelo de clasificación ya que tiene la ventaja de estar disponible de forma libre al público, además brinda una visualización y comprensión de datos, permite analizar con detalle los atributos ya que se pueden presentar valores redundantes o despreciables, visualizar la distribución de los datos, agrupar valores y presentar gráficas de dispersión de atributos numéricos y se pueden hacer combinaciones entre variables numéricas con el fin de establecer diferentes relaciones (Orallo, 2005).

3.1.2 Etapa de entendimiento de los datos

Las bases de datos genómicas contienen mas de 4.000 organismos secuenciados, mas de 52 millones de secuencias, mas de 56 mil millones de nucleótidos que conforman mas de 353 Gb en información (Gráfica 15). Las secuencias están disponibles en diferentes formatos entre ellos el Gbk donde se definen características como longitudes, posiciones, contenido de la secuencia y anotación de la secuencia. Los sitios oficiales de estas Bases Datos disponen de herramientas vía web, que permiten mostrar una información detallada de cada genoma o gen, además se puede hacer alineamientos Blast, entre otros análisis biológicos. Los archivos Gbk están disponibles vía FTP donde se pueden descargar. La secuencia completa de un genoma se encuentra dividido en varios archivos gbk comprimidos. Estos archivos están en texto legible con extensión .gbk .

Para la comprensión de los diferentes atributos descriptivos presentes en los archivos Gbk, se obtuvo el asesoramiento de Biólogos con conocimientos de bioinformática.



Gráfica 15: Estadística de la Base de Datos NCBI

Fuente:
http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

3.1.3 Etapa de preparación de los datos

En esta etapa se selecciona el conjunto datos, se seleccionan los atributos con cuales va a trabajar el modelo describiendo el procedimiento de obtención de cada uno y finalmente se hace una limpieza a los atributos, dejando los que realmente aportan información.

3.1.3.1 Selección de datos

Se diseñó un algoritmo en lenguaje Python para extraer las secuencias de nucleótidos en formato Fasta (secuencia de letras A, C, T y G) tanto para exones como para intrones; pero mucha información de los genomas consignada en los archivos gbk, no era relevante para este proyecto que sólo requería información biológica básica como longitud y contenido de las secuencias (secuencia de letras A, C, T y G). Por esta razón la información se filtró por conceptos básicos de la estructura del gen, tales como que todo gen comienza y termina por un exón, esto indica que la cantidad de exones, para genes interrumpidos, es igual a cantidad de intrones más uno, adicionalmente se presentan genes simples compuestos únicamente por un exón, bajo esta restricción se trabajó con genes interrumpidos que permitieran obtener información para discriminar exones de intrones, y son poco comunes en genomas eucariotes.

Algunas secuencias de los exones o intrones presentaban nucleótidos desconocidos, representados por la letra N, al encontrarse una secuencia con estas letras se omitió el gen, para preservar el equilibrio y robustez de los datos. Algunas secuencias demasiados cortas, menores a seis pares de bases, se descartaron las secuencias de exones e intrones ya que no aportaban información biológica relevante al análisis.

Los datos de los genomas trabajados (Cuadro 2) se presentan resumidos en el siguiente cuadro así, la especie humana (*H. sapiens*) tiene 23 cromosomas, se trabajó con todos ellos, los cuales contienen 19326 genes con 192.999 exones y 173.673 intrones, tomados de la base de datos Ensembl:

Cuadro 2: Generalidades de los organismos estudiados

Especie	**cromosomas**		**Genes**	**Exones**	**Intrones**	**Fuente**
	Reales	**Trabajados**				
H. sapiens	23	23	19326	192999	173673	Ensembl
M. musculus	21	16	15806	155724	139918	Ensembl
G. gallus	32	28	7888	66260	58372	Ensembl
D. melanogaster	6	6	11432	54836	43404	Ensembl
O. sativa	12	12	20323	127218	106895	NCBI
C. elegans	6	6	20066	124957	104891	Ensembl
A. thaliana	5	5	21218	136847	115629	NCBI
S. cerevisiae	16	16	283	575	292	Ensembl

3.1.3.2 Selección de atributos/medidas

Los atributos representan los criterios con los cuales el modelo tomará una decisión para clasificar secuencias codificantes. Las medidas que se aplicaron para la búsqueda de patrones pretenden establecer tendencias de comportamiento diferentes tanto para exones

como para intrones, obteniendo de cada medida uno o varios atributos descriptivos para el modelo de predicción de exones.

En este proyecto se consideraron criterios de predicción ***ab initio*** los cuales se basan en caracterizar cada señal de un gen de acuerdo a su estructura, contenido y longitud. En la exploración de los diferentes criterios ab initio para la búsqueda de patrones, se identificaron medidas estadísticas estándar usadas frecuentemente para el análisis de secuencias de ADN tales como uso de codón, ficket, hexámeros y contenido GC (según los protocolos de Guigó, 1998). Como el aporte de esta investigación fue explorar características fractales del ADN, se aplicaron medidas fractales usadas en teoría de la información: ley de Zipf-Mandelbrot (Mandelbrot, 1993); y sistemas complejos: exponente de Hurst (Yu y Chen, 1999) y dimensión fractal bidimensional (Xiao et al., 1995).

Para el desarrollo de esta etapa se utilizaron los genes interrumpidos del cromosoma 19 de *M. musculus* como conjunto inicial de datos para la búsqueda de patrones. A continuación se presenta cada uno de los análisis desarrollados para la búsqueda de patrones con el fin de lograr discriminar secuencias de ADN codificantes de las no codificantes (Secciones a, b, c y d).

a) Aplicación de las medidas estándar de predicción de genes

Los cálculos de las medidas estadísticas estándar para la predicción de exones, se llevaron a cabo a partir de los protocolos de Guigó (1998) para Uso de Codón, de Ficket (1982) para prueba de ficket, de Claverie y Bougueleret (1986) para hexámeros y Koski (2001) para contenido GC. Cada uno de los algoritmos aplicados para obtener los respectivos resultados se realizaron en python.

b) Aplicación del análisis de Rango Reescalado R/S

Una secuencia de ADN puede ser considerada como un alfabeto conformado por cuatro letras {A, C, G, T} que representan las cuatro bases que constituyen la secuencia de ADN: Adenina, Citosina, Guanina, Timina. Cualquier secuencia de ADN $s = s_1 s_2 \ldots s_N$, se define S en función de f de la siguiente manera $f : s \rightarrow x = x_1, x_2 \ldots x_N$ para cualquier $1 \leqslant k \leqslant N$ (Yu y Chen, 1999),

$$x_k = \begin{pmatrix} -2 & \text{si} & S_k = A \\ -1 & \text{si} & S_k = C \\ 1 & \text{si} & S_k = G \\ 2 & \text{si} & S_k = T \end{pmatrix}$$

De acuerdo con la definición de f , las cuatro bases {A, C, G, T} son sustituidas por cuatro valores diferentes (-2, -1, 1,2), lo importante es distinguir las purinas A y G de las pirimidinas C y T.

Se obtiene un valor de la secuencia dado por $x=\{x_k\}_{k=1}^{2}$, donde $x_k \in \{-2,-1,1,2\}$. Esta secuencia es tratada como un registro fractal en el tiempo. Para estudiar estos registros Hurst (Hurst, 1951) inventó un método estadístico -el análisis de rango reescalado (R/S)- que posteriormente fue modificado por Mandelbrot (Mandelbrot, 1982) y J. Feder (Feder, 1988), quienes introdujeron el análisis R/S para registros fractales. Para cualquier registro fractal en el tiempo $x=\{x_k\}_{k=1}^{2}$ y para cualquier $2 \leqslant n \leqslant N$, se puede definir:

$$\langle x \rangle_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$X(i,n) = \sum_{u=1}^{i} [x_u - \langle x \rangle_n]$$

$$R(n) = maxX(i,n) - minX(i,n)$$

$$S(n) = [\frac{1}{n} \sum_{i=1}^{n} (x_i - \langle x \rangle_n)^2]^{1/2}$$

Hurst (Hurst, 1951) encontró que: $R(n)/S(n) \sim (\frac{n}{2})^{H}$

H es el exponente de Hurst que se determina por medio de la regresión lineal de los puntos de *log*(R(*n*)/S(*n*)) vs *log*(*n*).

Se diseño un algoritmo en lenguaje python para el calculo del exponente de Hurst. Aquí un ejemplo del calculo del exponente de Hurst: el Cuadro 3 muestra el desarrollo de los pasos por columnas, para llegar al resultado de la Gráfica 16:

1) Cambio de bases por pesos

2) Cálculo de la media general

3) Cálculo de las desviaciones estándar

4) Acumulación de las desviaciones estándar

5) Cálculo de los rangos (R)

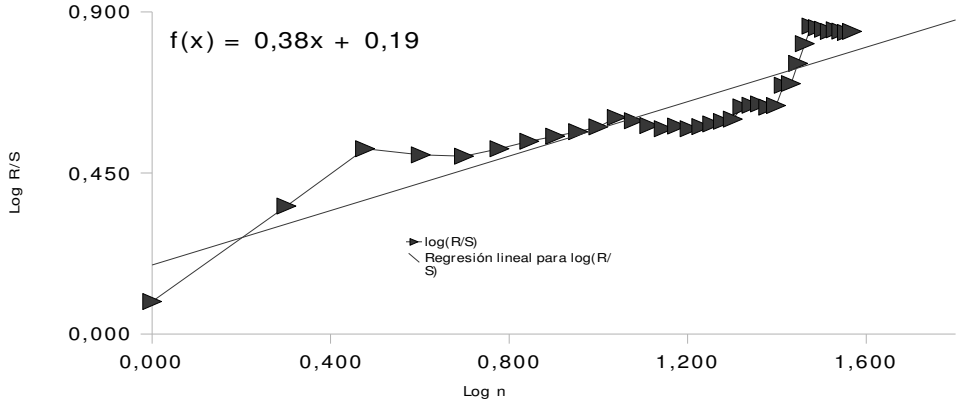
6) Cálculo de las desviaciones estándar (S)

7) Cálculo de los logaritmos de R/S y n

En el Cuadro 3 muestra que para la secuencia CTTTATCGCC el primer paso se representa por la columna “1)”, el segundo por la columna “2)Media” y asi sucesivamente.

Cuadro 3: Ejemplo del procedimiento para el cálculo del coeficiente de Hurst

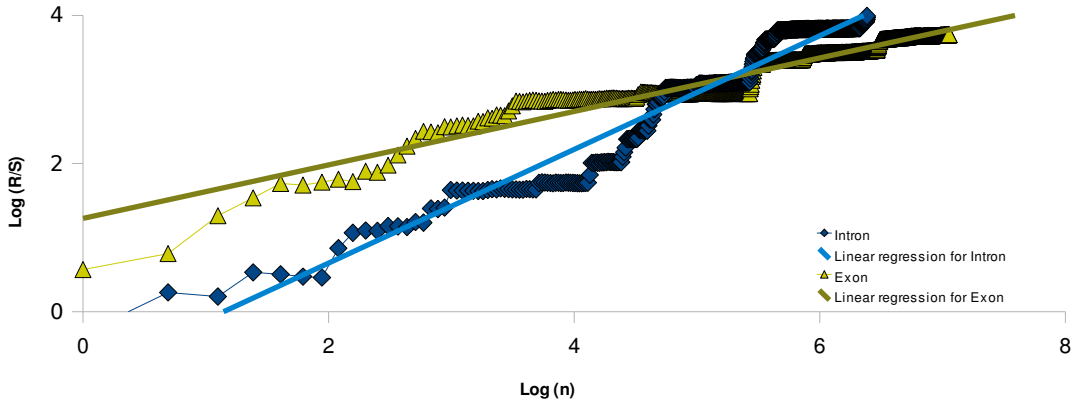
seq ADN	1)	2)Media	3)	4)	5)	6)	7) log(R/S)	7) log(n)
C	-1	0.079	-1.079	-1.079	1.921	1.56	0.091	0.000
T	2	0.079	1.921	0.842	3.842	1.69	0.357	0.301
T	2	0.079	1.921	2.763	5.763	1.75	0.518	0.477
T	2	0.079	1.921	4.684	5.763	1.82	0.501	0.602
A	-2	0.079	-2.079	2.605	5.763	1.84	0.497	0.699
T	2	0.079	1.921	4.526	5.763	1.75	0.518	0.778
C	-1	0.079	-1.079	3.447	5.763	1.67	0.538	0.845
G	1	0.079	0.921	4.368	5.763	1.61	0.553	0.903
C	-1	0.079	-1.079	3.289	5.763	1.57	0.565	0.954



Gráfica 16: Ejemplo del análisis R/S en una secuencia de ADN de *M. musculus*

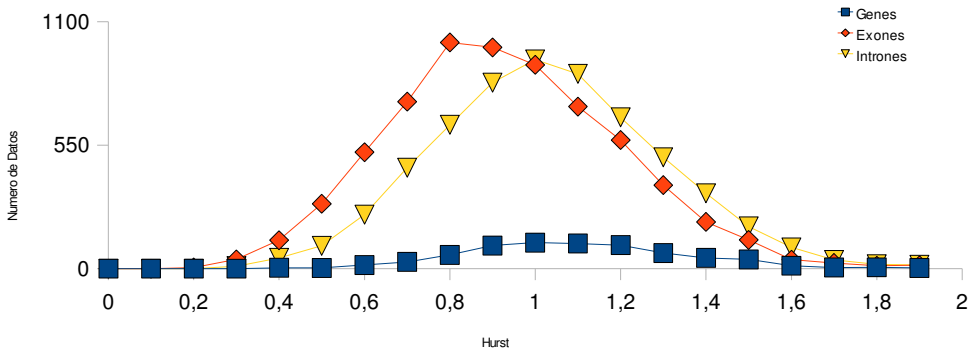
Ejemplo del calculo del exponente de Hurst: el siguiente cuadro muestra el desarrollo de los pasos para llegar al resultado de la Gráfica 16. Definimos a N como el número total de elementos de la secuencia y n variando desde 2 a N, se obtienen al final N-1 puntos para el gráfico *log*(n) vs *log*(R/S). Se calcula el exponente de Hurst H de la secuencia de ADN por medio de una regresión lineal y la pendiente resultante proporciona el valor del exponente.

La Gráfica 17 ilustra el comportamiento del análisis R/S para los dos tipos secuencias estudiadas. Valores pequeños en los parámetros de los exones contrastan con valores altos en los intrones. La causa principal de estos valores radica en la composición de bases y en el tamaño de las secuencias. El alto contenido de bases (A y T), repetición de bases y tamaños grandes de las secuencias de intrones (hasta 60000 pb) contrasta con el contenido de bases (C y G) y tamaño corto (hasta 7000 pb) de las secuencias de exones. Estas características marcan una caminata del mínimo al máximo más rápida en intrones que en exones por lo tanto la pendiente o exponente de Hurst tiende a ser más alto en intrones que en exones.



Gráfica 17: Diagrama de Pox de los datos del análisis R/S con secuencias de ADN de *M. musculus*: exón (H=0.36); intrón (H=0.76).

El registro de datos de exones, intrones y genes del cromosoma 19 de *M. musculus* presentan una distribución normal a lo largo del exponente de Hurst, esta tendencia se mantiene a lo largo del genoma, indicando que esta distribución no es al azar, reflejando un comportamiento diferencial entre exones e intrones (Gráfica 18). Los genes tienen una curtosis de -0.76 y una asimetría de 0.89, para intrones -0.87 y 0.8, para exones de -0.96 y 0.76 respectivamente.



Gráfica 18: Distribución normal de los datos del cromosoma 19 de *M. musculus*. Exones, intrones y genes

El análisis de rango reescalado R/S con su respectivo coeficiente de Hurst se realizó para cada secuencia de ADN del grupo de secuencias de exones y de intrones, que conforman cada cromosoma del genoma de *M. musculus*. El coeficiente de Hurst se dividió en 10 intervalos de igual amplitud, de tal manera que cada secuencia se clasificó en un intervalo según el coeficiente obtenido y se procedió a analizar los parámetros del análisis R/S (máximo, mínimo, rango y longitud) para cada intervalo. El Cuadro 4, muestra la distribución de las secuencias de exones por intervalos del coeficiente de Hurst para el

cromosoma 19, de la misma manera se procedió para el análisis de las secuencias de intrones de cada cromosoma.

Cuadro 4: Parámetros del conjunto de datos de las secuencias de exones en el genoma de *M. musculus*.

$H_{(exones)}$	Frec.	Max.		Min.		Rango		Long. (pb)	
0,1	215	7,89	±7,10	-4,84	±11,19	12,74	±12,89	160,29	±407,99
0,2	4804	10,28	±8,04	-7,47	±14,13	17,75	±16,52	199,54	±431,05
0,3	25981	11,86	±9,82	-10,21	±14,09	22,07	±18,68	231,97	±456,96
0,4	50941	13,55	±12,08	-12,93	±17,60	26,48	±23,70	253,83	±485,02
0,5	46703	15,53	±15,10	-15,82	±21,66	31,35	±29,07	270,56	±521,64
0,6	21255	17,76	±18,37	-19,62	±29,67	37,38	±37,74	293,77	±586,69
0,7	5051	20,74	±25,55	-25,49	±44,53	46,24	±54,24	332,18	±663,73
0,8	653	23,07	±31,58	-32,54	±67,63	55,61	±80,36	386,02	±980,33
0,9	75	19,27	±38,82	-25,50	±72,00	44,76	±83,56	257,99	±780,84
1	28	6,14	±11,69	-4,84	±9,67	10,99	±20,71	41,14	±103,79

Parámetros del análisis R/S: **H**: coeficiente de Hurst; **Max**: máximo; **Min**: mínimo; y rango; **long**: longitud.

El comportamiento de los parámetros del análisis R/S (máximos, mínimos, rango y longitud) por intervalos del coeficiente de Hurst difieren para cada uno de los grupos de secuencias estudiadas del cromosoma 19: exones e intrones (Cuadro 4 y Cuadro 5). Por ejemplo, una secuencia de ADN codificante con un coeficiente de Hurst ente 0.4 y 0.5 presenta un máximo de 13.55± 12.08, un mínimo de -12.93±17,60, un rango de 26.48±23,70 y una longitud promedio de 253.83±485,02. En contraste, una secuencia no codificante con un mismo coeficiente de Hurst, en el intervalo entre 0.4 y 0.5 presenta un máximo de 62.80±123,57; mínimo de -58,56±104,37; 121,36±192,78 y una longitud de 3235,77±11245,92.

Cuadro 5: Parámetros del conjunto de datos de las secuencias de intrones del genoma de *M. musculus*.

$H_{(Intrones)}$	Frec.	Max.		Min.		Rango		Long.	
0,1	103	37,63	±100,66	-26,25	±40,18	63,89	±118,64	2311,30	±7589,18
0,2	2622	37,43	±121,91	-34,07	±48,06	71,50	±142,33	2252,59	±9547,94
0,3	15785	47,16	±111,60	-45,74	±89,51	92,90	±168,02	2709,50	±11568,42
0,4	38615	62,80	±123,57	-58,56	±104,37	121,36	±192,78	3235,77	±11245,92
0,5	42974	92,24	±192,02	-83,32	±154,16	175,56	±292,16	4470,89	±16145,45
0,6	26481	140,21	±279,31	-128,88	±242,39	269,09	±438,99	6070,24	±18787,80
0,7	9939	230,67	±419,31	-221,21	±405,02	451,88	±689,39	9091,12	±24328,56
0,8	2673	439,05	±695,29	-419,73	±665,17	858,78	±1087,58	14695,92	±28830,73
0,9	598	762,57	±971,95	-721,52	±965,65	1484,09	±1407,18	22583,50	±35376,84
1	114	1193,16	±1178,62	-977,29	±1036,25	2170,46	±1444,85	27060,98	±24637,75

Parámetros del análisis R/S: **H**: coeficiente de Hurst; **Max**.: máximo; **Min**.: mínimo; y rango; **long**.: longitud

En total se evaluaron 16 cromosomas de los 21 que contiene el genoma de *M. musculus* para un total de 15803 genes: 155707 exones y 139904 intrones. El total de genes estudiados corresponde al 86% del genoma de *M. musculus*.

El conjunto de datos de las secuencias de exones presentaron un H<0.5 (p<0.001) (Cuadro 6), lo que indica que son series antipersistentes con correlaciones de corto alcance, alta complejidad e irregularidad. En contraste, los intrones presentaron un H>0.5 (p<0.001), indicando que son series persistentes relacionadas con correlaciones de largo alcance, baja complejidad y alta regularidad; estas características pueden ser debidas a la existencia de estructuras repetitivas en la secuencias de intrones. Estos resultados $H_{exon}<H_{intrón}$ coinciden con los obtenidos por Yu y Chen (2000), Luo y Lee (1998) y Shen et al (1993).

Cuadro 6: Resultados del análisis de rango reescalado R/S y del coeficiente de Hurst en grupos de secuencias de exones e intrones del genoma de *M. musculus*.

	Frec.	**H**	**DS**	**EE**	**Max.**	**Min.**	**Rango**	**R^2**
Exones	155707	0,496	0,111	0,000	1,589	0,024	1,566	0,888
Intrones	139904	0,535	0,124	0,000	1,264	0,123	1,141	0,893

Frec.: frecuencias; **H**: coeficiente de Hurst; **DS**: desviación estándar; **EE**.: error estándar; **R2**: coeficiente de determinación; parámetros del análisis R/S: **Max**.: Máximo; **Min**: mínimo y Rango (diferencia ente el máximo y el mínimo); **DF**: dimensión fractal, D=2-H. Tendencias estadísticamente significativas (p<0.001).

c) Aplicación del análisis bidimensional

El análisis del conjunto de datos se realizó aplicando el método de dimensión fractal bidimensional, según modificaciones hechas a los protocolos de Havlin y Ben-Avrham (1982); Gates (1986); Xia y colaboradores (1995).

Para cada una de las UI se obtuvo la suma promedio de las distancias ($<R^2_N>_{No}$), para cada escala de longitud (recorriendo las secuencias de a pasos de 1 nucleótido, de 2 nucleótido, etc; hasta N_o). A partir del $<R^2_N>_{No}$ de cada secuencia se seleccionaron los valores máximo y mínimo (R^2_{max} y R^2_{min}) del recorrido a diferentes escalas de longitud hasta N_o. A partir del gráfico $ln(N)$ y $ln(<R^2_N>_{No})^{1/2}$ se ajustó una recta por mínimos cuadros, en donde la pendiente es la dimensión fractal (DF) de cada secuencia de ADN o UI. Además de la DF obtenida, se adicionaron al grupo de atributos arrojados por el análisis bidimensional y con el objetivo de mejorar la clasificación, los siguientes parámetros: longitud de la secuencia (cantidad de pares de bases), R^2_{max} y R^2_{min} de los N pasos recorridos en la secuencia, amplitud ($R^2_{max} - R^2_{min}$) y finalmente, la dimensión fractal teórica propuesta por Gates (1986). Una vez obtenida todas las dimensiones para exones del cromosoma 19, se agruparon las secuencias por rangos de acuerdo con sus dimensiones fractales, el Cuadro 7 muestra el ordenamiento por dimensión fractal de las secuencias codificantes del cromosoma 19 de *M. musculus*, las dimensiones oscilan entre -4,6 a 10,40 así para la

primer secuencia se tiene una dimensión fractal bidimensional de -4,603 con longitud de 1267 nucleótidos, una distancia media máxima de 2,59, una distancia media mínima de 0,6 y una dimensión teórica fractal de 1,29.

Cuadro 7: Agrupamiento de las secuencias de ADN por Dimensión Fractal (Extracto) para el Cromosoma 19 *M. musculus*

DF	long	MaxRN	MinRN	Dtf
-4,60390	1267	2,59375	0,60000	1,29
-4,57873	1490	2,50000	0,75000	1,62
-4,21985	97	2,18056	1,00000	1,61
-4,16485	468	2,48913	0,83333	1,43
. . .	. . .	. . .	. . .	. . .
8,69420	2444	2,55556	1,00000	1,43
9,37683	1778	2,39130	1,00000	1,31
10,40142	1035	2,68182	0,50000	1,29

DF: dimensión fractal, **long**: longitud de la secuencia en pares de base (bp), **MaxRN**: valor máximo de ($<R^2_N>_{No}$), **MinRN**: valor mínimo de ($<R^2_N>_{No}$), **Dtf**: dimensión fractal teórica

A cada uno de los parámetros obtenidos en los rangos de dimensiones fractales para exones se les hizo un proceso de ajuste estadístico, generando con esto los atributos que van a servir para el predictor de exones, el Cuadro 8 muestra los 16 rangos, donde el primer corresponde a secuencias de dimensión fractal bidimensional entre -5 y -4 con una

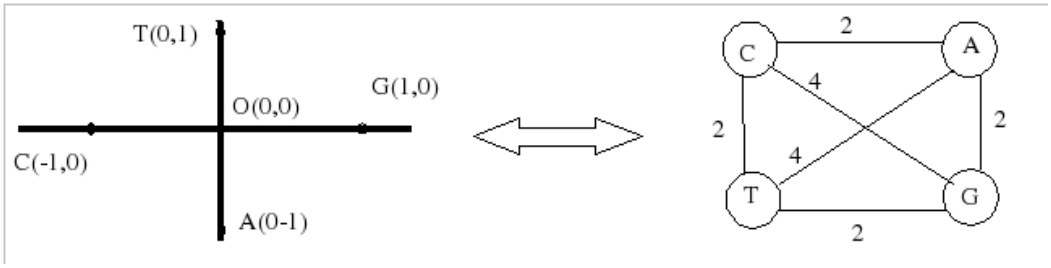
Cuadro 8: Parámetros ajustados del algoritmo de predicción para exones con base en el análisis bidimensional.

DF	Ampl	long	Max	Dtf
-5:-4	1.18056:1.99375	97:1490	2.18056:2.875	1.24506:1.62374
-4:-3	1.07317:2.05556	11:3125	2.07317:2.78947	1.20391:2.44566
-3:-2	1.25862:2.75	8:5161	2.13889:3.5	0.0:7.22882
-2:-1	0.88333:3.16667	7:4383	1.13333:3.5	0.0:8.88264
-1:0	1.25:3.0	8:7431	1.85714:3.57143	0.0:8.77479
. . .				
4:5	1.37143:2.5	57:4188	2.37143:3.4	1.22004:8.87036
5:6	1.875:2.15	123:6450	2.52439:3.0	1.49167:1.89562
6:7	1.5:2.33333	1292:3763	2.5:2.83333	1.31312:1.63242
7:8	1.625:2.25	418:1950	2.625:3.25	1.42164:1.96472
8:9	1.5:2.22222	1478:2703	2.5:3.22222	1.38547:1.57909
9:10	1.3913:1.3913	1778:1778	2.3913:2.3913	1.31352:1.31352
10:1000	2.18182:2.18182	1035:1035	2.68182:2.68182	1.29487:1.29487

DF: dimensión fractal, **Ampl**: amplitud, **Length**: longitud de la secuencia, **Max**: R^2_{max}, **Dtf**: dimensión fractal teórica. Los dos puntos indican un rango.

amplitud entre 1,18 y 1,99 , una longitud entre 97 y 1490 nucleótidos, una distancia media máxima que va de 2,18 a 2,87 y una dimensión teórica fractal entre 1,24 y 1,62.

Las interrelaciones entre los cuatro nucleótidos que interactúan en el plano cartesiano (representación de Gates, 1982), se puede ver como un grafo no dirigido de cuatro nodos, donde cada arista representa la distancia al cuadrado de un nodo a otro en el plano cartesiano (Gráfica 19). La distancia del centro del plano a cualquier nodo siempre es igual a uno, la distancia entre C y A es $\sqrt{2}$ en el plano cartesiano; pero elevada al cuadrado es 2 como se muestra en la arista que une los nodos C y A en el grafo.



Gráfica 19: Grafo representativo de las distancias al cuadrado ($R^2_{i,i+N}$) de un nodo a otro en el plano cartesiano

d) Aplicación de la Ley de Zipf-Mandelbrot

Del conjunto obtenido en la etapa de preparación de los datos de minería de datos y específicamente exones e intrones denominados Unidades de Información (UI), se aplicó la Ley experimental de Zipf-Mandelbrot tomando como alfabeto los cuatro nucleótidos de ADN y la longitud de las UI como el término por el cual hacer la jerarquización del conjunto de datos.

Inicialmente se realizó la codificación de los procesos y un análisis exploratorio de las tendencias de los datos aplicando la ley de Zipf-Mandelbrot. Este análisis se realizó con todas las UI (exones e intrones) del cromosoma 19 de *M. musculus*.

Cada UI se ordena de mayor a menor longitud (jerarquía de acuerdo a su longitud) en pares de bases, es decir se obtuvo un ordenamiento descendente en relación a la longitud para todo el conjunto de exones e igualmente para intrones. A cada una de las posiciones del listado se le asignó un número consecutivo en orden ascendente, que en adelante se denomina rango. De acuerdo a la modificación de Mandelbrot (1975), a este rango se le suma una constante, que en este caso es 1/N-1 donde N=4; por ser el alfabeto correspondiente a los cuatro nucleótidos que conforman las secuencias de ADN . La frecuencia relativa de cada UI es el cociente entre la longitud de la UI y la suma total de las longitudes de todos las UI (exones e intrones). La pendiente del gráfico log(rango) vs log(frecuencia) se denomina el exponente de Zipf. La dimensión fractal del conjunto es el inverso de la pendiente. Este procedimiento se realizó para cada uno de los genomas contemplados en este estudio.

Con el fin de obtener parámetros que permitieran discriminar exones de intrones, se agruparon las UI por rangos a intervalos de igual amplitud (0,0001), dando como resultado 196 rangos para el genoma *M. musculus* (Anexo A). Las UI se clasificaron de acuerdo con su rango; dentro de cada intervalo del rango se obtuvieron máximo y mínimo del rango, máximo y mínimo de las longitudes y, máximo y mínimo de las frecuencias de las UI ubicadas en el intervalo correspondiente. El mismo procedimiento fue aplicado a los demás genomas de estudio. Los parámetros del algoritmo fueron establecidos a partir del conjunto de búsqueda de patrones correspondiente al genoma de *M. musculus*. La implementación del algoritmo Zipf-Mandelbrot y demás scripts necesarios para este análisis, se realizaron en Python, un lenguaje de programación para prototipado rápido.

3.1.3.3 Análisis de los resultados de las medidas

Luego de la selección y aplicación de las medidas, se analizó la información obtenida de importancia biológica. Este apartado se desarrolla en el **capitulo 4**.

3.1.3.4 Selección y limpieza de atributos

En cada una de las medidas anteriores se obtuvo una tabla donde se colocaron los atributos calculados y probados de forma individual. Con el fin de que el modelo tuviera más criterios para tomar una decisión, se combinaron los atributos obtenidos; pero antes se evalúa la relación entre cada una de los atributos obtenidos, con el fin de generar un modelo lo más confiable posible, es decir, que el modelo no tenga atributos que den la misma información o que directa o indirectamente estén relacionados; lo que estadísticamente se conoce como multicolinealidad (Canavos, 1984).

Una alternativa para eliminar la multicolinealidad es determinar la matriz de correlación (Anexo B) para los 32 atributos (22 estadísticos estándar y 10 medidas fractales). Esta matriz contiene todas las posibles pares de combinaciones donde se seleccionaron aquellos atributos con coeficientes de correlación en el intervalo de -0,75 a 0,75. Según Canavos (Canavos, 1984) si la correlación es pequeña las consecuencias son de índole menor, sin embargo si existe una correlación entre dos a más atributos los resultados son ambiguos. Cuando dos o más atributos de predicción son colineales no se miden los efectos individuales sobre una respuesta sino que estos reflejan un efecto parcial, sujeto a todo lo que pase con los demás atributos de predicción del modelo. Se realizó el análisis confirmatorio de los atributos que por si solos aportan información importante en la construcción de un modelo desarrollando un análisis de regresión multivariado utilizando la técnica de stepwise, que aporta el verdadero valor de cada atributo que realmente debe pertenecer al modelo.

La matriz de correlación resultante se compone de 32 filas y columnas donde se calcula el coeficiente de correlación entre los conjuntos de datos de los atributos (Anexo C) que se cruzan en cada celda. Ya que la matriz es simétrica, sólo se calcula la parte inferior, bajo la

diagonal principal (Anexo B). Hecho esto, se seleccionan aquellos atributos que se crucen en la celda con menor valor; además se seleccionaron aquellos atributos que ofrecían información biológica relevante. Obteniendo una vista minable preliminar (Cuadro 9) con los atributos relevantes obtenidos de la matriz de correlaciones.

Reducida la cantidad de atributos mediante el coeficiente de correlación, la vista preliminar contiene 12 atributos numéricos (10 reales y 2 enteros) y un atributo nominal (ui) con las categorías de exón e intrón, las cuales permiten identificar lo que se quiere clasificar. La ecuación de regresión multivariada se realizó con el conjunto de entrenamiento utilizando MATLAB.

Cuadro 9: Vista minable preliminar (extracto)

ui	fv1	fv4	%gc	rcg	pusage	D	DmaxRN	DminRN	Dtfd	hurst	max	min
intron	0	0	32,26	0	0,030	-0,600	2,75	0,5	1,49	0,37	2,89	-46,08
intron	1	0	35,49	0,46	0,020	-2,380	2,75	0,5	1,52	0,5	21,8	-24,86
intron	0	0	43,7	0	0,050	-1,300	2,75	0,5	1,42	0,62	4,59	-21,69
intron	0	0	34,4	0,55	0,050	-2,500	2,7	0,5	1,46	0,59	10,14	-16,45
intron	0	0	32,82	0,87	0,050	-1,790	2,75	0,5	2,22	0,4	11,48	-4,55
intron	0	0	32,04	0,85	0,060	-0,590	2,63	0,5	1,33	0,48	3,04	-24,48
intron	0	0	26,04	0	0,070	-1,240	2,75	0,5	1,73	0,39	6,38	-11,54
intron	0	0	32,23	0,69	0,050	-1,620	2,75	0,5	1,77	0,49	7,58	-15,6
intron	2	0	36,11	0,63	0,030	-0,920	2,75	0,5	1,2	0,37	4,11	-19,71
intron	0	0	24,05	0	0,040	-1,830	3	0,5	1,83	0,52	15,72	-9,82
...												
exon	0	0	38,89	1,5	0,120	-2,010	2,76	0,85	1,73	0,33	2,4	-10,35
exon	0	0	39,21	0,23	0,030	-1,720	2,88	0,75	1,84	0,41	19,03	-19,24
exon	0	0	57,14	0	0,100	-2,020	2,63	1	1,73	0,72	11,55	-5,06
exon	0	0	38,89	0	0,060	-1,070	2,5	0,5	1,51	0,71	25,16	-3,96
exon	0	0	45	0,33	0,110	-1,330	3,25	0,5	2,95	0,59	8,52	-8,07
exon	6	6	50	0,42	0,030	-1,660	2,75	0,5	2,06	0,41	14,76	-8,73
exon	0	0	48,31	0,44	0,050	1,650	3,17	1	1,59	0,56	22,99	-1,2
exon	25	25	51,92	0,21	0,040	-2,410	2,5	0,5	1,55	0,4	10,62	-13,95
exon	0	0	50	0,45	0,060	-2,240	2,46	1	2,03	0,53	16,18	-3,29
exon	6	0	41,88	0,39	0,050	-1,890	2,6	0,56	2,17	0,49	9,46	-15,39
exon	0	0	52,85	0,24	0,050	-1,680	2,75	0,5	1,58	0,38	15,95	-2,66
...												
exon	0	0	37,2	1,27	0,030	-1,790	2,64	0,5	2,74	0,47	15,64	-5,49
exon	0	0	36,51	0,5	0,100	-1,190	2,72	0,33	1,62	0,65	3,33	-13,68
exon	0	0	55,26	0,36	0,060	-1,710	3	0,6	1,71	0,59	11,53	-10,48
intron	210	210	40,44	0,31	0,000	-2,190	2,46	0,5	1,56	0,6	79,5	-110,83
intron	0	0	34,46	0,15	0,010	-1,320	2,75	0,5	1,81	0,4	55,31	-13,66
intron	0	0	45,83	0,29	0,020	-1,180	2,78	0,5	2,11	0,31	14,31	-21,74
...												
intron	1	1	61,17	0,12	0,030	-0,620	2,75	0,5	1,24	0,5	16,83	-6,3
intron	0	0	31,34	0,19	0,010	-1,220	2,75	0,5	1,55	0,57	20,49	-33,19
intron	28	28	46,44	0,18	0,010	4,310	2,75	0,5	1,45	0,79	111,3	-60,77
intron	79	79	48,33	0,33	0,000	16,730	2,75	0,5	1,25	0,85	291,25	-80,22
intron	24	24	42,35	0,65	0,000	-0,990	2,58	0,5	1,81	0,52	81,24	-82,98

ui: unidad de información (exón/intrón), **fv1**: ficket ventana 1, **fv4**: ficket ventana 4, **%gc**: porcentaje de GC, **rcg**: índice de CG, **pusage**: porcentaje de uso de codón, **D**: dimensión fractal bidimensional, **DmaxRN**: dimensión fractal máxima, **DminRN**: dimensión fractal mínima, **Dtfd**: dimensión fractal teórica, **hurst**: coeficiente de Hurst, **max**: coeficiente máximo de Husrt, **min**: coeficiente mínimo de Hurst

Con el análisis multivariado se eliminó el atributo **fv4** (ficket ventana 4) de la vista minable preliminar, ya que el método de regresión con base en la técnica de stepwise indica que el intervalo de confianza del atributo **fv4** contiene a cero en su intervalo (Cuadro 10), haciendo que se elimine dicha variable, por lo tanto la vista minable resultante contiene 11 atributos numéricos (10 reales y 1 entero) de los 12 seleccionados con la matriz de correlación. La vista minable final para crear el modelo esta representada por el Cuadro 9 sin la columna fv4.

Cuadro 10: Resultados del análisis de atributos utilizando regresión multivariada con base en la técnica stepwise. **Parámetro**: coeficiente de regresión de la variable

Atributo	Parámetro	Intervalo de confianza	
		Inferior	Superior
fv1	0,02585	0,01203	0,03967
fv4	**0,00725**	**-0,007085**	**0,02159**
%GC	0,1517	0,1424	0,161
ratio_CG	0,04539	0,03618	0,0546
pusage	0,07296	0,06354	0,08239
D	-0,0434	-0,05527	-0,03154
DmaxRn	-0,0252	-0,03437	-0,01603
DminRn	0,1825	0,1732	0,1918
Dtf	-0,0921	-0,1013	-0,083
Hurst	0,01302	0,003604	0,02244
Max	-0,0298	-0,0431	-0,01655
Min	0,034	0,02125	0,04676

fv1: ficket ventana 1, **fv4:** ficket ventana 4, **%GC**: porcentaje de GC, **ratio_CG**: índice de CG, **pusage**: porcentaje de uso de codón, **D**: dimensión fractal bidimensional, **DmaxRN**: dimensión fractal máxima, **DminRN**: dimensión fractal mínima, **Dtf**: dimensión fractal teórica, **Hurst**: coeficiente de Hurst, **Max**: coeficiente máximo de Hurst, **Min**: coeficiente mínimo de Hurst.

Junto con el análisis de multicolinealidad y el análisis de regresión múltiple, se construyó una matriz de pesos (Cuadro 11), donde dependiendo del tipo de error que se cometa se tendrá un costo. Este tipo de matrices contrasta información de los costos para los casos predichos por el modelo y los casos reales, el objetivo de la matriz es multar la predicción cuando esta se equivoca, para que el algoritmo sea mas cuidadoso ya que tiene en cuenta el coste de una equivocación. Desde el punto de vista biológico se busca incrementar la precisión en la predicción de exones para reducir la perdida de información,

es decir, es más costoso perder un exón porque implica perder información (codificación de una proteína) que permitir cierto nivel de ruido en los datos al introducir intrones como exones ya que estos posteriormente pueden ser eliminados en el proceso experimental.

Cuadro 11: Matriz de costos

	Real		
Predicho		Exón	Intrón
	Exón	0	1
	Intrón	5	0

El proceso de identificación experimental que se lleva a cabo en el laboratorio es muy costoso cuando un predictor que ayuda en el proceso se equivoca, especialmente cuando da Falsos Negativos ya que se esta dejando por fuera del conjunto de secuencias aquellas que codifican proteínas, un costo muy elevado ya que no se podría inferir resultados mas acertados al producir una medicina por ejemplo, mientras que si el predictor arroja Falsos Positivos, el experto del laboratorio puede identificar y eliminar secuencias no codificantes por medios quimicos sin incurrir en grave perdida de información para posteriores etapas.

3.1.4 Etapa de modelamiento

Esta etapa comprende seleccionar la técnica de aprendizaje de máquina, seleccionar los conjunto de entrenamiento y prueba y, crear los modelo con las técnicas seleccionadas, esta última parte, por su importancia para este proyecto, se detalla en el **capitulo 5, sección 5.1**.

3.1.4.1 Selección de las técnicas de aprendizaje de máquina

Las diferentes técnicas de clasificación como Redes Neuronales (Neural Network), Máquinas de soporte vectorial (Support Vectorial Machine), Redes bayesianas (Bayesian Networ), k-esimo vecino más cercano (k-Nearest Neighbour), Reglas de aprendizaje (Rule Learners) y Árboles de decisión (AD) presentan ventajas y desventajas dependiendo de las características del problema a afrontar y no hay una regla general para su aplicación. Algunas de las técnicas anteriores son usualmente usadas en bioinformática (NN, SVM, BN y AD) para el procesamiento de secuencias de ADN; el Cuadro 12 muestra algunas de las características importantes de las técnicas de clasificación, resaltando las mas relevantes para este proyecto como velocidad en clasificación, aprendizaje, fácil explicación, manipulación de parámetros entre otras. Las que comparten mayoría de detalles son AD, BN, kNN y Reglas de aprendizaje, esta última se descartó ya que un conjunto de reglas de aprendizaje individuales forman un árbol de decisión, de igual manera se descartó kNN ya que clasifica buscando similaridades entre instancias (registros) de un conjunto de datos (Kotsiantis, 2007), lo que no fue adecuado para el enfoque de este proyecto que buscaba información obtenida de la propia secuencia (de su contenido y estructura) y no a partir de su relación con otras. Aunque la técnica SVM es rápida para clasificar no es adecuada ya

que en el proceso de entrenamiento no se tiene en cuenta todas las instancias (registros) del conjunto de datos (Bedoya, 2006), esto implica perdida de información ya que de un conjunto heterogéneo de secuencias no se toman todas las posibilidades que brinda dicho conjunto.

Los árboles de decisión presentan ventajas en la facilidad de implementación ya que comprende uso de reglas o sentencias sencillas de toma de decisiones, además los datos no son restringidos por operaciones o funciones estadísticas sino que son los mismos valores de los datos los encargados de dar información para poder clasificarlos. El árbol de decisión al componerse de reglas sencillas es fácil poder hacerle un seguimiento para corroborar empíricamente si las reglas sirven o modificarlos de acuerdo al problema. Las redes bayesianas comparte algunas características con los árboles de decisión aunque su implementación y estructura sean muy diferentes, buena técnica para confrontar modelos.

Cuadro 12: Características de las técnicas de Aprendizaje

Características	AD	NN	BN	kNN	SVM	Rule-learners
Exactitud en general	**	***	*	**	****	**
Velocidad de aprendizaje	***	*	****	****	*	**
Velocidad de clasificación	****	****	****	*	****	****
Tolerancia a valores faltantes	***	*	****	*	**	**
Tolerancia a valores irrelevantes	***	*	**	**	****	**
Tolerancia a atributos redundantes	**	**	*	**	***	**
Tolerancia a atributos altamente interdependientes	**	***	*	*	***	**
Tratar atributos discretos/binarios/continuos	****	***	***	***	**	***
Tolerancia al ruido	**	**	***	*	**	*
Tratar con peligros de sobre-ajuste	**	*	***	***	**	**
Posibilidades de aprendizaje incremental	**	***	****	****	**	*
Fácil explicación/Transparencia de conocimiento/Clasificaciones	****	*	****	**	*	****
Manipular parámetros del modelo	***	*	****	***	*	***

AD: árbol de decisión, **NN**: red neuronal, **BN**: red bayesiana, **kNN**: k-esimo vecino más cercano, **SVM**: máquina de soporte vectorial, **Rule-learnes**: reglas de aprendizaje. ***** bajo desempeño. ********* alto desempeño. Tomado de (Kotsiantis, 2007)

Finalmente se seleccionó redes neuronales para completar una terna de técnicas de diferente algoritmos de aprendizaje usados para afrontar el problema de clasificación de secuencias de ADN.

3.1.4.2 Selección conjunto muestral

Para determinar el tamaño muestral de una población finita basada en atributos se empleó el muestreo aleatorio simple (M. A. S) (Martinez, 2003). El nivel de confianza (Z_a^2) y el error de muestreo (d) son definidos por el investigador, por lo tanto, para este caso el nivel de confianza es de 95%, correspondiendo a un valor de Z de 1,96, según la tabla de distribución Z; el error de muestreo seleccionado fue de 0,02. El valor de p es la proporción esperada en la población de exones e introns estudiados, con base en el conocimiento que se tiene de los datos se utilizó un valor de p igual a 0,5 (50%), que así mismo maximiza el tamaño muestral (q).

$$n = \frac{N * Z_a^2 * p * q}{d^2 * (N-1) + Z_a^2 * p * q} \qquad (3.1)$$

Donde:

- N = Total de la población (ver Cuadro 2 de generalidades de los genomas estudiados; N será el total de Exones e Introns igual a 1'602.490 de secuencias)
- Z_a^2 = 1.962 (la seguridad es del 95%), nivel de confianza: 0,05
- p = proporción esperada (en este caso 50% = 0.5)
- q = 1 – p (en este caso 1 - 0.5 = 0.5)
- d = precisión (en este caso se desea un 2%), error de muestreo

Aplicando la formula 3.1 la muestra mínima para el análisis fue de 2397 secuencias:

$$n = \frac{1602490 * 1{,}96^2 * 0{,}5 * (1-0{,}5)}{0{,}02^2 * (1602490-1) + 1{,}96^2 * 0{,}5 * (1-0{,}5)} = 2397$$

La muestra seleccionada fue de 17318 secuencias aleatorias de la población total; la cual se dividió en dos conjuntos: uno de 86% para entrenamiento y otro de 14% para pruebas, no hay una teoría definitiva que indique el porcentaje para conformar conjuntos de entrenamiento y prueba, sin embargo el experto en el tema sugirió los porcentajes utilizados.

3.1.4.3 Creación del modelo

El modelo se creó mediante la herramienta de Aprendizaje de Máquina, Weka (Witten y Frank, 2005); con la técnica de árboles de decisión y su algoritmo de aprendizaje J4.8 (C4.5). Se usaron dos técnicas de Aprendizaje de Máquina: Perceptrón multicapa y Bayes Network para comparar la validez del modelo y tener como referente los resultados de los otros modelos.

Los árboles de decisión ofrecen visualizar las reglas conformadas, lo que permite que un proceso biológico complejo modelado con esta técnica, pueda conocer lo que esta pasando

internamente y como se están tomando las decisiones, algo importante en el área de la Biología.

La creación del modelo de predicción exones con más detalle se puede apreciar en el **capitulo 5, sección 5.1**.

3.1.5 Etapa de evaluación

La evaluación del modelo se hizo mediante la herramienta Weka con un conjunto de prueba de 2400 secuencias constituido por los diferentes atributos establecidos en la **etapa de preparación de los datos**.

Esta evaluación permite dar como premisa inicial que las medidas estadísticas pueden lograr una probabilidad de clasificar secuencias codificantes cercano al 80%, que las medidas fractales mejoran un poco la probabilidad de clasificar exones, cercana al 88% y, la combinación de medidas estadísticas y fractales se obtienen resultados cercanos al 91% de probabilidad de clasificación de secuencias codificantes (para más detalle de la evaluación del modelo ver **capitulo 5, sección 5.2**).

3.2 Metodología de desarrollo de Software

Programación extrema es una metodología ágil de desarrollo de Software, haciendo mayor énfasis en la adaptabilidad que en la previsibilidad, es decir, el proceso se debe adaptar a los cambios que se presenten durante la ejecución en lugar de predefinir todos los cambios de requisitos al inicio del proyecto ya que es desgastante y que por cualquier u otra razón siempre hay ajustes sobre el software, sin caer en el error de aceptar cambios al gusto del cliente (pues, nunca se terminaría) (Beck, 1999). Se adoptó esta metodología principalmente porque el equipo de trabajo estaba formado por dos desarrolladores y el director del proyecto, además de basarse en desarrollo de prototipos permitió que la codificación de cada una de las medidas se pudiera evaluar de manera rápida y probar su efectividad al momento de su implementación.

Los puntos que se adjudicaron para evaluar la dificultad de cada una de las historias de usuario esta entre 1 y 5, siendo 5 la de mayor dificultad y que requería más dedicación en recursos.

Para el desarrollo de la aplicación que interactúa con el usuario (Interfaz Gráfica de Usuario), se ajustó a un ciclo de desarrollo.

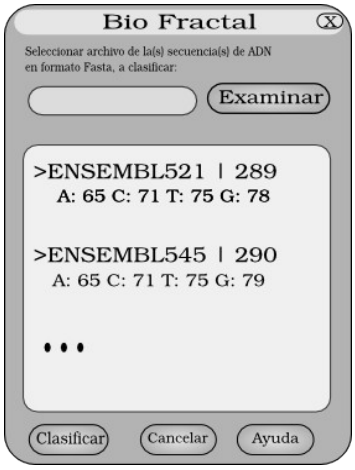
3.2.1 Planificación del proyecto

Planificación inicial:

El equipo de desarrollo esta compuesto principalmente por dos desarrolladores (también hacen tareas de testers), un manager (director del proyecto) y un usuario de profesión Biólogo (cliente).

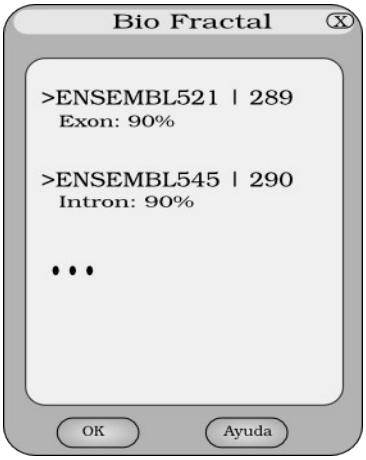
Prototipo de interfaces:

Selección archivos de secuencias de ADN:



Gráfica 20: Presentación de resultados de la clasificación

Presentación de resultados:



Gráfica 21: Seleccionar secuencias de ADN

Historias de usuario:

No	Nombre Historia	Prioridad	Riesgo	Esfuerzo	Iteración
	Implementación de las medidas estándar para la clasificación de secuencias codificantes				
1	Implementación y prueba uso de codón	Media	Bajo	4	1
2	Implementación y prueba de Ficket	Media	Bajo	4	1
3	Implementación y prueba Contenido G+C	Media	Bajo	4	1
4	Implementación y prueba de Hexámeros	Media	Bajo	4	1
	Implementación de las medidas fractales para la clasificación de secuencias codificantes				
5	Implementación y prueba análisis R/S	Alta	Alto	5	2
6	Implementación y prueba Zipf-Mandelbrot	Alta	Alto	5	2
7	Implementación y prueba análisis bidimensional	Alta	Alto	5	2
	Implementación Modelo de clasificación de exones				
8	Implementar Modelo de clasificación	Alta	Alto	4	3
	Interfaces Gráficas de Usuario				
9	Implementar GUI	Baja	Bajo	2	4

Las plantillas de las historias de usuario se encuentran en el **Anexo D**.

Plan de entrega:

Primera Iteración: se implementaron las medidas estándar para la clasificación de secuencias codificantes, este proceso se lleva a cabo en la **etapa de preparación de datos** del proceso de Minería de Datos.

1) Implementación de la medida Uso de codón: de una secuencia en formato Fasta se identifica la frecuencia de codones (ver sección 2.2.1)

Tareas:

- Comprobación del formato Fasta de las secuencias de ADN.
- Calculo de Uso de codón (Guigó, 1998).
- Pruebas y comprobación de resultados.

2) Implementación de la medida Ficket: proceso de identificación de patrones en las secuencias de ADN por método de Ficket (ver sección 2.2.3)

Tareas:

- Comprobación del formato Fasta de las secuencias de ADN.
- Calculo de Ficket (Fickett y Tung, 1992).
- Pruebas y comprobación de resultados.

3) Implementación de contenido GC: identificación del contenido de Guanina y Citosina en las secuencias de ADN (ver sección 2.2.4)

Tareas:

- Comprobación del formato Fasta de las secuencias de ADN.
- Identificación de contenido GC (Koski, 2001).
- Pruebas y comprobación de resultados.

4) Implementación de hexámeros : identificación del contenido y frecuencia de hexámeros en las secuencias codificantes y no codificantes (ver sección 2.2.2)

Tareas:

- Comprobación del formato Fasta de las secuencias de ADN.
- Identificación de hexámeros (Claverie y Bougueleret, 1986).
- Pruebas y comprobación de resultados.

Segunda Iteración: implementación de medidas fractales para la clasificación de secuencias codificantes: este proceso se llevó a cabo en la **etapa de preparación de datos** del proceso de Minería de Datos.

5) Implementación del análisis R/S: búsqueda de patrones mediante análisis R/S tanto en las secuencias codificantes como no codificantes (ver sección 3.1.3.2)

Tareas:

- Preparación de las secuencias de ADN.
- Identificación de patrones mediante el análisis R/S (Yu y Chen, 1999).
- Pruebas y comprobación de resultados.

6) Implementación de la medida Zipf-Mandelbrot: análisis de la distribución de rangos y frecuencias de las longitudes de las secuencias de ADN (ver sección 3.1.3.4)

Tareas:

- Preparación de las secuencias de ADN.
- Identificación de patrones mediante Zipf-Mandelbrot (Mandelbrot, 1993).
- Pruebas y comprobación de resultados.

7) Implementación del análisis bidimensional: búsqueda de patrones mediante el análisis bidimensional en las secuencias codificantes (ver sección 3.1.3.3)

Tareas:

- Preparación de las secuencias de ADN.
- Identificación de patrones mediante análisis bidimensional (Xiao et al., 1995).
- Pruebas y comprobación de resultados.

Tercera Iteración:

8) Implementación del modelo de clasificación: El modelo obtenido en la Minería de Datos se implementa en el lenguaje de programación Python.

Tareas:

- Seleccionar el modelo con mejores resultados, obtenido en la **etapa de modelamiento** del proceso de Minería de datos (3 puntos)

- Codificar el modelo (4 puntos)

Cuarta Iteración:

9) Implementación de la Interfaz Gráfica de Usuario (GUI): esta iteración consta de las siguientes fases:

a) Seleccionar archivo de las secuencias de ADN: El usuario accede a la aplicación, donde selecciona la(s) secuencia(s) a clasificar, estas deben estar en formato Fasta. Inmediatamente aparece la información de las secuencias seleccionadas como nombre, longitud, cantidad individual de nucleótidos (en porcentajes)

Tareas:

- Diseño interfaz selección de secuencia (1 puntos)
- Comprobación de entradas (1 puntos)
- Listar la información de las secuencias seleccionadas: nombre de la secuencia, longitud (1 puntos).

b) Clasificar secuencias de ADN: El usuario al presionar la opción "Clasificar", da inicio al proceso de clasificación de las secuencias Fasta. Este proceso consta de llamados a la Interfaz de Aplicaciones (API) del modelo de predicción implementado.

Tareas:

- Llamados a las interfaces de aplicación del modelo de predicción (1 punto)
- Comprobación de parámetros (1 punto)

c) Presentar resultados de la clasificación: El usuario vera la lista de las secuencias con el resultado de la clasificación.

Tareas:

- Diseño de la interfaz de presentación de resultados (1 punto)
- Comunicación con la interfaces del modelo de predicción (1 punto)

3.2.2 Implementación

Medidas de predicción de exones:

En la **etapa de preparación de datos** del proceso de Minería de Datos se implementa cada una de las medidas estadísticas estándar (Ficket, Uso de codón, hexámeros y contenido de

GC) y las medidas fractales (análisis reescalado R/S, ley de Zipf-Mandelbrot y análisis bidimensional) con la metodología planteada en el capitulo 3.

Modelo de clasificación:

El modelo obtenido en la etapa de **modelamiento** del proceso de Minería de Datos se implementó en el lenguaje de programación Python. El modelo conformado por reglas de decisión dentro de un árbol se generaron con la herramienta de Aprendizaje de Máquina, Weka; aplicando el algoritmo de "ganancia de información o reducción de entropía", J4.8 (Witten y Frank, 2005).

Clasificación de secuencias:

La clasificación de secuencias se divide en las siguientes actividades: seleccionar la(s) secuencia(s) a clasificar la(s) cual(es) debe(n) estar en formato Fasta, clasificación de las secuencias insertadas y mostrar información del resultado de la clasificación.

Para ello la aplicación cuenta con una ventana donde se presenta la opción de seleccionar el archivo en formato Fasta de la(s) secuencia(s). Bajo esta opción se presenta la información de las secuencias que se van a clasificar tal como nombre, longitud y cantidad pares de bases. Una vez seleccionada las secuencias el usuario puede hacer uso del botón "Clasificar" el cual ejecuta el algoritmo de predicción tomando la(s) secuencia(s) insertada(s) (Gráfica 21). Al terminar el proceso de clasificación, se presenta la información del resultado en otra ventana (Gráfica 20).

Código fuente:

El lenguaje de programación usado para implementar los índices o medidas de clasificación (Uso de codón, hexámeros, cpg, ficket, Hurst, Zipf, Bidimensional) fue Python.

El árbol de clasificación se modeló en la herramienta Weka (Witten y Frank, 2005).

La interfaz gráfica de aplicación se implemento en PyGtk 2.0

3.2.3 Pruebas

3.2.3.1 Casos de prueba: Implementación de las medidas estándar para la predicción secuencias codificantes

Descripción: en la **etapa de preparación de datos** del proceso de Minería de Datos se implementó y probó la efectividad de las medidas con el mismo conjunto con el que se obtuvo los parámetros o atributos, que en este caso fue el cromosoma 19 de *M. musculus;* luego se aplicó sobre los demás cromosomas y genomas. (Este proceso se describe con más detalle en la sección 2.2 del Marco Teórico y las secciones 4.2; 4.3; 4.6 y 4.7 de resultados).

Resultados con valores de sensibilidad altos:

- Condiciones de ejecución: Tener secuencias en formato Fasta y los parámetros obtenidos del cromosoma 19 de *M. musculus*.
- Entrada: tabla con los parámetros de clasificación y secuencias en formato Fasta.
- Resultado esperado: Valores altos de sensibilidad y valores bajos de especificidad.
- Evaluación de la prueba: **valor se sensibilidad de cada una de las medidas, en promedio, del 80%.**

3.2.3.2 Caso de prueba: Implementación de las medidas fractales para la predicción de secuencias codificantes

Descripción: en la etapa de preparación de datos del proceso de Minería de Datos se implementaron y se probaron las medidas fractales para medir la sensibilidad y especificidad, de manera individual sobre el mismo conjunto de donde se obtuvieron los parámetros, en este caso dicho conjunto fue el cromosoma 19 de *M. musculus*, luego sobre los demás cromosomas y genomas (Este proceso se describe con más detalle en la sección 2.3.6 del Marco Teórico y las secciones 3.1.3.2; 3.1.3.3; 3.1.3.4 de la Metodología).

Resultados con valores de sensibilidad altos:

- Condiciones de ejecución: Tener secuencias en formato Fasta y la tabla de parámetros obtenidos del cromosoma 19 de *M. musculus*.
- Entrada: tabla con los parámetros de clasificación y secuencias en formato Fasta.
- Resultado esperado: Valores altos de sensibilidad (mayores a 80%) y valores bajos de especificidad.
- Evaluación de la prueba: **valor se sensibilidad, en promedio, del 90%. A excepción del análisis de la ley de Zipf-Mandelbrot cuyos valores de sensibilidad fueron muy bajos.**

3.2.3.3 Casos de prueba: Implementación del modelo de clasificación

Descripción: **e**l conjunto de reglas generadas por Weka, se mapearon con las sentencias **if-elif-else** del lenguaje de programación Python, y se comparó el resultado obtenido con el arrojado por Weka.

Clasificación correcta de secuencias ADN:

- Condiciones de ejecución: Tener secuencia en formato Fasta.

- Entrada: La secuencia de prueba se pasa como parámetro al modelo implementado para comparar dicho resultado con el arrojado por Weka.
- Resultado esperado: Resultado igual o similar comparado con el arrojado por Weka.
- Evaluación de la prueba: **prueba satisfactoria**

3.2.3.4 Casos de prueba: Seleccionar archivos de secuencias de ADN

Descripción: en esta historia hay que comprobar la introducción del archivo de secuencias de ADN. Si el formato del archivo de secuencias no es correcta (no sigue el formato indicado o los valores son incorrectos respecto al estándar Fasta) se avisa al usuario y no se procesan las secuencias incorrectas.

Introducción de secuencias correctas:

- Condiciones de ejecución: Tener la secuencia en un archivo con formato Fasta. Si son varias secuencias a clasificar deben estar en un único archivo.
- Entrada: En la ventana principal se mostrará un botón para seleccionar el archivo con la(s) secuencia(s) de ADN a procesar. Al seleccionar el archivo, se procesará internamente y se mostrará un mensaje indicando que secuencias se han ingresado. El proceso de introducción de secuencias se considera finalizado.
- Resultado esperado: Una vez seleccionado el archivo de secuencias de ADN, muestra un mensaje con la información de las secuencias como nombre y longitud.
- Evaluación de la prueba: **prueba satisfactoria**

Introducción de secuencias de ADN con errores:

- Condiciones de ejecución: Tener la secuencia en un archivo con formato Fasta con errores. Si son varias secuencias a clasificar deben estar en un único archivo.
- Entrada: Tras seleccionar el archivo con la(s) secuencia(s) se verifica que cumpla con el formato Fasta formado por los nucleótidos A, C, G y T o que tenga nucleótidos desconocidos representados por N.
- Resultado esperado: Se muestra un mensaje de error y no se tiene en cuenta el archivo seleccionado.
- Evaluación de la prueba: **prueba satisfactoria**

Introducción de archivos inválidos (distintos de Fasta):

- Condiciones de ejecución: Tener el archivo con secuencias invalidas.

- Entrada: Luego de haber seleccionado el archivo de las secuencias, se hace un pre-procesamiento interno y se verifica que el archivo esta en formato incorrecto y/o que no corresponda con un archivo de secuencias de ADN.
- Resultado esperado: Se muestra un mensaje de error y no tiene en cuenta el archivo ingresado.
- Evaluación de la prueba: **prueba satisfactoria**

3.2.3.5 Casos de prueba: Clasificar secuencia de ADN

Descripción: al seleccionar la opción "clasificar" se debe verificar que haya seleccionado algún archivo, de lo contrario se avisa al usuario que seleccione uno. En segundo plano se calcula, para cada secuencia seleccionada, los valores relacionados con las medidas estándares y fractales que sirven para hacer la clasificación con base en ellas.

Se ha seleccionado un archivo valido:

- Condiciones de ejecución: Haber seleccionado un archivo con las secuencias de ADN.
- Entrada: Como procesamiento interno, el archivo de secuencias se discretizan de tal forma que a la interface del modelo de clasificación se le pase secuencia por secuencia, para su procesamiento.
- Resultado esperado: Mensaje de espera mientras procesa. Mensaje de terminado.
- Evaluación de la prueba: prueba satisfactoria

No se ha seleccionado archivo:

- Condiciones de ejecución: No haber seleccionado un archivo con las secuencias de ADN.
- Entrada: El usuario presiona el botón "clasificar" sin haber seleccionado un archivo, o haber ingresado el path del mismo en la caja de texto.
- Resultado esperado: Mensaje de que seleccione un archivo.
- Evaluación de la prueba: prueba satisfactoria

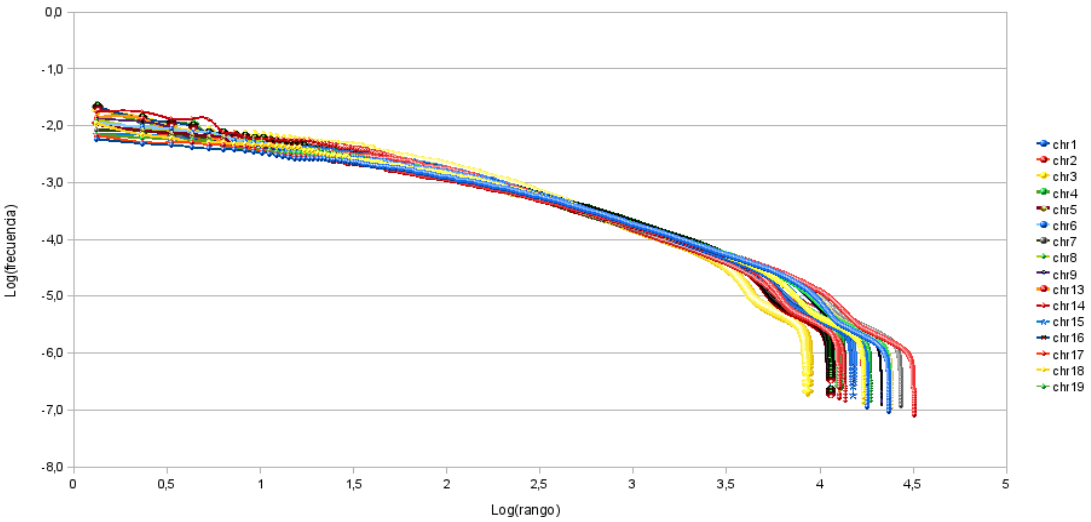
3.2.3.6 Casos de prueba: Presentar resultados de la clasificación

Descripción: En la presentación de resultados se visualiza la decisión que tomó el modelo de clasificación en cada una de las secuencias de ADN.

4. Análisis de los resultados de las medidas estadísticas y fractales

4.1 Ley de Zipf-Mandelbrot

La Gráfica 22 muestra el comportamiento de la jerarquización de las UI por cromosoma del genoma de *M. musculus*, donde se visualiza una tendencia similar de ordenamiento de las UI a través de los diferentes cromosomas. El cromosoma más interno corresponde al más pequeño en longitud total, mientras que el cromosoma más grande es el más externo en la gráfica.



Gráfica 22: Ley de Zipf-Mandelbrot. Distribución rango-frecuencia de los 16 cromosomas de *M. musculus*

Un análisis detallado a nivel del cromosoma 19 de *M. musculus* permite entender estructuralmente el genoma a través del ordenamiento de las UI en los cromosomas. El cromosoma 19 contiene 736 genes con 12710 UI, la proporción de UI es similar: 6723 exones (53%) y 5987 intrones (47%); sin embargo, los intrones contienen el 93% del total de bases nitrogenadas del cromosoma (Cuadro 14). El total de genes interrumpidos (GI) (con 2 o más exones) fue 619 que abarcan el 99,5% del total de bases que codifican el cromosoma. El 16% de los genes (117) son genes simples (GSI), es decir, están conformadas por una sola unidad de información (exón) y contienen el 0.5% del total de bases del cromosoma 19 (Cuadro 13). Estas proporciones fueron similares en todos los cromosomas del genoma de *M. musculus* (Cuadro 14). Sin embargo, estas proporciones

fueron específicas para cada especie estudiada y de acuerdo a su nivel de evolución (Cuadro 15).

Cuadro 13: Estadística descriptiva del cromosoma 19 de *M. musculus*. Genes interrumpidos, GI (exones+intrones); Genes simples, GSI (1 exón); UI (Unidades de Información)

	#	Long. totl	Long. Prom.	Desv. St.	Max (Long)	Min(Long)	Pendiente (ZM)	R^2
Genes	736	26806954	36422.492	82189.995	1197272	66	a:-1.5032 b:0.277	0.79
Exones	6723	1783168	265.234	488.683	7431	7	a:-0.8421 b:-1.2084	0.94
Intrones	5987	25023786	4179.687	15191.268	558120	5	a:-1.461 b:0.5086	0.86
totl UI	12710	26806954	2109.123	10613.660	558120	5	**a:-1.5296 b:0.765**	0.94
GI								
Genes	619	26674294	43092.559	88045.731	1197272	520	a:-1.2558 b:-0.2142	0.85
Exones	6606	1650508	249.850	468.399	7431	7	a:-0.8106 b:-1.3019	0.94
Intrones	5987	25023786	4179.687	15191.268	558120	5	a:-1.461 b:0.5086	0.86
totl UI	12593	26674294	2118.184	10662.186	558120	5	**a:-1.534 b:0.7735**	0.94
GSI								
Genes	117	132660	1133.846	753.017	3927	66	**a:-0.6101 b:-1.1465**	0.67
Exones	117							
Intrones								
totl UI	117							

El Cuadro 13 muestra algunas características del cromosoma 19 de *M. musculus* como por jejemplo que contiene 6723 exones y en su conjunto están compuestos por 1'783.168 nucleótidos, que la longitud promedio de los exones es de 265,23 nucleótidos y que dicha longitud varia mas o menos 488,68 nucleótidos con respecto al promedio, que la máxima longitud de un exón del cromosoma es de 7431 nucleótidos y la mínima longitud es de 7 nucleótidos, que este conjunto de exones tiene una pendiente negativa de 0.84 con coeficiente de confianza del 94%.

Cuadro 14: Relación cantidad de UI, % de bp (par base) y DF (Dimensión fractal) por cada cromosoma del genoma *M. musculus*

Cromosomas	Exones		Intrones		DF
	#	% bp	#	% bp	
chr18	4569	4,73	4043	95,27	0,6
chr16	6060	5,37	5363	94,63	0,62
chr13	6822	5,43	5936	94,57	0,62
chr19	6833	6,77	6076	93,23	0,66
chr14	7300	4,59	6436	95,41	0,61
chr15	8016	6,12	7186	93,88	0,64
chr3	9193	6,03	8116	93,97	0,63
chr17	9373	7,23	8274	92,77	0,67
chr6	9670	5,13	8475	94,87	0,62
chr8	9947	6,12	8831	93,88	0,64
chr9	11392	6,45	10127	93,55	0,64
chr1	12559	5,44	11078	94,56	0,62
chr4	12539	6,28	11141	93,72	0,64
chr5	12814	6,11	11488	93,89	0,65
chr7	14723	8,28	12681	91,72	0,68
chr2	17050	6,52	15116	93,48	0,64

El Cuadro 14 muestra relaciones entre el porcentaje de nucleótidos y DF de cada cromosoma del genoma de *M. musculus* donde el cromosoma 18 tiene 4569 exones que conforman el 4,73% del total de nucleótidos o pares de bases del cromosoma y 4043 intrones que conforman el 95,27% del total de nucleótidos del cromosoma con una DF de 0,6.

Las pendientes de las curvas hiperbólicas (Cuadro 13, columna pendiente(ZM)), tanto de genes como de UI indican que la pendiente es directamente proporcional al nivel de complejidad de la estructura en estudio: -0,84 para exones, -1,46 para intrones, -1,5 para genes. Es decir a medida que el valor de la pendiente aumenta, la riqueza del vocabulario es mayor, hay mayor cantidad de información codificada por cada una de las estructuras y mayor uso de secuencias de exones con longitudes pequeñas (Cuadro 14). Sin embargo, estas pendientes son específicas para cada especie dependiendo del nivel de complejidad evolutivo de éstas (Cuadro 16), a medida que la especie es menos evolucionada las pendientes tienden a ser mayores, por ejemplo para *H. sapiens* la pendiente es de -1.65 mientras que para *A. thaliana* es de -0,769 mucho mayor que la del primero.

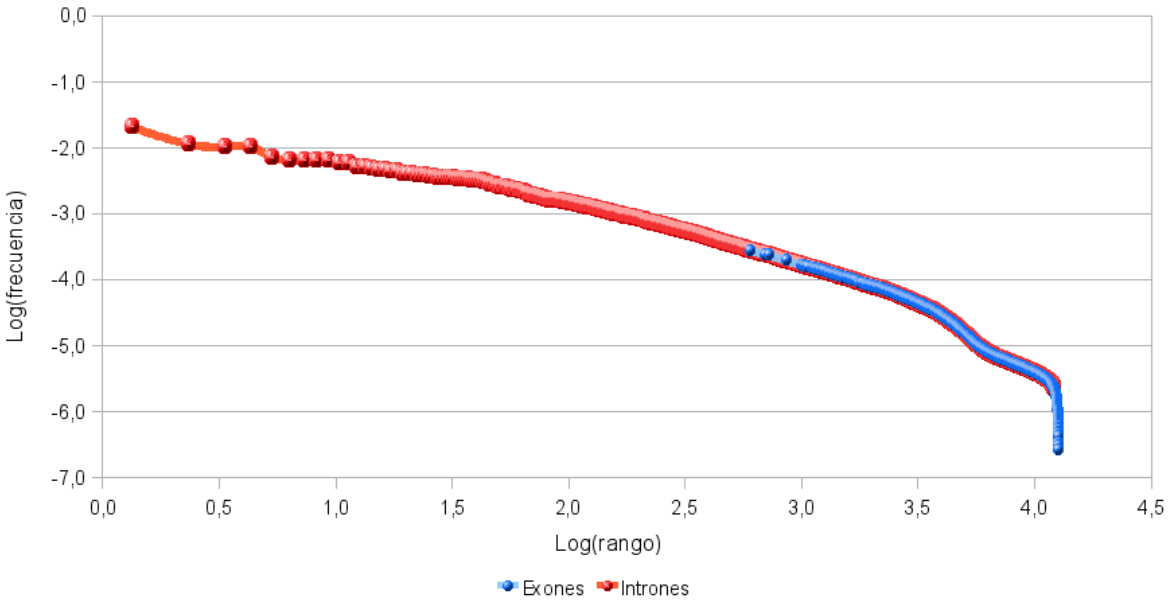
Cuadro 15: Comparación de la cantidad de UI y porcentaje de pares de bases (%bp) para los genomas en estudio.

	Exones		**Intrones**	
	#	**% bp**	**#**	**% bp**
H. sapiens	66708	5,33	55991	94,67
M. musculus	158860	6,07	140367	93,93
G. gallus	69980	13,38	56851	86,62
D. melanogaster	58043	38,66	43404	61,34
C. elegans	132240	47,37	104891	52,63
S. cerevisiae	7258	99,41	292	0,59
O. sativa	133605	48,05	106895	51,95
A. thaliana	143555	73,85	85298	26,15

Cuadro 16: Pendiente de la Ley Zipf-Mandelbrot, coeficiente de determinación (R^2) y Dimensión Fractal (DF) de los genomas estudiados

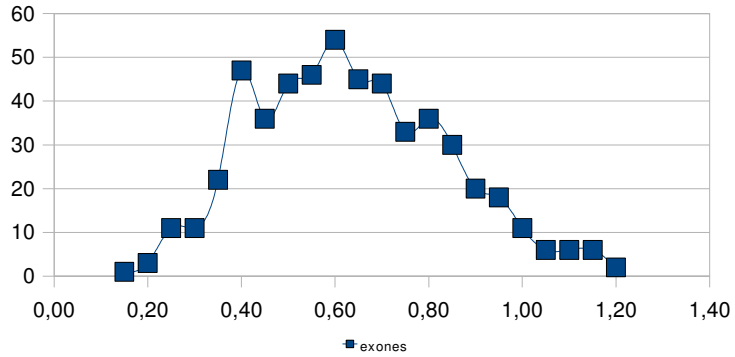
	pendiente	**R^2**	**DF**
H. sapiens	-1,6559	0,9258	0,6
M. musculus	-1,5661	0,9284	0,64
G. gallus	-1,2327	0,918	0,81
D. melanogaster	-1,2201	0,9373	0,82
C. elegans	-0,9414	0,8976	1,06
S. cerevisiae	-0,8552	0,6114	1,17
O. sativa	-0,9587	0,9031	1,04
A. thaliana	-0,769	0,9402	1,3

En un análisis detallado de los rangos de Zipf y frecuencias de longitudes a nivel de genoma, se observa que los primeros rangos pertenecen a intrones y los últimos rangos a exones, con un trayecto de transición donde se mezclan los dos tipos de UI (Gráfica 23). Esto se relaciona con el estudio de Zipf sobre textos, donde palabras que conectan las frases están en primer lugar, y son las que le dan sentido al mensaje de la frase, con el hecho que los intrones están en medio de los exones y son ellos los que barajan la combinación de exones para dar como resultado una proteína.

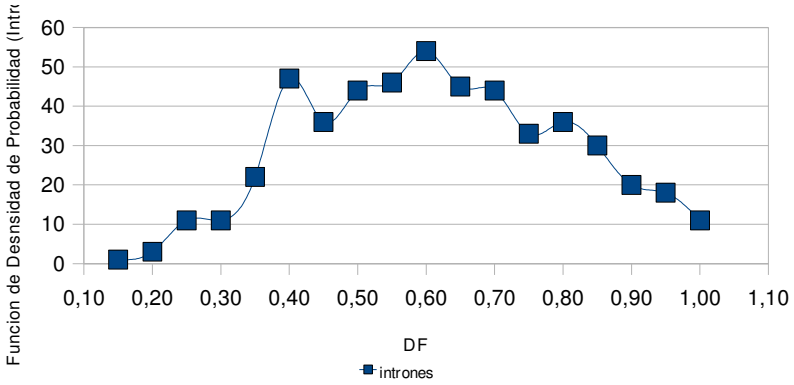


Gráfica 23: Distribución de exones e intrones en el cromosoma 19 de *M. musculus*

El análisis densidad de probabilidad de las UI del cromosoma 19 de *M. musculus*, muestran que están se distribuyen normalmente con simetría y curtosis (Gráfica 24), esta propiedad se cumple a nivel de los genomas en estudio. Las dimensiones fractales (DF) obtenidas para los 619 GI del cromosoma 19 de *M. musculus,* se discriminaron en rangos de 0.05 obteniendo para los GI (0.05<Rango DF<1.55) un 45% de los datos fue menor de 0.5, mientras que el 50% varió entre 0.5 y 1.0 y el 5% restante es no fractal (Anexo E). Para los exones (0.10<Rango DF<5.15) el 17% fue menor de 0.5, 47% varió entre 0.5 y 1.0. Para los intrones (0.10<Rango DF<19.10) el 31% fue menor de 0.5, el 60% varió entre 0.5 y 1.0. Esta tendencia fue consistente a través de todo el genoma de *M. musculus*, al igual que para todos los genomas.



Exones: Asimetría: 0,19; Curtosis: -1,54



Intrones: Asimetría: -0,19; Curtosis: -1,32

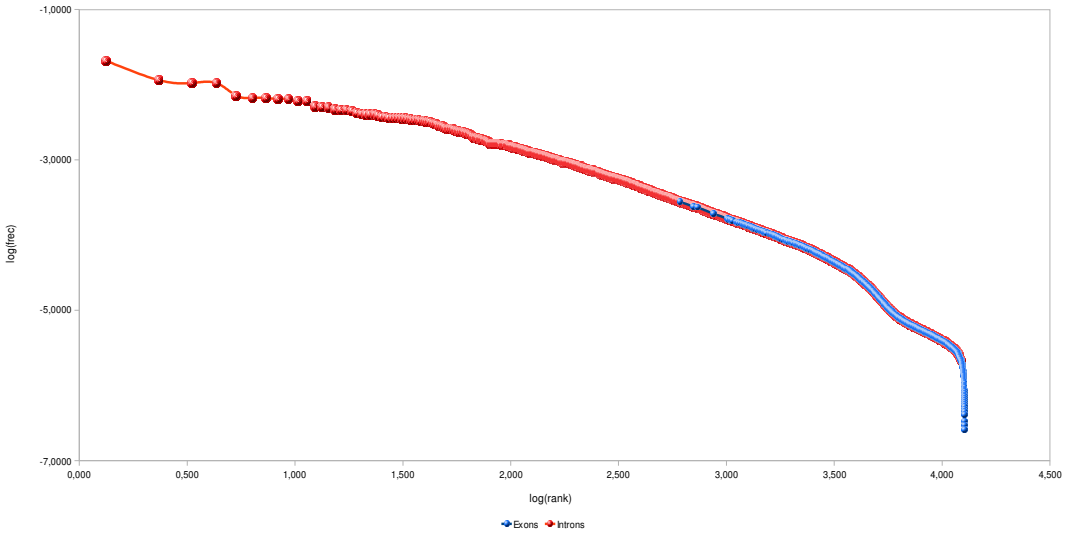
Gráfica 24: Función de densidad de probabilidad para exones e intrones del cromosoma 19 de *M. musculus*.

La eficiencia de los parámetros de clasificación del algoritmo que se obtuvieron a partir del análisis general y detallado del genoma de *M. musculus*, se evaluó mediante las medidas de sensibilidad (Sn) y especificidad (Sp) (sección 2.7.1). Con los parámetros (longitud , rango y frecuencia) obtenidos a través del cromosoma 19 de *M. musculus*, y aplicados en este genoma se obtuvo una Sn y Sp del 71% y 77% respectivamente (Cuadro 17); estas medidas disminuyen al aplicarlos en los demás genomas por varias razones, la primera que los parámetros están basados en el genoma de *M. musculus;* segundo, los genomas difieren en longitud, número de cromosomas y distribución de las UI a través de los cromosomas y tercero son especies eucariotes evolutivamente diferentes. Adicionalmente, en algunos genomas como *G. gallus* la información en la BD todavía no esta completa.

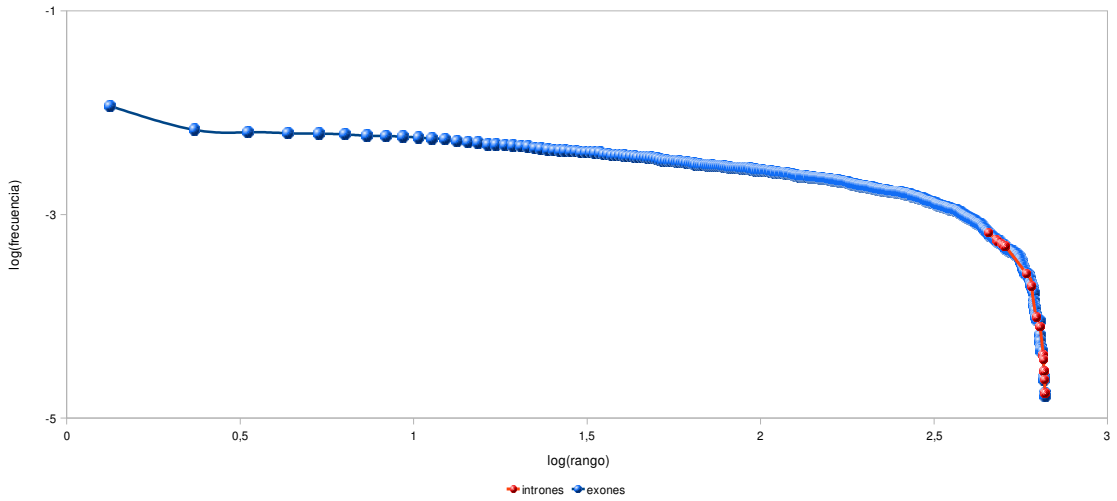
Cuadro 17: Resultados de la Ley de Zipf-Mandelbrot por genoma mediante las medidas de Sensibilidad (Sn) y Especificidad (Sp).

	Sn	Sp
H. sapiens	0,3702	0,8890
M. musculus	0,7083	0,7715
G. gallus	0,0436	0,9920
D. melanogaster	0,1010	0,9263
C. elegans	0,0004	0,9999
S. cerevisiae	0,0000	1,0000
O. sativa	0,0003	0,9999
A. thaliana	0,0004	1,0000

La Gráfica 25 ilustra la distribución general del cromosoma 19 de *M. musculus* donde se presentan regiones definidas para exones e intrones; los exones están en la parte inferior de la curva mientras que los intrones están en la parte superior, esto contrasta con las regiones de exones e intrones del cromosoma 15 de *S. cerevisiae* (Gráfica 26), donde los exones están en la parte superior y los intrones en la parte inferior. Esto se explica en el hecho que *M. musculus* es un organismo superior frente a *S. cerevisiae*. En los Anexos F y G se muestra el comportamiento en los demás genomas.



Gráfica 25: Distribución de exones e intrones a lo largo del cromosoma 19 de *M. musculus*



Gráfica 26: Distribución de exones e introns a lo largo del cromosoma 15 de *S. cerevisiae*

4.2 Uso de codón

Una característica universal presente en cualquier genoma es el uso igual de codones en secuencias codificantes. Esta característica se puede usar para diferenciar regiones codificantes de no codificantes del genoma (Maniatis, 1982). Con base en el procedimiento establecido por Guigó (Guigó, 1998) se desarrolló una tabla de uso de codón para cada secuencia. Este procedimiento se realizó para secuencias en fase 0 (secuencia cuya longitud es múltiplo de 3), los resultados para el cromosoma 19 de *M. musculus* fueron todos positivos en exones y todos negativos para introns en él, es decir, los valores de los exones variaron entre 0 a 1, mientras los valores en introns variaron entre 0 y -1. A nivel de genoma de *M. musculus* el 69% de los exones presentaron valores positivos y 75% de lo introns presentaron sus respectivos valores negativos (Cuadro 18). Todos los exones del genoma de *A. thaliana* presentaron valores positivos mientras que *S. cerevisiae* tiene menor porcentaje de exones con valores positivos. El Cuadro 18 presenta una tendencia proporcional entre el porcentaje de exones con valores positivos y el porcentaje de introns con valores negativos; por ejemplo el 78% de los exones e introns presentan valores positivos y negativos respectivamente en el genoma de *H. sapiens*. Estos porcentajes decrecen a medida que los genomas animales descienden en la escala evolutiva hasta el hongo (*S. cerevisiae*). Por otra parte el genoma de *O. sativa* presenta un 58% de exones con valores positivos en contraste con *A. thaliana* que presenta el 100% de exones e introns con valores positivos. Para más detalle de los resultados obtenidos del análisis de contenido de Uso de codón ver Anexo H.

Cuadro 18: Porcentaje de los exones e intrones con valores positivos y negativos, respectivamente, en genomas eucariotes de estudio.

	Exones		Intrones	
	+ (%)	- (%)	+ (%)	- (%)
H. sapiens	0,78	0,22	0,12	0,78
M. musculus	0,69	0,31	0,25	0,75
G. gallus	0,75	0,25	0,46	0,54
D. melanogaster	0,33	0,67	0,50	0,50
C. elegans	0,83	0,17	0,17	0,83
S. cerevisiae	0,19	0,81	0,94	0,06
O. sativa	0,58	0,42	0,67	0,33
A. thaliana	1,00	0,00	1,00	0,00

4.3 Porcentaje de GC

El contenido de Guanina (G) y Citosina (C) es un parámetro estadístico universal de los genomas. Este contenido es diferente para cada especie (Nussinov, 1984). Con base en el porcentaje y razón de GC para los genomas de estudio se presentan resultados de sensibilidad y especificidad que varían entre 0,88 y 0,93. El Cuadro 19 muestra que los parámetros de contenido de G y C en las secuencias son buenos para predecir regiones codificantes (valores de sensibilidad altos), sin embargo no es bueno para predecir intrones (valores de especificidad bajos). Para más detalle de los resultados obtenidos del análisis de contenido de GC ver Anexo I.

Cuadro 19: Medidas de Sensibilidad (Sn) y especificidad (Sp) del porcentaje de GC para cada genoma en estudio.

	Sn	Sp
H. sapiens	0,931	0,344
M. musculus	0,934	0,309
G. gallus	0,939	0,187
D. melanogaster	0,734	0,387
C. elegans	0,768	0,453
S. cerevisiae	0,734	0,361
O. sativa	0,858	0,146
A. thaliana	0,886	0,182

4.4 Análisis bidimensional

La caminata bidimensional a través de la secuencia de ADN de los genomas estudiados, presentó parámetros (Rmax, Rmin, Amplitud, longitud y Dimensión teórica) con medidas

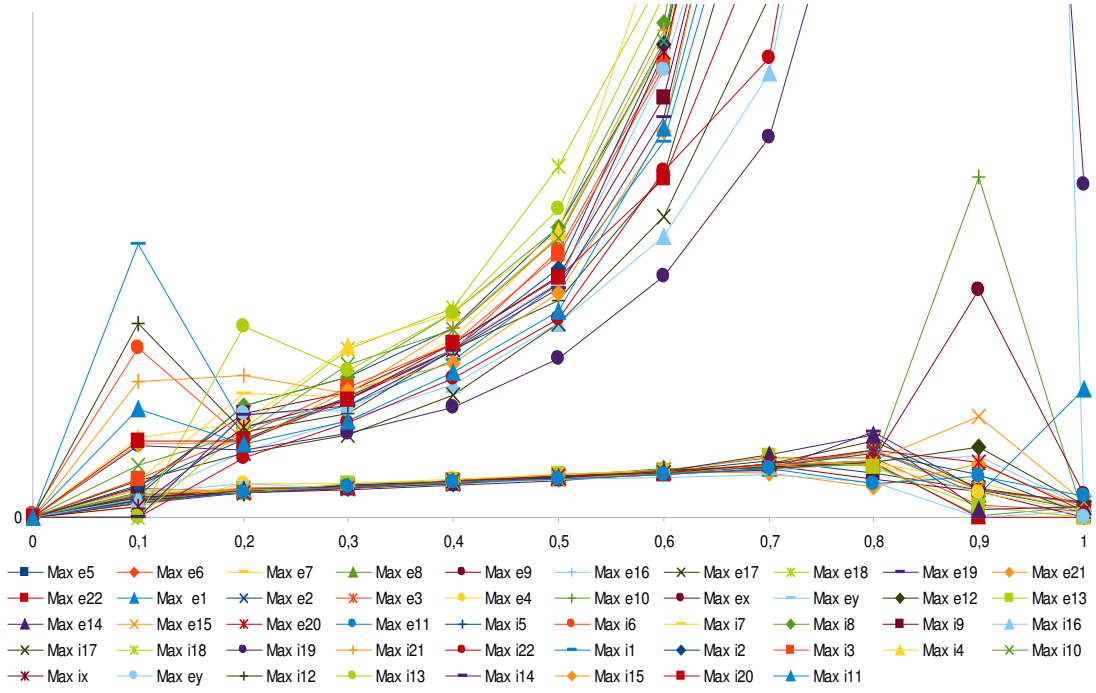
de sensibilidad altas (mayores de 94%) y especificidad bajas (menores de 0,274) (Cuadro 20). Este resultado muestra sólo la caminata en toda la secuencia de ADN, es decir, las distancias al cuadrado de un nucleótido a otro (recorriendo en pasos de un nucleótido, dos nucleótidos, etc.) son diferentes tanto para regiones codificantes como para regiones no codificantes, de tal manera que las sensibilidades y especificidades reflejan que los parámetros obtenidos del cromosoma 19 de *M. musculus* ofrecen información que ayuda a discriminarlos. Para mayores detalles de los resultados del análisis bidimensional ver **Anexo J**.

Cuadro 20: Medidas de sensibilidad (Sn) y especificidad (Sp) para los genomas estudiados, con base en análisis bidimensional

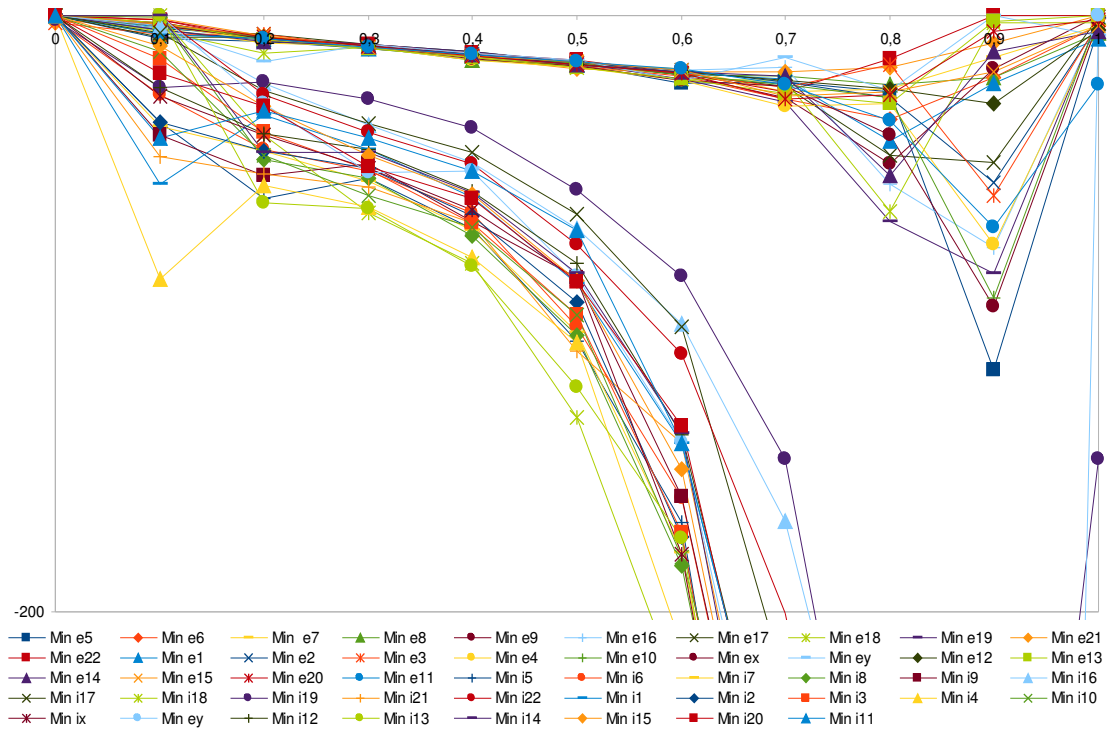
	Sn	Sp
H. sapiens	0,984	0,274
M. musculus	0,981	0,200
G. gallus	0,987	0,067
D. melanogaster	0,966	0,063
C. elegans	0,985	0,007
S. cerevisiae	0,948	0,007
O. sativa	0,975	0,033
A. thaliana	0,978	0,011

4.5 Análisis de rango Reescalado R/S

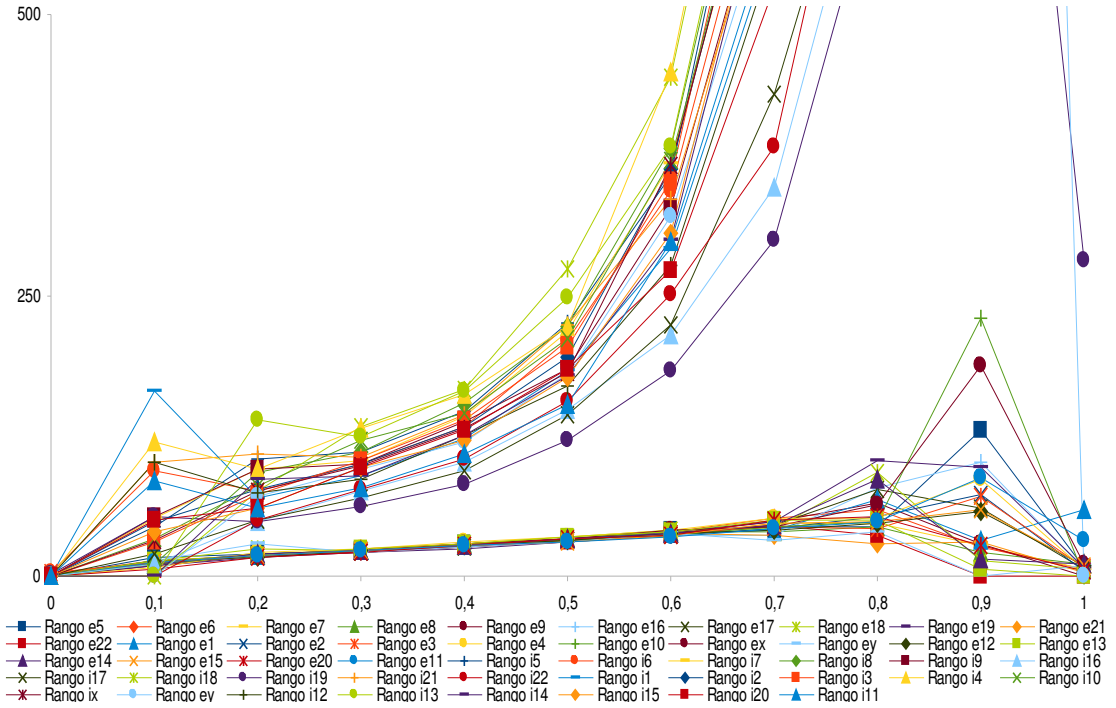
Los parámetros del análisis R/S muestran un comportamiento diferencial para los 8 genomas estudiados. En especies animales superiores (*H. sapiens, M. musculus, G. gallus, D. melanogaster, C. elegans*), el comportamiento de los parámetros del análisis R/S permiten separar claramente exones de intrones; la Gráfica 27 ilustra la distribución de los exones y de los intrones en el genoma humano a partir de los parámetros: máximo (a), mínimo (b), rango (c) y longitud (d); estos resultados indican que los parámetros R/S son buenos predictores de las diferencias entre secuencias de exones e intrones. El análisis de los parámetros en los otros genomas de estudio permitió inferir que estos parámetros son menos discriminatorios a medida que se desciende en la escala evolutiva, la estructura de los exones e intrones es cada vez menos estructurada y más aleatoria. La Gráfica 28 ilustra la distribución de los exones e intrones en el genoma de *C. elegans*, que es uno de los organismos inferiores dentro de los organismos eucarióticos superiores. Estos resultados permiten concluir que no existe un patrón común de parámetros para el genoma eucariote. El Anexo K presenta las distribuciones de exones e intrones con base en el análisis R/S para los ocho genomas estudiados.



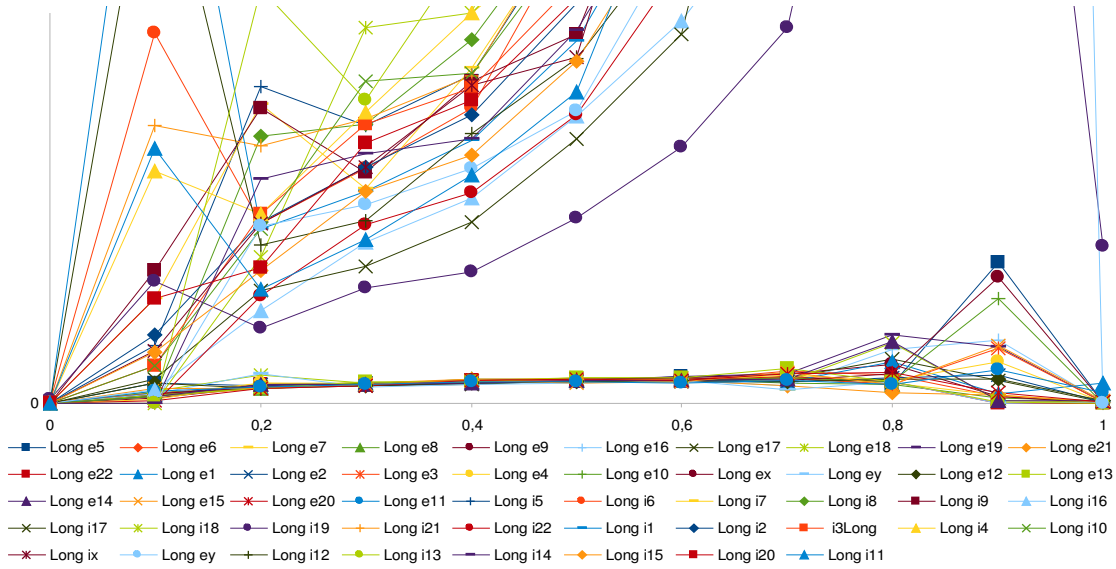
(a) Máximos del genoma *H. sapiens*



(b) Mínimos del genoma *H. sapiens*

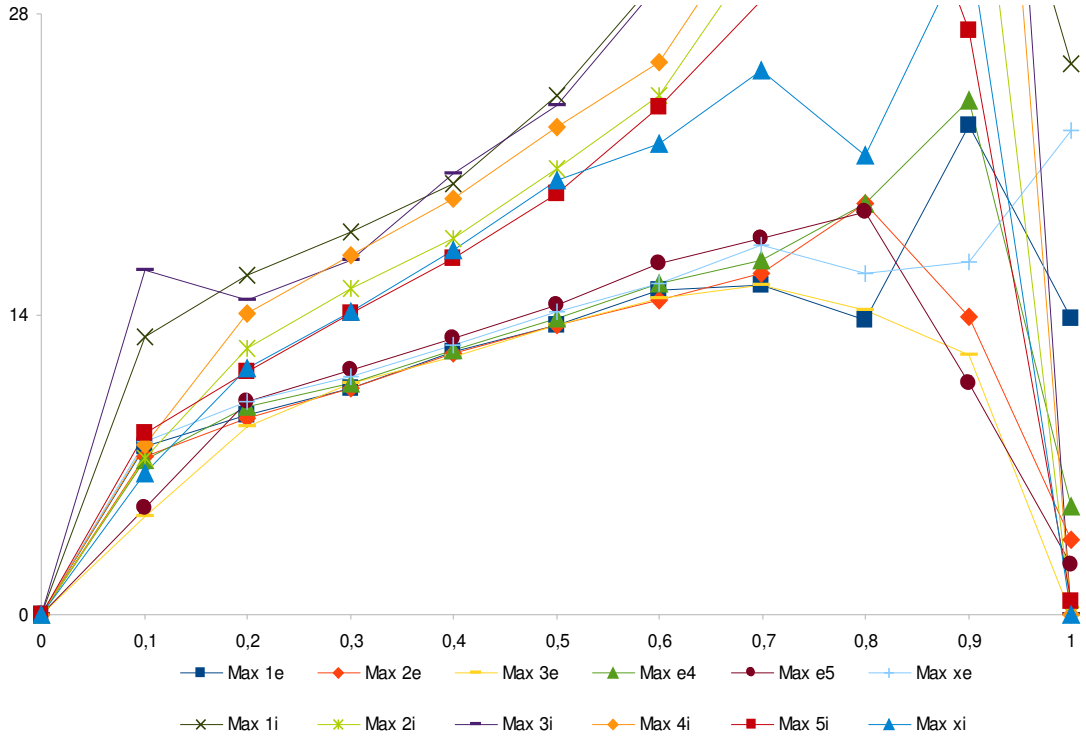


(C) Rangos para el genoma *H. sapiens*

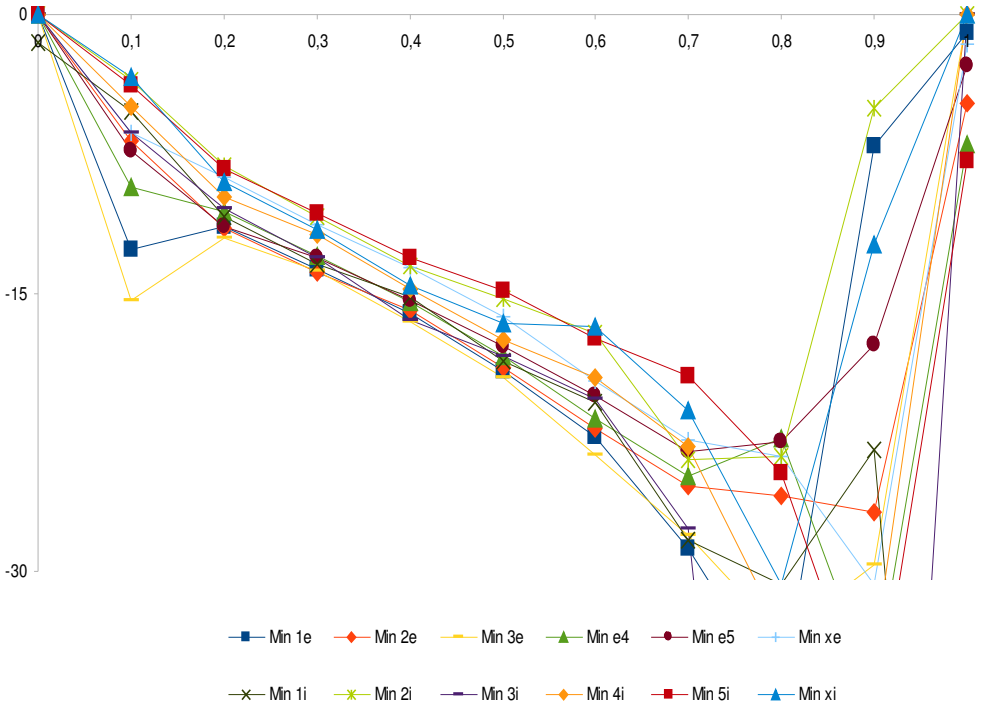


(d) Longitudes para el genoma *H. sapiens*

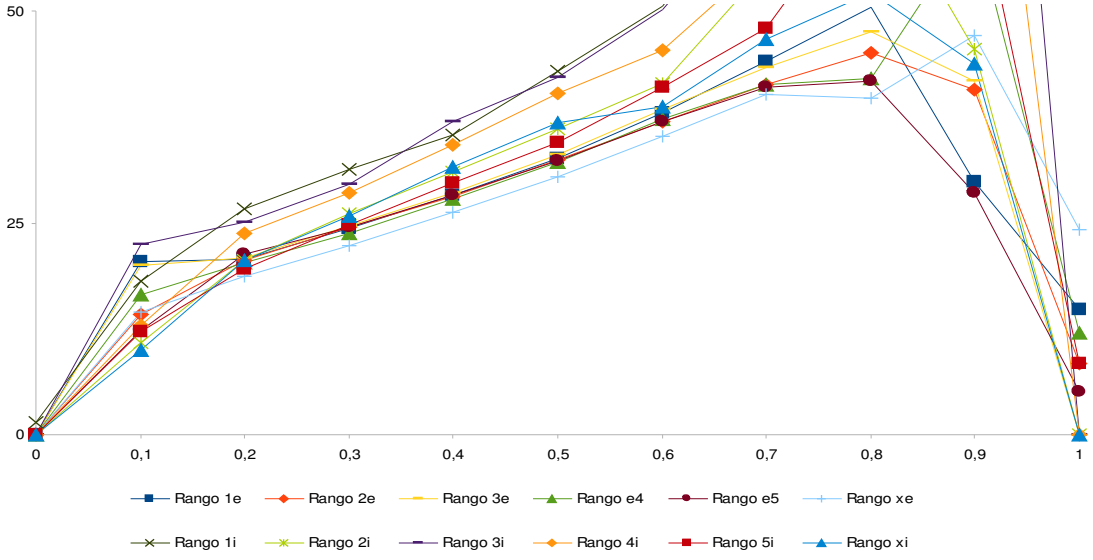
Gráfica 27: Distribución de exones e intrones del genoma humano a través de los parámetros del análisis R/S. **max**: máximo, (a); **min**: mínimo (b) y rango (c) y el parámetro **long**.: longitud (d). Eje Y: frecuencia de datos vs intervalos del coeficiente de Hurst, 0<H<1. En cada gráfica 'e' representa los exones e 'i' los intrones, el número que acompaña estas letras indicará el cromosoma al que pertenece.



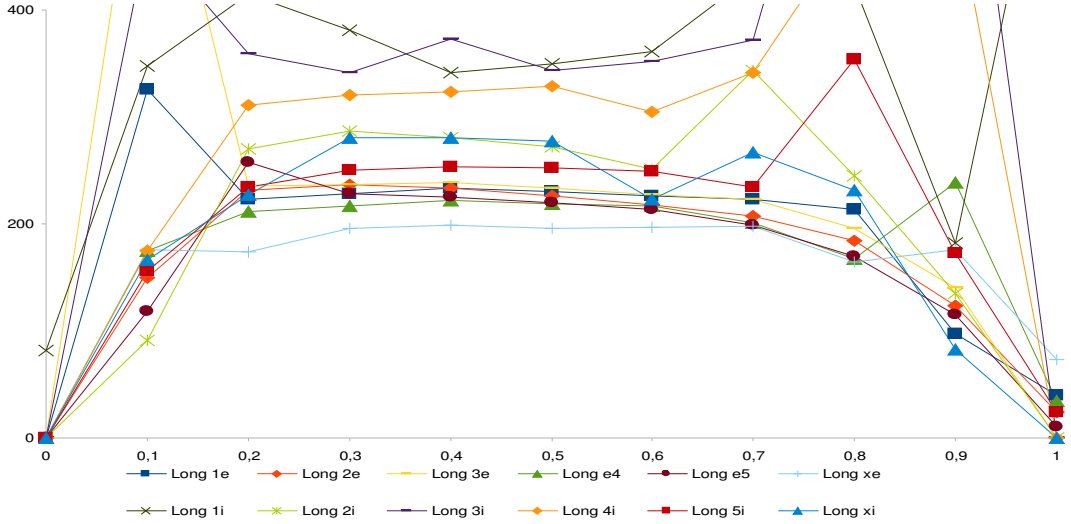
(a) Máximos para el genoma *C. elegans*



(b) Mínimos para el genoma *C. elegans*



(c) Rangos para el genoma *C. elegans*



(d) Longitudes para el genoma *C. elegans*

Gráfica 28: Distribución de exones e intrones del genoma de *C. elegans* a través de los parámetros del análisis R/S. **max**: máximo, (a); **min**: mínimo (b) y rango (c) y el parámetro **long**: longitud (d). Eje Y: frecuencia de datos vs intervalos del coeficiente de Hurst, 0<H<1. En cada gráfica 'e' representa los exones e 'i' los intrones, el número que acompaña estas letras indicará el cromosoma al que pertenece.

En especies vegetales como el arroz (*O. sativa*) de amplia manipulación y domesticación por el hombre, el comportamiento de los parámetros del análisis R/S es similar al de las especies animales superiores; sin embargo, la especie vegetal silvestre *A. thaliana* aunque

los parámetros del análisis R/S son discriminatorios para exones e intrones, los valores de los parámetros se invierten, siendo mayores para exones que para intrones pero conservando el poder de predicción de los parámetros.

4.6 Ficket

Ficket divide una secuencia en ventanas y asigna un porcentaje de codificación a cada una de acuerdo a una tabla obtenida del estudio de regiones codificantes (Fickett y Tung, 1992). Para regiones codificantes el porcentaje de la ventana es mayor a 95%, ventanas entre un 74% y 95% es incierta la codificación y no codifica si es menor a 74%; dependiendo de la fase de la secuencia se obtuvieron 3, 6 o 9 resultados, para una secuencia en fase cero se obtienen 3 resultados: el número de ventanas que alcanzaron un porcentaje de codificación mayor a 95% (FV1), el numero de ventanas entre 74 y 95% de codificación (FV2) y por último las ventanas menores a 74% de codificación (FV3), estos resultados son acumulativos si la secuencia presenta mas fases; así, una secuencia en fase cero tendrá una sola lectura y tres resultados, una en fase 1, dos lecturas con 6 resultados y una de fase dos, tres lecturas con 9 resultados. Para el cromosoma 19 de *M. musculus* aplicó el procedimiento anterior para determinar las características de exones e intrones y así poder discriminar entre secuencias codificantes y no codificantes. El siguiente cuadro (Cuadro 21) muestra que los exones presentan mayor porcentaje en las primeras ventanas de cada fase de lectura (FV1, FV4 y FV7), en contraste con los intrones, por ejemplo para exones se presenta un 30% de ventanas en FV1 (ventana 1), en contraste los intrones presentan un 6% para la misma ventana, de manera similar para las demas ventanas, este comportamiento persista en FV4 y FV7.

Cuadro 21: Calculo de Ficket para el cromosoma 19 de *M. musculus*

	FV1	FV2	FV3	FV4	FV5	FV6	FV7	FV8	FV9
exones	97696	137004	91312	59742	89660	59440	28216	46542	29977
% ventanas	0.300	0.420	0.280	0.286	0.429	0.285	0.269	0.444	0.286
intrones	488199	5460339	2043680	311449	3395019	1284758	155884	1735992	654813
%ventanas	0.061	0.683	0.256	0.062	0.680	0.257	0.061	0.682	0.257

FV1 a **FV9**: ficket ventana 1 a 9

Las tres fases de lectura de una secuencia para ficket (fase 0: de FV1 a FV3, fase 1: FV1 a FV6 y fase 2: FV1 a FV9) para todas las secuencias de exones presenta un porcentaje promedio de 28,33% en las ventanas (FV1, FV4, FV7) mientras que para intrones el porcentaje promedio es de 6%, lo que indica que en las secuencias de exones hay mayor porcentaje codificación aplicando la prueba de ficket. Este comportamiento se espera que se mantenga a lo largo de todos los exones e intrones de los demás genomas estudiados.

4.7 Hexámeros

La disposición de los 4096 hexámeros en una matriz de 64x64 (Anexo L), manteniendo fijo en las filas el primer codón y variando el segundo codón por columnas, permitió seleccionar los codones de mayor y menor frecuencia tanto en filas como en las columnas, esto se aplicó para el cromosoma 19 de *M. musculus* logrando distinguir un comportamiento diferente en las tres primeras y últimas letras de los hexámeros tanto para secuencias codificantes como para secuencias no codificantes. Los siguientes cuadros muestran el primer codón de los hexámeros más frecuentes encontrados en los exones de los 8 organismos estudiados :

Cuadro 22: Frecuencia máxima del primer codón de los hexámeros mas frecuentes de los exones en los organismos estudiados.

S.cerevisiae		M.musculus		G.gallus		O.sativa		D.melano		C.elegans		A.thaliana		H.sapiens	
exon		Exon		exon		exon		exon		exon		exon		exon	
1150	TTC	207180	CTG	65966	CTG	134351	AAA	94015	CTG	157399	AAA	164884	AAA	259398	CTG
1111	TTT	192337	CAG	59951	CAG	134306	TTT	92193	CAG	154132	TTT	164489	TTT	241369	CAG
1095	AAA	148690	CTC	50041	TTT	120443	CTC	89063	TTT	137213	ATT	154243	CTT	218408	TTT
1029	CAA	148582	CTT	49868	AAA	118538	CTT	87927	AAA	134879	AAT	144696	GAA	216972	AAA
943	CTT	147047	CCA	48417	CTT	115958	GAG	83634	CCA	130763	GAA	143491	AAG	191644	CCA
926	AAT	145820	TGG	44288	GAA	112641	TTG	82829	TGG	128162	TTC	143482	TTC	187857	TGG
904	TTG	145261	TTT	43624	CTC	112414	GAA	81776	GCA	114222	CAA	141254	AGA	187680	CTT
887	ATC	145188	AAA	43316	TTC	112259	CAA	81368	TGC	111206	TTG	140070	TCT	186961	CTC
877	ATT	141825	GAG	42694	AAG	111488	ATG	81078	TTG	100413	ATC	129017	TTG	178465	GAG
857	GAA	138698	AGA	42035	GCT	111215	TTC	79975	CAA	100393	GAT	128011	CAA	176666	CCT

El Cuadro 22 ilustra el primer codón de los hexámeros mas frecuentes al recorrer horizontalmente la matriz 64x64 (Anexo L); se logra apreciar que no hay un patrón presente en todas las secuencias codificantes de los 8 organismos; sin embargo se puede ver que algunos organismos comparten el mismo codón inicial, ejemplo los codones CTG y CAG se presenta en M. musculus, G. gallus, D. melanogaster y H. sapiens; y el codón AAA y TTT se presenta en *O. sativa*, *C. elegans* y *A. thaliana*.

Cuadro 23: Frecuencia máxima del primer codón de los hexámeros mas frecuentes de los intrones en los organismos estudiados.

S.cerevisiae		A.thaliana		D.melanogaste		O.sativa		C.elegans		G.gallus		M.musculus		H.sapiens	
Intron		Intron		intron		intron		intron		intron		intron		intron	
470	TTT	160983	AAA	320113	AAA	291374	TTT	421528	TTT	541312	TTT	1574327	TTT	3573390	TTT
451	AAA	159595	TTT	319298	TTT	289737	AAA	409726	AAA	533661	AAA	1552422	AAA	3545151	AAA
290	AAT	103903	ATT	226947	ATT	216435	AAT	229362	ATT	381828	CTG	1052838	TCT	2706008	TCT
277	ATT	103843	AAT	225816	AAT	214701	ATT	224887	AAT	362557	CAG	1029695	AGA	2699180	AGA
263	ATA	89072	ATA	174480	TTA	189603	ATA	137038	TTA	334511	TGT	835632	TGT	2623834	TGT
257	TTA	88973	TTA	173819	TAA	188246	TAT	135693	TAA	330475	ACA	828233	ACA	2604935	ACA
248	TAA	88674	TAT	170756	ATA	170936	TAA	134921	TTC	330254	ATT	806867	CTG	2527914	CTG
237	GTA	88564	TAA	170105	TAT	170287	TTA	132535	GAA	330039	TCT	802966	CAG	2469000	CAG
219	TAT	77687	TTG	168890	TTG	169651	TTG	122099	TTG	329424	AGA	790064	CTT	2364473	CTT
209	GAA	77170	CAA	168601	CAA	168154	CAA	122084	CAA	329069	CTT	787908	AAG	2350521	AAG

El Cuadro 23 muestra que el primer codón de los hexámeros mas frecuentes en intrones es similar a los hexámeros mas frecuentes presentes en los exones; por lo tanto no hay un patrón bien definido tanto para exones como para intrones y es poco informativo para clasificar secuencias codificantes de las no codificantes. El análisis de los resultados de hexámeros señala que los porcentajes de hexámeros mas frecuentes y menos frecuentes son específicos por cada especie, y en general ofrece poca información de diferenciación para la predicción de exones e intrones en el genoma eucariote.

5. Modelo de predicción de exones

5.1 Creación del modelo

Para la creación del modelo de predicción se usó una muestra de 14917 secuencias aleatorias seleccionadas de una población de 1'602.490 secuencias, basados en el calculo muestral de la **etapa de modelamiento** del proceso de minería de datos. En la fase de selección y limpieza de atributos (Sección 3.1.3.5), se definió una matriz de correlaciones y de regresión múltiple para solucionar el problema de multicolinealidad, que permitieron seleccionar de manera rigurosa los atributos del modelo a partir de los cuales iniciar la etapa de modelamiento.

Se utilizaron tres técnicas de aprendizaje de máquina (TAM) para desarrollar los modelos: árboles de decisión (AD), Redes Bayesianas (BN) y Redes Neuronales (PM). Se crearon modelos agrupados en tres categorías: medidas estadísticas, medidas fractales y medidas mixtas. El método utilizado para generar los modelos por categoría se basó en la creación de modelos con todas las medidas para cada categoría y luego verificar el aporte de información de cada una de estas, y éste se evaluó con base en la comparación de resultados de las otras TAM.

La categoría uno consta de cinco modelos combinando las siguientes cuatro medidas estadísticas: ficket ventana 1 (**fv1**), porcentaje de G+C (**%gc**), índice CG (**ratio_cg**) y probabilidad de uso de codón (**pusage**) (Cuadro 24). Por ejemplo el modelo m1 consta de todas las medidas estadísticas alcanzando porcentajes de predicción del 85%, 81% y 81% para árboles de decisón, redes bayesianas y perceptrón multicapa respectivamente.

Cuadro 24: Categoría uno, modelos con medidas estadísticas

MP	Medidas Estadísticas estándar				Medidas fractales							TAM		
	fv1	%gc	ratio_cg	pusage	D	DMaxRN	DMinRN	Dtf	Hurst	Max	Min	AD	BN	PM
m1	X	X	X	X								85	81	81
m2	X	X	X									76	76	74
m3		X	X	X								83	80	83
m4		X		X								79	78	78
m5		X	X									76	75	73

MP: modelo de predicción, **TAM**: técnica de aprendizaje de máquina, **m1**: modelo 1, **m2**: modelo 2, **m3**: modelo 3, **AD**: árboles de decisión, **BN**: Bayes Network, **PM**: perceptrón multicapa, **fv1**: ficket ventana 1, **%gc**: porcentaje de GC, **ratio_cg**: índice de CG, **pusage**: probabilidad de uso de codón

Los modelos creados con medidas estadísticas presentaron en promedio porcentajes de predicción del 79.8% para los AD, 78% para BN, 77.8% para PM; lo que nos indica consistencia en la predicción, mostrando que la combinación de medidas estadísticas seleccionadas presentan óptima información para la clasificación (Cuadro 24). El modelo uno (m1) presentó el valor más alto de predicción al combinar las tres medidas estadísticas para AD y en general para las tres técnicas de aprendizaje de máquina.

Los modelos de la categoría dos, constituidos por medidas fractales presentaron un porcentaje promedio de predicción del 88% con AD, 86.3% con BN, 85% con PM (Cuadro 25), indicando consistencia en la predicción, tanto a niveles de las TAM como a través de los modelos; estos promedios son significativamente altos en comparación con los modelos de la categoría uno. Por ejemplo para el modelo m6 consta de todas las medidas fractales alcanzando porcentajes de predicción del 88%, 87% y 85% para árboles de decisón, redes bayesianas y perceptrón multicapa respectivamente.

Cuadro 25: Categoría dos, modelos con medidas fractales

MP	Medidas Estadísticas estándar				Medidas fractales							TAM		
	fv1	%gc	ratio_cg	pusage	D	DMaxRN	DMinRN	Dtf	Hurst	Max	Min	AD	BN	PM
m6					X	X	X	X	X	X	X	88	87	85
m7					X	X	X	X	X	X		88	86	85
m8					X	X	X	X		X		88	86	85

MP: modelo de predicción, **TAM**: técnica de aprendizaje de máquina, **m6**: modelo 6, **m7**: modelo 7, **m8**: modelo 8, **AD**: árboles de decisión, **BN**: Bayes Network, **PM**: perceptrón multicapa, **D**: dimensión fractal bidimensional, **DMaxRN**: bidimensional máxima, **DMinRN**: bidimensional mínima, **Dtf**: dimensión teórica fractal, **Hurst**: exponente de Hurst, **Max**: valor máximo de Hurst, **Min**: valor mínimo de Hurst.

La categoría tres formada por la combinación de las medidas estadísticas y fractales, muestran un promedio de 91.8% para AD, 89% para BN, 89.4% para PM, confirmando la consistencia de la predicción de estas medidas, en general las TAM presentan un incremento significativo en el nivel de predicción en comparación con las anteriores categorías (Cuadro 26). Por ejemplo para el modelo m9 consta de todas las medidas estadísticas y fractales alcanzando porcentajes de predicción del 95%, 89% y 91% para árboles de decisón, redes bayesianas y perceptrón multicapa respectivamente.

Cuadro 26: Categoría tres, modelos mixtos (medidas estadísticas y fractales)

MP	Medidas Estadísticas estándar				Medidas fractales							TAM		
	fv1	%gc	ratio_cg	pusage	D	DMaxRN	DMinRN	DTfd	Hurst	Max	Min	AD	BN	PM
m9	X	X	X	X	X	X	X	X	X	X	X	95	89	91
m10	X	X	X		X	X	X	X	X	X		92	89	89
m11		X			X	X	X	X	X	X		91	90	89
m12		X			X	X	X	X		X		91	89	91
m13		X	X		X	X	X		X	X		90	88	87

MP: modelo de predicción, **TAM**: técnica de aprendizaje de máquina, **m9**: modelo 9, **m10**: modelo 10, **m11**: modelo 11, **m12**: modelo 12, **m13**: modelo 13, **AD**: árboles de decisión, **BN**: Bayes Network, **PM**: perceptrón multicapa

Los resultados generales de los modelos agrupados por categorías señalan de manera evidente que los modelos que integran atributos estadísticos y fractales son buenos predictores de secuencias codificantes y no codificantes del ADN, la calidad de estos atributos se refleja en la poca variabilidad de los porcentajes de clasificación arrojados por las tres TAM.

5.2 Evaluación de los modelos

El conjunto de prueba esta conformado por 2400 secuencias, seleccionadas al azar en la **etapa de modelamiento** (Sección 3.1.4) del proceso de minería de datos. Para medir la exactitud de los resultados de los modelos de predicción se utilizaron las medidas comunes en la predicción de genes (Burset y Guigó, 1996): sensibilidad (Sn) y especificidad (Sp). La categoría uno formada por las medidas estadísticas señala que la combinación de las tres medidas (fv1, %gc, ratio cg), representadas en el modelo dos (m2) presentaron los mejores porcentajes de sensibilidad y especificidad del 84% y 69%, respectivamente (Cuadro 27).

Cuadro 27: Categoría uno, evaluación de los modelos estadísticos

Modelos	UI	Predictor		total	Sensibilidad	Especificidad	AD	BN	PM
		exon	intron						
Real	Exons	**vp**	**fn**	**vp+fn**	vp/(vp+fn)	vn/(fp+vn)			
	Introns	**fp**	**vn**	**fp+vn**					
m1	Exons	71	1232	1303	0.054	0.991	48	57	65
	Introns	10	1087	1097					
m2	Exons	1101	202	1083	0.845	0.694	76	74	74
	Introns	336	761	1097					
m3	Exons	348	955	1303	0.267	0.940	58	82	81
	Introns	66	1031	1097					
m4	Exons	48	1255	1303	0.037	0.992	47	78	78
	Introns	9	1088	1097					
m5	Exons	998	305	1303	0.766	0.711	74	74	74
	Introns	317	780	1097					

m1 a **m5**: modelos de predicción 1 a 5, **UI**: unidades de información, **vp**: verdaderos positivos, **fn**: falsos negativos, **fp**: falsos positivos, **vn**: verdaderos negativos, **AD**: árboles de decisión, **BN**: Bayes Net, **PM**: Perceptrón Multicapa

La evaluación de los modelos de la categoría dos muestran que la combinación de todas las medidas fractales generan modelos óptimos de predicción, sin embargo el modelo 6 (m6) presenta resultados significativamente altos de sensibilidad y especificidad del 77% y 89%, respectivamente (Cuadro 28); resaltando que el modelo 6 incluye la combinación de todas las medidas fractales. En general, los modelos fractales presentan valores altos de sensibilidad en comparación que los modelos estadísticos.

Cuadro 28: Categoría dos, evaluación de los modelos fractales

Modelos	UI	Predictor		total	Sensibilidad	Especificidad	AD	BN	PM
		exon	intron						
m6	Exons	1014	289	1303	0.778	0.896	83	84	78
	Introns	114	983	1097					
m7	Exons	990	313	1303	0.760	0.884	82	82	78
	Introns	127	970	1097					
m8	Exons	990	313	1303	0.760	0.882	82	82	78
	Introns	129	968	1097					

m6 a **m8**: modelos de predicción 6 a 8, **UI**: unidades de información, **AD**: árboles de decisión, **BN**: Bayes Net, **PM**: Perceptrón Multicapa

La evaluación de los modelos de la categoría tres muestra que la combinación de medidas estadísticas con medidas fractales genera modelos óptimos de predicción, sin embargo la combinación idónea se encontró en el modelo 13 (m13), al combinar porcentaje de GC, índice de CG, medidas bidimensionales (D, Dmax, Dmin) y medidas de Hurst (Hurst, max), que presenta una sensibilidad y especificidad de 90.6% y 84.3% (Cuadro 29).

Cuadro 29: Categoría tres, evaluación de los modelos mixtos

Modelos	UI	Predictor		total	Sensibilidad	Especificidad	AD	BN	PM
		exon	intron						
m9	Exons	749	554	1303	0.575	0.954	75	79	84
	Introns	51	1046	1097					
m10	Exons	1121	182	1303	0.860	0.870	86	87	86
	Introns	143	954	1097					
m11	Exons	1104	199	1303	0.847	0.884	86	86	84
	Introns	127	970	1097					
m12	Exons	1113	190	1303	0.854	0.879	87	87	84
	Introns	133	964	1097					
m13	Exons	1180	123	1303	0.906	0.843	88	87	87
	Introns	172	925	1097					

m9 a **m13:** modelos de predicción 9 a 13, **AD**: árboles de decisión, **BN**: Bayes Net, **PM**: Perceptrón Multicapa

Estos resultados muestran que la combinación de medidas estadísticas estándar con medidas fractales, es la mejor opción para clasificar exones que junto con la técnica de AD se logra un 90% de sensibilidad.

El porcentaje de instancias correctamente clasificadas de cada una de las TAM (AD, BN, PM) durante el proceso de entrenamiento y prueba son consistentes, así el porcentaje de clasificación en entrenamiento del modelo trece (m13) fue del 90% con AD, 88% con BN y 87% con PM, mientras que en prueba el mismo modelo, obtuvo un 88% con AD, 87% con BN y 87% con PM; los cambios en porcentaje de clasificación no fueron significativos lo que muestra la estabilidad del modelo para clasificar un conjunto de secuencias diferente al de entrenamiento.

La técnica C4.5 conocida en weka como J4.8, fue la utilizada para la construcción de los modelos de AD dado que contiene mejoras como la poda (prunning) y el criterio de ganancia (gain information) para igualar la competición de las variables (atributos).

Los resultados de los porcentajes de predicción para los AD fueron altos y similares: el conjunto de entrenamiento presento un 89.8371% de aciertos (Cuadro 30) y el conjunto de prueba alcanzo un 87.7083% (Cuadro 31). El alto porcentaje de acierto y similitud en los resultados de los conjuntos de datos indican a un alto grado de confiabilidad en los resultados.

Cuadro 30: Evaluación del conjunto de entrenamiento para AD (Resultados Weka)

=== Evaluation on training set ===
Correctly Classified Instances 13401 89.8371 %
Incorrectly Classified Instances 1516 10.1629 %
Kappa statistic 0.7958
Total Cost 5080
Average Cost 0.3406
Mean absolute error 0.1624
Root mean squared error 0.2849
Relative absolute error 32.5656 %
Root relative squared error 57.0663 %
Total Number of Instances 14917

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.92	0.126	0.89	0.92	0.905	0.945	exon
0.874	0.08	0.908	0.874	0.891	0.945	intron

=== Confusion Matrix ===

a	b	<-- classified as
7236	625	a = exon
891	6165	b = intron

Cuadro 31: Validación del conjunto de prueba para AD (Resultados Weka)

=== Evaluation on test set ===
Correctly Classified Instances 2105 87.7083 %
Incorrectly Classified Instances 295 12.2917 %
Kappa statistic 0.7515
Total Cost 983
Average Cost 0.4096
Mean absolute error 0.1794
Root mean squared error 0.3129
Relative absolute error 36.0448 %
Root relative squared error 62.7775 %
Total Number of Instances 2400

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.906	0.157	0.873	0.906	0.889	0.921	exon
0.843	0.094	0.883	0.843	0.862	0.921	intron

=== Confusion Matrix ===

a	b	<-- classified as
1180	123	a = exon
172	925	b = intron

Los estimadores de los errores estadísticos fueron bajos tanto en la fase de entrenamiento como en la fase de prueba; así, el error absoluto relativo fue de 32.5656% y 36.0448% en los conjuntos de entrenamiento y prueba respectivamente (Cuadro 30 y 31), mostrando poca variabilidad del error al pasar de una fase a otra, este comportamiento fue similar para la raíz cuadrada del error relativo donde se presento un 57.0663% y 62.7775% para el conjunto de entrenamiento y prueba respectivamente (Cuadro 30 y 31), confirmando que no hay cambios significativos en los porcentajes de error al usar el conjunto de entrenamiento y de prueba, mostrando que los atributos (medidas estadísticas y fractales) seleccionados no generan ruido al pasar de una fase a otra.

Con la técnica de BN se obtuvo resultados altos y similares, el conjunto de entrenamiento presento un 87.6115% de aciertos (Cuadro 32) y el conjunto de prueba alcanzo un 87.0417 % (Cuadro 33). El alto porcentaje de acierto y similitud en los resultados de los conjuntos de datos indican a un alto grado de confiabilidad en los resultados.

Los estimadores de los errores estadísticos fueron muy bajos tanto en la fase de entrenamiento como en la fase de prueba; así, el error absoluto relativo fue de 35.5173% y 36.7778% en los conjuntos de entrenamiento y prueba respectivamente (Cuadro 32 y 33), mostrando poca variabilidad del error al pasar de una fase a otra, este comportamiento fue similar para la raíz cuadrada del error relativo donde se presento un 60.4862% y 61.4413% para el conjunto de entrenamiento y prueba respectivamente (Cuadro 32 y 33), confirmando que no hay cambios significativos en los porcentajes de error al usar el conjunto de entrenamiento y de prueba, mostrando que los atributos (medidas estadísticas y fractales) seleccionados no generan ruido al pasar de una fase a otra.

Cuadro 32: Evaluación del conjunto de entrenamiento para BN (Resultados Weka)

=== Evaluation on training set ===
Correctly Classified Instances 13069 87.6115 %
Incorrectly Classified Instances 1848 12.3885 %
Kappa statistic 0.7518
Total Cost 5204
Average Cost 0.3489
Mean absolute error 0.1771
Root mean squared error 0.302
Relative absolute error 35.5173 %
Root relative squared error 60.4862 %
Total Number of Instances 14917

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.872	0.119	0.891	0.872	0.881	0.943	exon
0.881	0.128	0.86	0.881	0.871	0.943	intron

=== Confusion Matrix ===

a	b	<-- classified as
6852	1009	a = exon
839	6217	b = intron

Cuadro 33: Evaluación del conjunto de prueba para BN (Resultados Weka)

=== Evaluation on test set ===
Correctly Classified Instances 2089 87.0417 %
Incorrectly Classified Instances 311 12.9583 %
Kappa statistic 0.7397
Total Cost 851
Average Cost 0.3546
Mean absolute error 0.183
Root mean squared error 0.3062
Relative absolute error 36.7778 %
Root relative squared error 61.4413 %
Total Number of Instances 2400

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.865	0.123	0.893	0.865	0.879	0.94	exon
0.877	0.135	0.845	0.877	0.861	0.94	intron

=== Confusion Matrix ===

a	b	<-- classified as
1127	176	a = exon
135	962	b = intron

Cuadro 34: Evaluación del conjunto de entrenamiento para PM (Resultados Weka)

```
=== Evaluation on training set ===
Correctly Classified Instances       12999               87.1422 %
Incorrectly Classified Instances      1918                12.8578 %
Kappa statistic                          0.7422
Total Cost                            5678
Average Cost                             0.3806
Mean absolute error                      0.1936
Root mean squared error                  0.3106
Relative absolute error                 38.8342 %
Root relative squared error             62.207  %
Total Number of Instances            14917

=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
 0.876    0.133      0.88      0.876     0.878      0.936    exon
 0.867    0.124      0.862     0.867     0.864      0.936    intron

=== Confusion Matrix ===
   a    b   <-- classified as
 6883  978 |    a = exon
  940 6116 |    b = intron
```

Cuadro 35: Validación del conjunto de prueba para MP (Resultados Weka)

```
=== Evaluation on test set ===
Correctly Classified Instances        2089               87.0417 %
Incorrectly Classified Instances       311                12.9583 %
Kappa statistic                          0.7388
Total Cost                             943
Average Cost                             0.3929
Mean absolute error                      0.1977
Root mean squared error                  0.3135
Relative absolute error                 39.7151 %
Root relative squared error             62.9067 %
Total Number of Instances            2400

=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
 0.883    0.144      0.879     0.883     0.881      0.932    exon
 0.856    0.117      0.86      0.856     0.858      0.932    intron

=== Confusion Matrix ===
   a    b   <-- classified as
 1150  153 |    a = exon
  158  939 |    b = intron
```

Con la técnica de PM se obtuvo resultados altos y similares, el conjunto de entrenamiento presento un 87.1422% de aciertos (Cuadro 34) y el conjunto de prueba alcanzo un 87.0417% (Cuadro 35).

Los estimadores de los errores estadísticos fueron muy bajos tanto en la fase de entrenamiento como en la fase de prueba y similares al comportamiento presentado con BN; así, el error absoluto relativo fue de 38.8342% y 39.7151% en los conjuntos de entrenamiento y prueba respectivamente (Cuadro 34 y 35), mostrando poca variabilidad del error al pasar de una fase a otra, este comportamiento fue similar para la raíz cuadrada del error relativo donde se presento un 62.207% y 62.9067% para el conjunto de entrenamiento y prueba respectivamente (Cuadro 34 y 35), confirmando que no hay cambios significativos en los porcentajes de error al usar el conjunto de entrenamiento y de prueba.

El Cuadro 36 presenta un resumen de los resultados de predicción y de calidad del modelo para las tres TAM; La técnica AD presenta valores de predicción (89.8371% y 87.7083%, entrenamiento y prueba, respectivamente) y calidad del modelo (0.92 y 0.874 de sensibilidad y especificidad para los datos de entrenamiento y 0.906 de sensibilidad y 0.843 de especificidad para los datos prueba) mas altos que las otras dos técnicas (BN y PM); sin embargo, la información anterior (Cuadro 31 a 35) muestran que BN y PM presentan errores estadísticos mas bajos que AD, lo que indica que los tres predictores son buenos en la predicción de exones.

Se desea obtener un modelo de predicción de exones que ofrezca alto porcentaje de predicción y que su estructura interna sea expresiva, de tal forma que permita aprender más acerca de los fenómenos biológicos que ayuden a clasificar secuencias codificantes de las no codificantes. Los AD, entre las tres técnicas utilizadas, brindan mayor expresividad, posibilidad de visualizar el modelo, comprender su estructura y comportamiento interno de manera sencilla.

Cuadro 36: Cuadro comparativo de resultados de las TAM

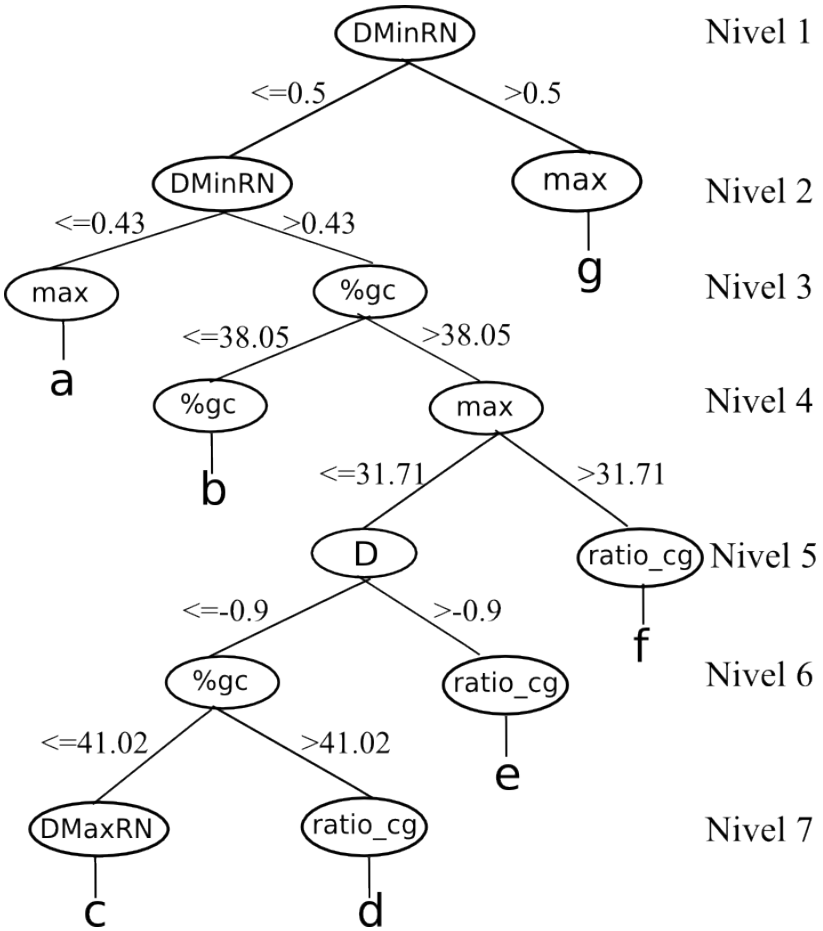
	Porcentaje de predicción		Entrenamiento		Prueba	
Algoritmo	C. entrenamiento	C. prueba	Sn	Sp	Sn	Sp
AD	89.8371 %	87.7083 %	0.92	0.874	0.906	0.843
BN	87.6115 %	87.0417 %	0.872	0.881	0.865	0.877
PM	87.1422 %	87.0417 %	0.876	0.867	0.883	0.856

AD: árboles de decisión, **BN**: Redes Bayesianas, **PM**: Perceptrón Multicapa, **C.entrenamiento**: conjunto de entrenamiento, **C.prueba**: conjunto de prueba, **Sn**: sensibilidad, **Sp**: especificidad.

El modelo de mejores resultados, modelo trece con AD (m13 con AD del cuadro 28), con base en la técnica de árboles de decisión, fue el modelo seleccionado para desplegar y visualizar. El árbol general se dividió en 7 ramas: a, b, c, d, e, f y g, con el fin de facilitar la

visualización, ya que el árbol tiene 85 hojas, 15 niveles y 169 nodos (entre hojas y nodos intermedios) (Anexo M):

La Gráfica 29 muestra el esquema general del modelo trece con AD (m13 con AD) ilustrando una organización alterna entre medidas estadísticas y fractales que se refuerzan mutuamente para lograr clasificar secuencias ADN codificantes y no codificantes. Cada nivel del árbol presenta medidas estadísticas acompañadas de medidas fractales por ejemplo en el nivel 3 se encuentra max (máximo de Hurst) con %gc, en el nivel 4 esta %gc con max (máximo de Hurst), en el nivel 5 esta D con ratio_cg, ésta estructura tan particular se presenta de forma similar en los demás niveles del árbol de decisión.



Gráfica 29: Esquema general del árbol del decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (**%gc**), índice de CG (**ratio_cg**); medidas fractales bidimensionales: bidimensional fractal (**D**), bidimensional mínimo (**DMinRN**), bidimensional máximo (**DMaxRN**), ; medidas fractales de Hurst: exponente de hurst (**hurst**), Hurst máximo (**max**)

Para cada una de las unidades de información se forman ciertos rangos de medidas que permiten clasificar mayor cantidad de secuencias en los primeros niveles que en los niveles más profundos, por ejemplo una secuencia con DMinRN menor de 0.5 y menor a 0.43; un max menor 44.33 y un DMinRN mayor a 0.2 se clasifica como exón (rama "**a**", Gráfica 30) mientras que si max es mayor a 44,33 y DMinRN es menor a 0.2 se clasifica como intrón (rama a). De igual manera se puede ver que una secuencia con DMinRn mayor a 0.5, max menor a 37.89 se clasifica como exón (rama "**g**", Gráfica 32).

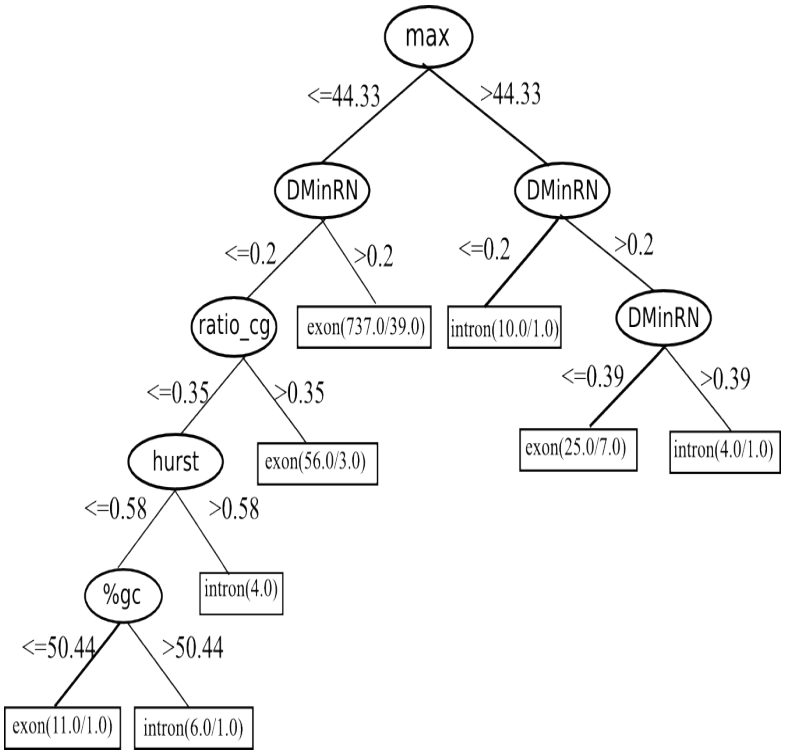
El árbol esta representado por óvalos (nodos) y rectángulos (clases), donde los primeros indican las medidas que permiten clasificar y los segundos representan la unidad de información y cantidad de datos o secuencias que cumplieron con las reglas (aristas) donde los números indican la cantidad de instancias que pertenecen a cierta clase ya sea exón o intrón, definiendo la clase de la hoja por la mayor cantidad de secuencias que pertenecen a exones o intrones.

Las ramas tienden a especializarse en la clasificación de cierto tipo de secuencias de acuerdo a la combinación de atributos; a continuación se muestra algunas ramas que tienen dicho comportamiento:

En la rama "**a**" (Gráfica 30) se observa una tendencia a clasificar mayor cantidad de exones que introns, 829 exones y 24 introns; definidos por la combinación de los valores de las medidas estadísticas y fractales específicas para secuencias codificantes.

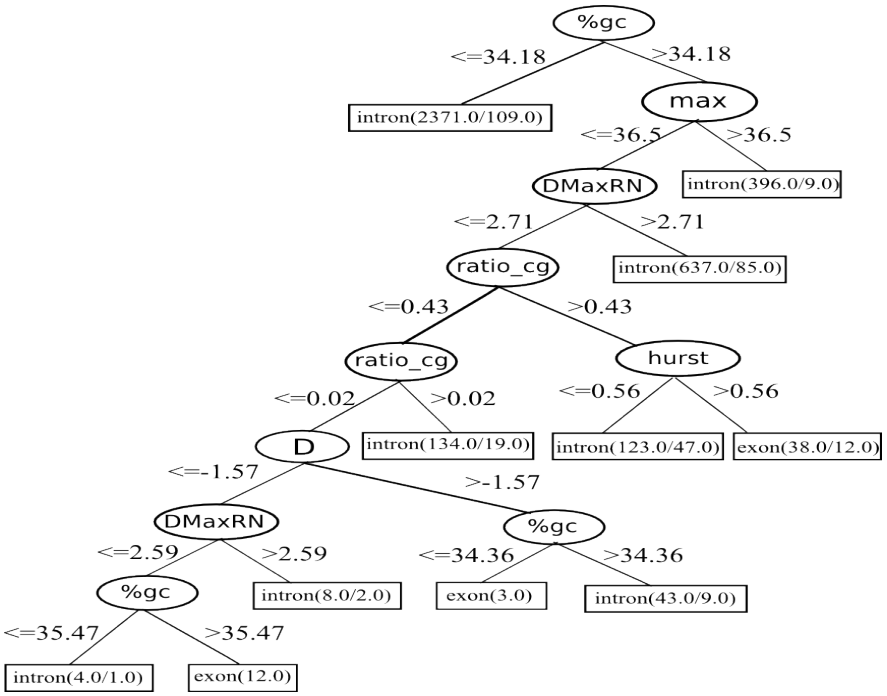
En la rama "**b**" (Gráfica 31) se presenta una tendencia a clasificar mayor cantidad de introns que exones, 3716 introns y 53 exones, definiendo valores para la combinación de las medidas estadísticas y fractales específicas para secuencias no codificantes.

Un caso particular se presenta en la rama "**g**" (Gráfica 32), ya que se sólo utiliza medidas fractales para la clasificación: máximo de Hurst (max) y bidimensional (D) que define una combinación de valores para clasificar exones.

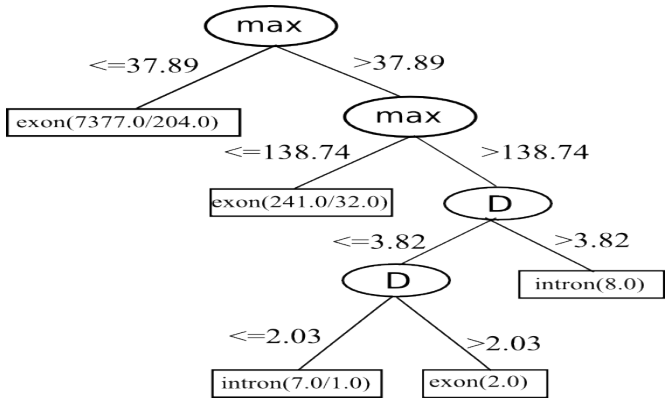


Gráfica 30: **Rama "a" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de GC (**%gc**), índice de CG (**ratio_cg**); medidas fractales bidimensionales: bidimensional fractal (**D**), bidimensional mínimo (**DMinRN**), bidimensional máximo (**DMaxRN**), ; medidas fractales de Hurst: exponente de Hurst (**hurst**), Hurst máximo (**max**)

En la rama **g** (Gráfica 32) se puede apreciar una posible reducción, donde el nodo **max** se encuentra en el primer y segundo nivel de la rama. El rango de valores del primer nodo **max** esta incluido en el rango del nodo del nivel dos. El error cometido en el primer nodo **max** es del 4,4% (204/(4377+204)), en el segundo es 11,7% (32/(241+32)); lo mas factible sería dejar el nodo con mínimo error y eliminar el segundo; pero esto incurre en la perdida de un 8,8 % de información, lo que implica una gran pérdida ya que se quiere ser estricto en la definición de los rangos para tener en cuenta el mayor número de valores posibles en los que se mueven las unidades de información, y como se analizó en cada una de las medidas, los rangos deben ser los mas estrechos posibles para lograr establecer una diferencia entre exones e intrones.



Gráfica 31: **Rama "b" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de GC (**%gc**), índice de CG (**ratio_cg**); medidas fractales bidimensionales: bidimensional fractal (**D**), bidimensional mínimo (**DMinRN**), bidimensional máximo (**DMaxRN**), medidas fractales de Hurst: exponente de Hurst (**hurst**), Hurst máximo (**max**)



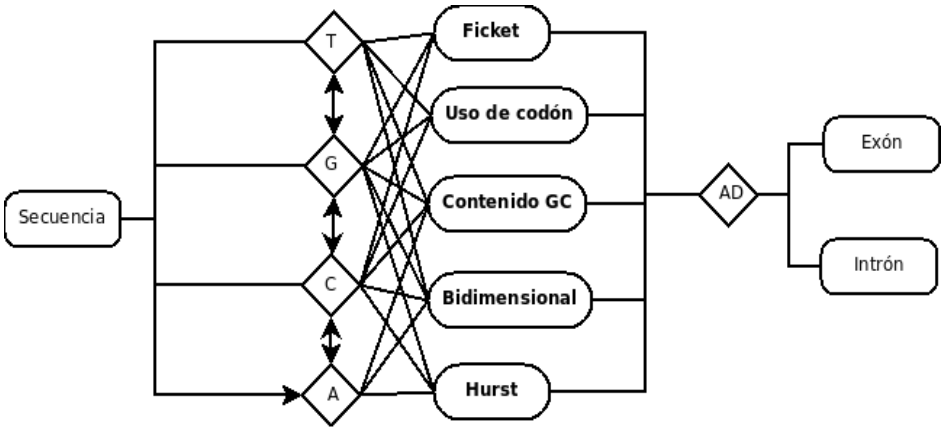
Gráfica 32: **Rama "g" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (**%gc**), índice de CG (**ratio_cg**); medidas fractales bidimensionales: bidimensional fractal (**D**), bidimensional mínimo (**DMinRN**), bidimensional máximo (**DMaxRN**), ; medidas fractales de Hurst: exponente de Hurst (hurst), Hurst máximo (max)

El modelo refleja combinaciones de valores de los atributos que permiten clasificar cantidades significativas tanto de exones como de intrones repartidas en diferentes ramas del árbol. Este modelo ayuda a resaltar mediante las medidas fractales y estadísticas, ciertas características presentes en exones e intrones que permiten cierto grado de diferencia tanto a nivel de estructura como de contenido.

Las medidas que ayudan a definir la estructura de las secuencias son porcentaje de CG e índice CG y entre las que definen el contenido están exponente de Hurst (max, hurst) y medida fractal bidimensional (DMaxRN, DMinRN, D). Las estructuras de las secuencias de ADN no son fijas y aun así las medidas utilizadas para la predicción alcanzan un porcentaje del 70% de efectividad. Igualmente el contenido de las secuencias de ADN es variable para cada especie, lo que hace difícil el proceso de búsqueda de patrones, por esta razón se alcanza el 88% efectividad en la identificación de secuencias codificantes.

El modelo definido, combina la información de estructura y contenido de las secuencias de ADN para alcanzar un 91% de efectividad en la predicción, esta unión se refleja en todo el modelo donde se presentan tanto medidas estadísticas como fractales, alcanzando un nivel alto en comparación con los modelos por separado.

5.3 Modelo de predicción de exones



Gráfica 33: Modelo de predicción de exones

El modelo de predicción de exones resultante toma una **secuencia** ADN anónima en formato Fasta para analizarla mediante medidas estadísticas (**Ficket**, **Uso de Codón** y **Contenido G+C**) y fractales (**Hurst** y **Bidimensional**), las cuales estudian el contenido y estructura de la secuencia, obteniendo valores que indican cierto comportamiento en la disposición de los nucleótidos presentes a lo largo de la secuencia, diferentes para cada medida con el fin de catalogarla como codificante o no codificante. Estos valores miden el comportamiento del tercer par base de un codón o el comportamiento de la frecuencia de codones o el comportamiento de la cantidad de Guanina y Citosina tanto de forma

separada (G+C) como los dos par base juntos (GC), además de tener en cuenta otro tipo de comportamiento como la dimensionalidad de la secuencia, que mediante un análisis más fino, exprese cambios notables al recorrer la secuencia en diferentes escalas muy pequeñas tratando de definir propiedades geométricas o al menos expresadas mediante una ecuación matemática. Todos los valores obtenidos anteriormente son los atributos de una secuencia de ADN que mediante una técnica supervisada de aprendizaje de maquina como los árboles de decisión (**AD**) clasifica dicha secuencia en las alguna de las dos categorías definidas: **exón** o **intrón**.

6. Conclusiones

6.1 Conclusiones generales del trabajo

El genoma eucariote es un sistema complejo en su estructura e información genética y se puso en evidencia esta premisa a través de la aplicación de los medidas estadísticas estándar junto con nuevas medidas matemáticas aplicadas al análisis de objetos fractales y usarlas para lograr un modelo de predicción que presenta importantes niveles de información y estructura que tienen los genomas.

El genoma eucariote varía de acuerdo a la especie, haciendo presente diferencias notables en la composición del genoma de una especie a otra; por ejemplo en el genoma humano se tienen exones cortos e intrones largos mientras que en *S. cerevisiae* se presentan exones largos con intrones cortos, debido a estas y otras diferencias no es fácil construir un predictor general para todos los genomas eucariotes.

Este estudio logra mostrar la capacidad de predicción de exones de manera generalizada en el genoma eucariote, utilizando un modelo que integra medidas estadísticas estándar y medidas fractales.

La capacidad de procesamiento de los equipos de computo aún es insuficiente en la búsqueda de patrones sobre la gran cantidad de información disponible en las Bases de datos genómicas, ésta insuficiencia se nota tanto en la capacidad de almacenamiento como en la capacidad de CPU para procesar largas cadenas de ADN, esto presenta inconvenientes en la elaboración de modelos ya que la costo computacional es elevada aún para las arquitecturas actuales.

El desarrollo de la genómica como área de estudio de los procesos relacionados con los genomas es reciente, aun más la disponibilidad de expertos de la informática y de las áreas de las ciencias biológicas. Los trabajos interdisciplinario son importantes ya que se comparten muchos puntos de vista de un mismo problema, abordado desde la experiencia de cada miembro del grupo, logrando con esto que a partir de los resultados obtenidos se puedan beneficiar varias áreas al mismo tiempo. Los problemas que se pueden resolver como la falta de alimentos o enfermedades en el hombre y animales domésticos, pueden al mismo tiempo desarrollar sofisticados algoritmos en el área de la Inteligencia Artificial y realimentarse de los resultados obtenidos.

6.2 Conclusiones del modelo

El modelo resultante para predicción de exones en genomas eucariotes se obtuvo de la combinación de medidas estadísticas estándar con medidas fractales e implementado con la técnica de Árboles de Decisión (AD) donde se obtiene un porcentaje promedio de predicción del 91.8% y resultó mejor que los otras técnicas de aprendizaje: Árboles de Decisión (AD) con 91.8%; Redes Bayesianas (BN) con 89% y Perceptrón Multicapa (PM) con 89.4%.

Lo anterior concluye que la mejor técnica de Aprendizaje de Máquina en la clasificación fue el Árbol de Decisión (AD) ya que sus resultados en los porcentajes de predicción supera las otras técnicas usadas como Redes Bayesianas (BN) y Redes Neuronales (Perceptrón Multicapa), además los árboles de decisión permiten apreciar los ajustes hechos en cada uno de los nodos, el comportamiento de las reglas a medida que los valores se van filtrando en su paso hacia las hojas o clases.

Los resultados con árboles de decisión presentan altos porcentajes de predicción y bajos porcentajes de error estadístico, mientras que las dos técnicas de aprendizaje de máquina BN y PM presentan un porcentaje relativamente menor de predicción y un poco más bajo en los errores estadísticos; el modelo m13 presenta con las tres técnicas resultados óptimos al combinar las medidas de predicción estadísticas y fractales. La similitud del comportamiento de las tres técnicas usadas en este modelo confirman que los conjuntos seleccionados de entrenamiento y prueba brindan información relevante para diferenciar secuencias codificantes de las no codificantes, gracias a la rigurosidad con la que se llevo a cabo la etapa de preparación de datos de la minería de datos.

Los altos porcentajes de predicción del modelo generado con AD, revelan que las medidas seleccionadas se complementan en la información que ofrecen, puesto que las medidas fractales hacen un análisis profundo del contenido y estructura de las secuencias de ADN, tal es el caso de bidimensional y Hurst donde recorren nucleótido por nucleótido en busca de regularidades y autosimilitudes estáticas o en el tiempo; mientras que las medidas estadísticas estándar como el contenido de Guanina y Citosina buscan elementos bien definidos y conservados en las estructuras del ADN y así definir comportamientos de diferenciación entre secuencias codificantes y no codificantes, todo éste análisis es orientado por una metodología que se enfoca en la búsqueda de información no disponible a primera vista.

La Minería de Datos es una metodología robusta para analizar conjunto de datos grandes como las secuencias de genomas eucariotes, permitiendo agilizar procesos en la búsqueda de patrones aplicando estadística y técnicas de Aprendizaje de Máquina.

El proceso riguroso de limpieza de los datos y de selección del conjunto muestral (conjunto de entrenamiento y prueba) permite que el modelo sea robusto y congruente en los resultados. Así los resultados fueron similares en las tres técnicas de aprendizaje de

máquina a través de todos los modelos previos al modelo trece (el modelo final) tanto en la etapa de entrenamiento como de evaluación.

La metodología de desarrollo de software, programación extrema (XP), fue adecuada para este tipo de investigación donde el desarrollo de prototipos, en este caso implementaciones de las medidas estadísticas o fractales, fue la principal forma de saber si el procedimiento a seguir era adecuado para alcanzar buenos resultados de predicción, es decir, si lograba clasificar secuencias codificantes y no codificantes, y así no desperdiciar esfuerzos planeando actividades con rigurosidad para finalmente llegar a resultados poco satisfactorios. Los buenos resultados en los prototipos permitieron seguir profundizando y así de forma sistemática llegar a los objetivos planteados.

6.3 Recomendaciones

Dada la complejidad de los genomas eucariotes y la abundancia de información biológica en las BD públicas, se requiere profundizar en la minería de datos, con el fin de lograr un mayor detalle de la información relacionada entre muchas de las características, fenómenos y funciones biológicas que los rodean. Se debe limitar el estudio de genomas eucariotes a especies, por ejemplo: estudiar los mamíferos, peces o plantas ya que estos genomas sí guardan una relación ya que están en la misma escala evolutiva.

Con el conocimiento biológico y computacional es posible obtener un modelo de predicción específico para cada genoma eucariote y relacionarlo con su escala evolutivamente cercana, en lugar de obtener un modelo que explique el comportamiento de todo un conjunto de organismos heterogéneos como si fueran organismos homogéneos en características y funciones.

La costo computacional presentada en el análisis bidimensional se puede abordar con un cambio de arquitectura de procesamiento donde se pueden explorar alternativas como procesamiento en cluster de CPU's o con un abordaje más novedoso que es el aprovechamiento de los procesadores gráficos (GPU's), usadas por las tarjetas gráficas como NVIDIA, para computo científico.

Fortalecer el área investigativa y formación de jóvenes profesionales en el campo científico, propiciando espacios donde se pueda socializar y compartir conocimientos de diferentes áreas como en este caso de biología e informática, los cuales pueden propiciar avances en el beneficio de la vida humana, animal y vegetal, contribuyendo de esta manera al progreso de la región y la Universidad.

Desarrollar líneas de fundamentación teórica que permitan un diálogo de saberes, resaltando la aplicabilidad y nuevos desarrollos tanto en el área biológica como informática, además de complementar las actitudes y aptitudes como seres humanos y sociables.

Continuar dando apoyo a grupos interdisciplinarios como GTI y BIMAC y líneas de investigación donde hay muchas cosas por explorar y contribuir al desarrollo del conocimiento a la par con lo desarrollado por las potencias mundiales; aprovechando todo el potencial en recursos naturales y humanos que tiene Colombia en beneficio propio y de la humanidad.

Bibliografía

Adams M.D. et al. (**2000**). (Celera Genomics). The genome sequence of Drosophila melanogaster. Science 287: 2185–2195.

Almeida, J. (**2002**). Ferramentas para comparaçao genômica. Tese de Doutorado. Instituto de Computaçao Universidade Estadual de Campinas.

Arabidopsis Genome Initiative. (**2000**). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.

Aranda, V. T . (**1998**) . El universo fractal. Número de Manual Formativo: 10 Tema: Matemáticas.

Argote, J.I. (**2004**). Dimensión Fractal. URL: http://www.campured.net. Consultado en Mayo de 2008.

Baldi, P. y Brunak, S. (**2001**) . Bioinformatics: The machine learning approach. MIT Press, Cambridge MA.

Beck, Kent. (**1999**) . Extreme Programming Explained: Embrace Change. Addison-Wesley.

Bedoya, O. (**2006**). Aplicando algoritmos de clasificación para la construcción de modelos de exones. Universidad del Valle.

Bennett, S. H. (**2000**). Origin of Fractal Branching Complexity in the Lung. Technical Report.University of California.

Biotech Magazine. (**2007**) . Bases de datos Bionformáticas. Biotech 2, Enero - Febrero 2007.

Breslau, L. et al. (**1999**). Web Caching and Zipf-Like Distributions:Evidence and Implications. Proc. Int'l Joint Conf. IEEE Com-puter and Comm. Societies (IEEE Infocom99), IEEE Comput-er Soc. Press, Los Alamitos, Calif. pp. 126-134.

Burge, C. y Karlin, S. (**1996**). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.

Burgos, J. D. y Moreno, T.P. (**1996**). Zipf scaling behavior in the immune system. BioSystem. 39:227-232.

Burgos, J., Breton, C.E., Moreno, P., Isaacs, R., Rodriguez, J. y Verdugo, A. (**1996**). Geometría fractal en biología molecular e inmunología. Revisata cientifica uninca, Bogota, 2, 7-16 .

Burset, M. y Guigó, R. (**1996**). Evaluation of gene structure prediction programs. Genomics, 34, 353 - 367.

C. elegans Sequencing Consortium. (**1998**). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012–2018.

Canavos, G.C. (**1984**) . Probabilidad y estadística: aplicaiones y métodos. McGraw Hill.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T. (**2000**) . CRISP DM 1,0, step by step dataming guide. SPSS, Inc.

Claverie, J. M. (**1997**). Computational methods for the identification of genes in vertebrate genomic sequences. Human Molecular Genetics, Vol. 6, No. 10:1735-1744.

Claverie, J.-M. y Bougueleret, L. (**1986**). Heuristic informational analysis of sequences. Nucleic Acids Res., 14, 179-196.

Cuello, L., Carretero, T., Conejero, R. y Toro, A. (**2005**). Obtener Información en Bases de datos de Biología Molucular. Biblioteca Nacional de Ciencias de la salud. Instituto de Salud Carlos III.

Deutsch M y Long M. (**1999**). Intron–exon structures of eukaryotic model organisms. Nucleic Acids Research, 1999, Vol. 27, No. 15 3219–3228.

Feder, Jens. (**1988**). Fractals. New York and London: Plenum Press, 1988. xv and 283 pp.

Fickett, J.W. y Tung, C.S. (**1992**). Assessment of protein coding measures. Nucleic Acids Res., 20. 6441-6450.

Gabaix, Xavier. (**1999**). Zipf's Law for Cities: An Explanation. The Quarterly Journal of Economics, Vol. 114, No. 3, pp. 739-767.

Gao, J., Cao, Y., Qi, Y. y Hu, J. (**2005**). Building Innovative Representations of DNA Sequences to Facilitate Gene Finding. IEEE Intelligent Systems. 1541-1672.

Gao, J., Qi, Y., Cao, Y. y Tung, W. (**2004**). Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences. Journal of Biomedicine and Biotechnology 2005:2 139-146.

Gates, M. A. (**1986**). A simple way to look at DNA. J. theor. Biol. 119, 316.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, y Oliver SG. (**1996**). Life with 6000 genes. Science 274:546, 563–567.

González, R. (**1996**). Física para juristas, economistas ... y demás gente curiosa. Barcelona: Crítica.

Guigó, Roderic. (**1998**). DNA Composition, Codon Usage and Exon Prediction. Departament d'Estadística, Universitat de Barcelona. España.

Hamori, E. y Ruskin, J. (**1983**). H Curves, Anovell Method of representation of nucleotides series especially suited for long DNA sequences.. J. Mol. Biol. Chem. 258, 1318-1327.

Hao B. L., Lee H. C., and Zhang S. Y. (**2000**). Fractals related to long DNA sequences and complete genomes. Chaos, Solitons and Fractals, 11(6), 825-836.

Havlin, S. y Ben-Avraham, D. (**1982**). Fractal dimensionality of polymer chains. Phys. A: Math. Gen. 15, L311.

Henderson, J., Salzberg, S. y Fasman, K. H. (**1998**). Finding Genes in DNA with a Hidden Markov Model. .

Hurst, H.E. (**1951**). Long-term storage of reservoirs. Trans. Amer. Soc. Civ. Eng. 116, 770-808.

International Chicken Genome Sequencing Consortium. (**2004**). International Chicken Genome Sequencing Consortium. Nature 432, 695-716.

Ivanov, P. Ch., Nunes, Luís A., Goldberger, A. L., Havlin , S., Rosenblum , M. G., Struzik, Z. R. y Stanley, H. E.. (**1999**). Multifractality in Human Heartbeat Dynamics. Nature 399. 461-465.

Ivars, P. (**1992**). El Turista Matemático. Publicat, Madrid Alianza. España.

Jeffrey, H. J. (**1990**). Chaos game representation of gene structure. Nucleic Acids Research18(8), 2163-2170.

Katsumi, V.. (**2002**). Bioinformática de Projetos Genoma de Bactérias. Dissertaçao de Mestrado. Instituto de Computaçao Universidade Estadual de Campinas..

Koski, T. (**2001**). Probabilistic and Statistical Modelling in Bioinformaics. University of Linkoping. Department of Mathematics, Sweden.

Kotsiantis, S. B. (**2007**). Supervised machine learning: A review of classification techniques. University of Peloponnese, Greece.

Majoros, W., Pertea, M. y Salzberg, S.L. (**2005**). Efficient decoding algorithms for generalized hidden Markov model gene finders. BMC Bioinformatics. 5:616.

Majoros, W., Pertea, M. y Salzberg, S.L. (**2004**). TIGRscan and GlimmerHMM: two open source abinitio eukaryotic gene finders. Bioinformatics 20, 2878-2879.

Mandelbrot, B.. (**1993**) . Los objetos fractales. Forma, azar y dimensión. Tusquets Editores, S.A.

Mandelbrot, B. y Hudson, R.L. (**2004**). The (mis) Behavoir of Markets. A fractal view of Risk, Ruin and Reward. Tusquets editrores. 322 p.

Mandelbrot, B.B. (**1982**). The fractal geometry of Nature. W. H. Freeman. New York.

Maniatis, T., Fritsch, E.F. y Sambrook, J. (**1982**). Molecular cloning: a laboratory manual. Cold spring Harbour Laboratory, NY.

Martinez, B. Ciro. (**2003**) . Estadística básica aplicada. Colección: Textos universitarios. Ecoe ediciones.

Mathé, C., Sagot, M., Schiex, T. y Rouzé, P. (**2002**). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. 30, 4103-4117.

Medrano-Soto, Arturo, Moreno, G., Vinuesa, P. (**2004**). Successful laterla transfer requires codon usage compability between foreign genes y recipient genomes. Mol Biol Evol, 21(10):1884-94.

Mitchell, T. M. (**1997**) . Machine Learning. Mc Graw Hill. 432 p.

Moreno, P.A., Velez, P.E. y Tischer, I. (**2006**) . Bioinformática para Biologos, Químicos, Ingenieros y profesionales de la Ciencias de la Salud. Memorias II Seminario Internacional en Genómica, Bioinformática y Biología de Sistemas. p 156-167. Universidad del Cauca.

NCBI. (**2008**). Genome database. URL: http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome (Consultado: 20 Mayo de 2008).

Nilsson, Nils J. (**1997**) . Introduction to Machine Learnin. Department of Computer Science . Stanford University. 209 p.

Nussinov, R. (**1984**). Strong doublet preferences in nucloetide sequences and DNA geometry. J. Mol. Evol., 20, 111.

Orallo, J.H., Ramírez, M.J., y Ramírez, C.F. (**2005**) . Introducción a la Minería de datos. Editorial Pearson, 2004. ISBN: 84 205 4091 9.

Peng C. K., Buldyrev S., Goldberg A. L., Havlin S., Sciortino F., Simons M., and Stanley H. E. (**1992**). Long-range correlations in nucleotide sequences. Nature 356, 168-170.

Pérez, J. A. (**2000**). Música fractal: el sonido del caos. Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante.

Pita F.S. y Pértegas D.S. (**2003**). Pruebas diagnósticas: Sensibilidad y especificidad. Cad Aten Primaria 2003; 10: 120-124.

Pyle, Dorian. (**1999**) . Data Preparation for Data Mining. Morgan Kaufmann, San Francisco, CA.

Salzberg, S., Delcher, A., Fasman, K., y Henderson, J. (**1999**). A decision tree system for finding genes in DNA. J. Comp. Biol. 5(4): 667-680.

Saupe, D. y Heinz-Otto, P. (**1988**). The Science of Fractal Images. Springer-Verlag.

Sole, R. V. y Manrubia, S. (**2001**). Orden y caos en sistemas complejos. Fundamentos. Barcelona: Edicions UPC.

Stanley, H. E., S. V. Buldyrev, A. L. Goldberger, J. M. Hausdorff, S. Havlin, J. Mietus, C. K. Peng, F. Sciortino, y M. Simons. (**1992**). Fractal Landscapes in Biological Systems: Long-Range Correlations in DNA and Interbeat Heart Intervals. Hamburg, Germany; Physica A 191, 1-12.

Suárez, O.R. (**2004**). Métodos de Predicción para Series Tiempo Fractales. Tesis Licenciatura. Matemáticas y Economía. Departamento de Física y Matemáticas, Escuela de Ciencias, Universidad de las Américas Puebla.

Tan, A.C., y Gilbert, D. (**2001**). Machine Learning and its application to bioinformatics: An overview. Bioinformatics Reseach Centre. University of Glasgow.

Uberbacher, E. C., Xu, Y. y Mural, R. J. (**1996**). Discovering and understanding genesin human DNA sequence using GRAIL. Meth. Enzymol., 266:259-281.

Uberbacher, E.C. y Mural, R.J. (**1991**). Locating protein-coding region in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl. Acad. Sci. USA, 88: 11261-11265.

Vélez, P., Moreno P.A., Gutierrez, J.M., Naik, A.K, Burgos, J., Garreta, L.E., Martinez, E., Amador, S., Flechas, A. y Tisher, I. (**2004**). Análisis Multifractal del Genoma Humano para la Búsqueda de Regularidades con Significado Biológico y una Contribución a la Generación de Biotecnología de la Información. Universidad del Cauca. COLCIENCIAS Código No. 1103-12-16765..

Venter JC et al. (**2001**). (Celera Genomics). The sequence of the human genome. Science 291: 1304–1351.

Wallis, J. W. et al. (**2004**). A physical map of the chicken genome. Nature 432, 761-764.

Wang Z, Chen Y. y Li, Y. (**2004**). A brief review of computational gene prediction methods. Geno. Prot. Bioinfo. 2(4):216-221.

Waterston RH et al. (**2002**). (Mouse Genome Sequencing Consortium). Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.

Witten, I.H y Frank, E. (**2005**) . Data mining: Practical machine learning tools and techniques . Elsevier Inc.

Xiao, Y., Runsheng Chen, Ruqun Shen, Jian Sun y Jun Xu. (**1995**). Fractal Dimension of Exon and Intron Sequences. J. theor. Biol. 175, 23-26.

Xu, Y., Einstein, J.R., Mural, R.J., Shah, M., y Uberbacher, E.C. (**1994**). An improved system for exon recognition and gene modeling in human DNA sequences. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (ed. Altman, R., Brutlag, D., Karp, P., Lathrop, R., and Searls, D.), pp. 376-384. AAAI Press, Menlo Park, CA.

Yu J, et. al. (**2002**). A draft sequence of the rice genome (Oryza sativa L. ssp. Indica). Science 296: 79–92.

Yu, Z.G. y Chen, G.Y. (**1999**). Rescaled range and transition matrix analysis of DNA sequences. Physics, 1.

Zhang, M. (**2002**). Computational Prediction of Eukaryotic Protein-Coding Genes. Nature Genetics. 3: 698-709.

Zipf, G. K. (**1949**). Human Behavior and the Principle of least Effort. Addison-Wesley Press.

Zipf, G.K. (**1932**). Selected Studies of the Principle of Relative Frequency in Language. Cambridge (Mass.).

Anexos

Anexo A: Parámetros de clasificación de exones con base en la Ley de ZM

#Min_r	Max_r	Min_length	Max_length	Min_P	Max_P
3,8028	3,8031	126	1852	-5,3549	-4,5986
3,8043	3,8046	126	1844	-5,3549	-4,6005
3,8127	3,8130	122	1805	-5,3689	-4,6098
3,8200	3,8203	119	1762	-5,3797	-4,6203
3,8239	3,8243	117	1741	-5,3871	-4,6255
3,8253	3,8257	116	1734	-5,3908	-4,6272
3,8269	3,8272	115	1728	-5,3946	-4,6287
3,8311	3,8314	113	1710	-5,4022	-4,6333
3,8364	3,8367	111	1678	-5,4100	-4,6415
3,8409	3,8412	108	1655	-5,4219	-4,6475
3,8417	3,8423	108	1650	-5,4219	-4,6488
3,8448	3,8451	106	1630	-5,4300	-4,6541
3,8478	3,8483	105	1611	-5,4341	-4,6592
3,8487	3,8492	105	1603	-5,4341	-4,6613
3,8522	3,8525	103	1580	-5,4425	-4,6676
3,8526	3,8529	103	1579	-5,4425	-4,6679
3,8536	3,8542	102	1571	-5,4467	-4,6701
3,8550	3,8553	101	1563	-5,4510	-4,6723
3,8561	3,8564	101	1558	-5,4510	-4,6737
3,8575	3,8579	99	1552	-5,4597	-4,6754
3,8592	3,8595	99	1543	-5,4597	-4,6779
3,8605	3,8613	98	1536	-5,4641	-4,6799
3,8620	3,8626	98	1526	-5,4641	-4,6827
3,8628	3,8631	97	1522	-5,4685	-4,6838
3,8648	3,8652	96	1511	-5,4730	-4,6870
3,8663	3,8668	96	1503	-5,4730	-4,6893
3,8674	3,8677	95	1498	-5,4776	-4,6908
3,8690	3,8695	94	1487	-5,4822	-4,6940
3,8696	3,8701	94	1482	-5,4822	-4,6954
3,8706	3,8710	93	1478	-5,4868	-4,6966
3,8715	3,8724	93	1473	-5,4868	-4,6981
3,8726	3,8729	93	1469	-5,4868	-4,6992
3,8730	3,8737	92	1468	-5,4915	-4,6995
3,8738	3,8741	92	1464	-5,4915	-4,7007
3,8744	3,8747	91	1461	-5,4963	-4,7016
3,8752	3,8760	91	1458	-5,4963	-4,7025
3,8768	3,8771	90	1450	-5,5011	-4,7049
3,8775	3,8781	90	1446	-5,5011	-4,7061
3,8784	3,8788	89	1442	-5,5059	-4,7073
3,8794	3,8797	89	1435	-5,5059	-4,7094
3,8798	3,8805	89	1434	-5,5059	-4,7097
3,8806	3,8811	88	1431	-5,5108	-4,7106
3,8812	3,8818	88	1428	-5,5108	-4,7115
3,8819	3,8833	87	1426	-5,5158	-4,7121
3,8841	3,8848	86	1418	-5,5208	-4,7146
3,8849	3,8853	85	1414	-5,5259	-4,7158
. . .					
4,4436	4,4451	88	88	-5,9218	-5,9218
4,4456	4,4473	87	87	-5,9267	-5,9267
4,4480	4,4490	86	86	-5,9318	-5,9318
4,4497	4,4510	85	85	-5,9368	-5,9368
4,4515	4,4536	84	84	-5,9420	-5,9420
4,4542	4,4555	83	83	-5,9472	-5,9472
4,4560	4,4575	82	82	-5,9524	-5,9524
4,4579	4,4599	81	82	-5,9578	-5,9524
4,4603	4,4615	80	81	-5,9632	-5,9578
4,4620	4,4633	79	79	-5,9686	-5,9686
4,4637	4,4654	78	78	-5,9742	-5,9742
4,4658	4,4671	77	77	-5,9798	-5,9798
4,4674	4,4686	76	76	-5,9854	-5,9854
4,4691	4,4705	75	75	-5,9912	-5,9912
4,4707	4,4717	74	75	-5,9970	-5,9912

Distribución de las UI del genoma de *M. musculus* por rangos. Las UI agrupadas en cada intervalo con su rangos máximos y mínimos, longitud máxima y mínima y frecuencia máxima y mínima

Anexo B: Matriz de correlaciones

	fv1	fv2	fv3	fv4	fv5	fv6	fv7	fv8	fv9	hcdo1	hcdo2	hcdv1	hcdv2	hcdof1	hcdof2	hcdvf1	hcdvf2	longitud	%gc	ratio_gc	ratio_cg	pusage	D	DMaxRN	DMinRN	Dtf	ZminL	ZmaxL	Hurst	Max	Min	Rango
fv1	1																															
fv2	0,82	1																														
fv3	0,86	0,98	1																													
fv4	0,73	0,71	0,78	1																												
fv5	0,67	0,78	0,81	0,91	1																											
fv6	0,69	0,77	0,83	0,95	0,98	1																										
fv7	0,61	0,58	0,66	0,83	0,74	0,8	1																									
fv8	0,57	0,62	0,67	0,78	0,79	0,82	0,94	1																								
fv9	0,59	0,61	0,68	0,8	0,78	0,83	0,97	0,99	1																							
hcdo1	0,87	0,97	0,99	0,79	0,83	0,84	0,66	0,68	0,69	1																						
hcdo2	0,87	0,97	0,99	0,8	0,83	0,84	0,67	0,68	0,69	0,99	1																					
hcdv1	0,84	0,99	0,98	0,72	0,77	0,77	0,58	0,62	0,61	0,97	0,97	1																				
hcdv2	0,84	0,99	0,99	0,73	0,78	0,78	0,6	0,62	0,62	0,97	0,97	0,99	1																			
hcdof1	0,85	0,88	0,9	0,75	0,77	0,77	0,58	0,59	0,59	0,89	0,9	0,87	0,88	1																		
hcdof2	0,86	0,88	0,9	0,74	0,77	0,77	0,61	0,63	0,63	0,9	0,9	0,87	0,87	0,92	1																	
hcdvf1	0,85	0,88	0,89	0,75	0,76	0,76	0,57	0,58	0,57	0,89	0,89	0,86	0,87	0,94	0,92	1																
hcdvf2	0,82	0,77	0,81	0,71	0,7	0,72	0,59	0,58	0,59	0,82	0,82	0,76	0,76	0,89	0,89	0,88	1															
longitud	0,86	1	0,99	0,74	0,79	0,79	0,61	0,64	0,64	0,98	0,98	0,99	0,99	0,9	0,89	0,89	0,89	1														
%gc	0,03	-0,03	-0,01	0,03	-0,02	-0,01	0,02	-0,01	0	0	0	-0,02	-0,02	0,01	0,02	0,01	0,09	-0,02	1													
ratio_gc	-0,05	0	-0,02	-0,05	-0,01	-0,02	-0,03	-0,01	-0,01	-0,03	-0,03	-0,01	-0,01	-0,04	-0,04	-0,04	-0,09	-0,01	-0,92	1												
ratio_cg	-0,05	-0,08	-0,08	-0,05	-0,07	-0,07	-0,03	-0,04	-0,04	-0,09	-0,09	-0,09	-0,09	-0,02	-0,03	-0,03	0,01	-0,08	0,04	0,01	1											
pusage	-0,14	-0,14	-0,14	-0,13	-0,12	-0,12	-0,08	-0,08	-0,07	-0,15	-0,15	-0,14	-0,14	-0,16	-0,17	-0,17	-0,18	-0,15	-0,11	0,18	0,08	1										
D	0,31	0,33	0,34	0,32	0,32	0,32	0,23	0,23	0,22	0,33	0,33	0,34	0,34	0,3	0,3	0,3	0,28	0,34	0,05	-0,05	-0,1	-0,17	1									
DMaxRN	0	0	0	0	0,01	0,01	0	0	0	0,01	0,01	0	0	0,01	0	0	0	0	-0,01	0,04	-0,01	-0,06	0,11	1								
DMinRN	-0,03	-0,07	-0,07	-0,03	-0,06	-0,06	-0,02	-0,04	-0,03	-0,07	-0,07	-0,07	-0,07	-0,06	-0,06	-0,06	-0,05	-0,07	0,16	-0,18	0,14	0,06	-0,05	0	1							
Dtf	-0,01	0	0	-0,01	0	0	-0,01	0	0	0	0	0	0	-0,01	-0,01	-0,01	-0,02	0	-0,05	0,09	-0,08	0,02	-0,05	0,06	-0,12	1						
ZminL	0,85	1	0,99	0,74	0,79	0,79	0,61	0,64	0,64	0,98	0,98	0,99	0,99	0,9	0,89	0,89	0,79	1	-0,02	-0,01	-0,08	-0,13	0,34	0	-0,07	0	1					
ZmaxL	0,85	1	0,99	0,74	0,79	0,79	0,61	0,64	0,64	0,98	0,98	0,99	0,99	0,89	0,89	0,89	0,79	1	-0,01	-0,01	-0,07	-0,13	0,32	0	-0,06	1	1					
Hurst	0,1	0,1	0,1	0,09	0,09	0,08	0,06	0,06	0,06	0,1	0,1	0,1	0,1	0,09	0,1	0,09	0,1	0,1	-0,01	-0,01	-0,02	-0,07	0,25	-0,02	0,02	0,1	0,09	1				
Max	0,63	0,74	0,75	0,66	0,71	0,7	0,55	0,59	0,58	0,73	0,73	0,73	0,73	0,68	0,67	0,68	0,61	0,74	-0,03	-0,01	-0,09	-0,18	0,52	0	-0,08	0	0,74	0,73	0,18	1		
Min	-0,61	-0,66	-0,65	-0,44	-0,47	-0,45	-0,31	-0,33	-0,32	-0,64	-0,62	-0,66	-0,65	-0,61	-0,63	-0,61	-0,54	-0,66	0,04	0	0,09	0,21	-0,48	0,01	0,08	0	-0,66	-0,65	-0,2	-0,42	1	
Rango	0,73	0,83	0,83	0,66	0,71	0,7	0,52	0,56	0,54	0,82	0,81	0,83	0,83	0,77	0,78	0,77	0,69	0,84	-0,04	0	-0,11	-0,23	0,6	0	-0,1	0	0,84	0,82	0,23	0,87	-0,81	1

Matriz de correlaciones de 32x32 atributos

Anexo C: conjunto de datos

ui	fv1	fv2	fv3	fv4	fv5	fv6	fv7	fv8	fv9	hexo1	hexo2	hexv1	hexv2	hexof1	hexof2	hexvf1	hexvf2	long	%gc	rgc	rcg	pusage	D	DmaxRN	DminRN	Dtfd	ZminL	ZmaxL	hurst	max	min	rango
intron	0	17	0	0	17	0	0	17	0	1	1	0	0	0	0	0	0	248	32,26	2,1	0	0,030	-0,600	2,75	0,5	1,49	126	1852	0,37	2,89	-46,08	48,97
intron	1	44	28	0	0	0	0	0	0	0	1	1	2	0	0	1	1	417	35,49	1,82	0,46	0,020	-2,380	2,75	0,5	1,52	126	1852	0,5	21,8	-24,86	46,65
intron	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	119	43,7	1,29	0	0,050	-1,300	2,75	0,5	1,42	119	1762	0,62	4,59	-21,69	26,28
intron	0	15	0	0	15	0	0	0	0	1	1	1	0	0	1	0	0	125	34,4	1,91	0,55	0,050	-2,500	2,7	0,5	1,46	122	1805	0,59	10,14	-16,45	26,59
intron	0	0	0	0	0	0	0	0	0	1	0	2	1	0	0	0	0	131	32,82	2,05	0,87	0,050	-1,790	2,75	0,5	2,22	126	1852	0,4	11,48	-4,55	16,04
intron	0	12	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	103	32,04	2,12	0,85	0,060	-0,590	2,63	0,5	1,33	103	1580	0,48	3,04	-24,48	27,52
intron	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	26,04	2,84	0	0,070	-1,240	2,75	0,5	1,73	96	1511	0,39	6,38	-11,54	17,92
intron	0	6	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	121	32,23	2,1	0,69	0,050	-1,620	2,75	0,5	1,77	119	1762	0,49	7,58	-15,6	23,19
intron	2	0	4	0	0	0	0	0	0	1	1	0	1	0	0	1	0	216	36,11	1,77	0,63	0,030	-0,920	2,75	0,5	1,2	126	1852	0,37	4,11	-19,71	23,82
intron	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	158	24,05	3,16	0	0,040	-1,830	3	0,5	1,83	126	1852	0,52	15,72	-9,82	25,54
...																																
exon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	54	38,89	1,57	1,5	0,120	-2,010	2,76	0,85	1,73	54	1248	0,33	2,4	-10,35	12,75
exon	0	2	26	0	2	26	0	2	26	0	0	0	2	0	0	0	0	227	39,21	1,55	0,23	0,030	-1,720	2,88	0,75	1,84	126	1852	0,41	19,03	-19,24	38,27
exon	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	63	57,14	0,75	0	0,100	-2,020	2,63	1	1,73	63	1281	0,72	11,55	-5,06	16,61
exon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108	38,89	1,57	0	0,060	-1,070	2,5	0,5	1,51	108	1655	0,71	25,16	-3,96	29,12
exon	0	2	9	0	0	0	0	0	0	3	0	0	0	0	0	0	0	60	45	1,22	0,33	0,110	-1,330	3,25	0,5	2,95	60	1268	0,59	8,52	-8,07	16,59
exon	6	0	5	6	0	5	6	0	5	1	2	1	2	0	0	0	1	230	50	1	0,42	0,030	-1,660	2,75	0,5	2,06	126	1852	0,41	14,76	-8,73	23,49
exon	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	118	48,31	1,07	0,44	0,050	1,650	3,17	1	1,59	117	1741	0,56	22,99	-1,2	24,19
exon	25	0	0	25	0	0	0	0	0	1	3	0	0	0	0	0	0	156	51,92	0,93	0,21	0,040	-2,410	2,5	0,5	1,55	126	1852	0,4	10,62	-13,95	24,57
exon	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	108	50	1	0,45	0,060	-2,240	2,46	1	2,03	108	1655	0,53	16,18	-3,29	19,47
exon	6	0	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	117	41,88	1,39	0,39	0,050	-1,890	2,6	0,56	2,17	117	1741	0,49	9,46	-15,39	24,85
exon	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	123	52,85	0,89	0,24	0,050	-1,680	2,75	0,5	1,58	122	1805	0,38	15,95	-2,66	18,61
...																																
exon	0	64	2	0	64	2	0	64	2	0	0	1	0	1	0	1	0	207	37,2	1,69	1,27	0,030	-1,790	2,64	0,5	2,74	126	1852	0,47	15,64	-5,49	21,13
exon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	63	36,51	1,74	0,5	0,100	-1,190	2,72	0,33	1,62	63	1281	0,65	3,33	-13,68	17,01
exon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	55,26	0,81	0,36	0,060	-1,710	3	0,6	1,71	113	1710	0,59	11,53	-10,48	22,01
intron	210	1068	409	210	1068	409	210	1068	409	20	19	17	12	1	4	5	3	5081	40,44	1,47	0,31	0,000	-2,190	2,46	0,5	1,56	5081	5081	0,6	79,5	-110,83	190,32
intron	0	129	31	0	129	31	0	0	0	2	3	1	3	0	0	0	0	679	34,46	1,9	0,15	0,010	-1,320	2,75	0,5	1,81	126	1852	0,4	55,31	-13,66	68,97
intron	0	14	8	0	0	0	0	0	0	2	0	0	0	1	0	0	0	264	45,83	1,18	0,29	0,020	-1,180	2,78	0,5	2,11	126	1852	0,31	14,31	-21,74	36,05
...																																
intron	1	0	31	1	0	31	0	0	0	2	1	1	0	0	0	0	0	206	61,17	0,63	0,12	0,030	-0,620	2,75	0,5	1,24	126	1852	0,5	16,83	-6,3	23,14
intron	0	76	3	0	76	3	0	76	3	1	1	0	1	1	0	0	0	434	31,34	2,19	0,19	0,010	-1,220	2,75	0,5	1,55	126	1852	0,57	20,49	-33,19	53,69
intron	28	89	65	28	89	65	0	0	0	4	5	3	5	0	0	1	0	745	46,44	1,15	0,18	0,010	4,310	2,75	0,5	1,45	126	1852	0,79	111,3	-60,77	172,08
intron	79	658	196	79	658	196	0	0	0	13	12	24	7	1	3	1	2	2998	48,33	1,07	0,33	0,000	16,730	2,75	0,5	1,25	2998	2998	0,85	291,25	-80,23	371,47
intron	24	698	226	24	698	226	24	698	226	8	15	5	9	5	1	4	4	3041	42,35	1,36	0,65	0,000	-0,990	2,58	0,5	1,81	3041	3041	0,52	81,24	-82,98	164,22

Conjunto de datos : 14917 filas (registros) con sus 32 columnas (atributos)

Anexo D: Historias de usuario

Historia de Usuario	
Número: 1	**Usuario:** Desarrollador
Nombre historia: Implementación de Uso de Codón	
Prioridad en negocio: Media	Riesgo en desarrollo: Bajo
Puntos estimados: 4	**Iteración asignada: 1**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Se identifica la frecuencia de codones en las secuencias codificantes y no codificantes de ADN.	
Observaciones: Este proceso se desarrolla en la **etapa de preparación de datos** de la metodología de Minería de Datos (Minería de Texto)	

Historia de Usuario	
Número: 2	**Usuario:** Desarrollador
Nombre historia: Implementación de la prueba de Ficket	
Prioridad en negocio: Media	Riesgo en desarrollo: Bajo
Puntos estimados: 4	**Iteración asignada: 1**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Se aplica la prueba de Ficket a las secuencias codificantes y no codificantes de ADN.	
Observaciones: Este proceso se desarrolla en la **etapa de preparación de datos** de la metodología de Minería de Datos (Minería de Texto)	

Historia de Usuario	
Número: 3	**Usuario:** Desarrollador
Nombre historia: Implementación del contenido G+C	
Prioridad en negocio: Media	**Riesgo en desarrollo:** Bajo
Puntos estimados: 4	**Iteración asignada: 1**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Se análisa el contenido de Guanina y Citosina en las secuencias codificantes y no codificantes de ADN	
Observaciones: Este proceso se desarrolla en la **etapa de preparación de datos** de la metodología de Minería de Datos (Minería de Texto)	

Historia de Usuario	
Número: 4	**Usuario:** Desarrollador
Nombre historia: Análisis del contenido de hexámeros en las secuencias de ADN	
Prioridad en negocio: Media	**Riesgo en desarrollo:** Bajo
Puntos estimados: 4	**Iteración asignada: 1**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Se análisa el contenido de hexámeros en las secuencias codificantes y no codificantes.	
Observaciones: Este proceso se desarrolla en la **etapa de preparación de datos** de la metodología de Minería de Datos (Minería de Texto)	

Historia de Usuario	
Número: 5	**Usuario:** Desarrollador
Nombre historia: Implementación del análisis reescalado R/S	
Prioridad en negocio: Alta	**Riesgo en desarrollo:** Alto
Puntos estimados: 5	**Iteración asignada: 2**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Desarrollar algoritmos para aplicar análisis R/S a las secuencias de ADN.	
Observaciones: Este proceso se desarrolla en la **etapa de preparación de datos** de la metodología de Minería de Datos (Minería de Texto)	

Historia de Usuario	
Número: 6	**Usuario:** Desarrollador
Nombre historia: Implementación del análisis de Zipf-Mandelbrot	
Prioridad en negocio: Alta	**Riesgo en desarrollo:** Alto
Puntos estimados: 5	**Iteración asignada: 2**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Desarrollar algoritmos para aplicar análisis Zipf-Mandelbrot a las secuencias de ADN.	
Observaciones: Este proceso se desarrolla dentro de la **etapa de preparación de datos** de la metodología de Minería de Datos (Minería de Texto)	

Historia de Usuario	
Número: 7	**Usuario:** Desarrollador
Nombre historia: Implementación del análisis bidimensional	
Prioridad en negocio: Alta	**Riesgo en desarrollo:** Alto
Puntos estimados: 5	**Iteración asignada: 2**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Desarrollar algoritmos para aplicar análisis bidimensional a las secuencias de ADN.	
Observaciones: Este proceso se desarrolla en la **etapa de preparación de datos** de la metodología de Minería de Datos (Minería de Texto)	

Historia de Usuario	
Número: 8	**Usuario:** Desarrollador
Nombre historia: Implementación del modelo de clasificación	
Prioridad en negocio: Alta	**Riesgo en desarrollo:** Alto
Puntos estimados: 4	**Iteración asignada: 3**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Implementación del modelo de clasificación: El modelo obtenido en la Minería de Texto se pasa a código Python.	
Observaciones:	

Historia de Usuario	
Número: 9	**Usuario:** Usuario
Nombre historia: Seleccionar archivos de las secuencias de ADN	
Prioridad en negocio: Baja	**Riesgo en desarrollo:** Bajo
Puntos estimados: 3	**Iteración asignada: 4**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: El usuario accede a la aplicación, donde selecciona la(s) secuencia(s) a clasificar, estas deben estar en formato Fasta. Inmediatamente aparece la información de las secuencias seleccionadas como nombre, longitud, cantidad individual de nucleótidos.	
Observaciones:	

Historia de Usuario	
Número: 10	**Usuario:** Usuario
Nombre historia: Clasificar secuencia de ADN	
Prioridad en negocio: Media	**Riesgo en desarrollo:** Medio
Puntos estimados: 3	**Iteración asignada: 4**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: El usuario al presionar la opción “Clasificar”, da inicio al proceso de clasificación de las secuencias Fasta. Este proceso consta de llamados a la Interfaz de Aplicaciones (API) del modelo de predicción implementado.	
Observaciones:	

Historia de Usuario	
Número: 11	**Usuario:** Usuario
Nombre historia: Presentar resultados de la clasificación	
Prioridad en negocio: Media	**Riesgo en desarrollo:** Bajo
Puntos estimados: 2	**Iteración asignada: 4**
Programador responsable: Carlos Téllez – Edwin Caldón	
Descripción: Una vez procesado las secuencias de ADN, se presenta un listado con el resultado de cada secuencia, en una ventana diferente.	
Observaciones:	

Plantillas de las Historias de usuario de la metodología de Programación Extrema

Anexo E: Rangos de dimensiones Fractales con base en la Ley de ZM

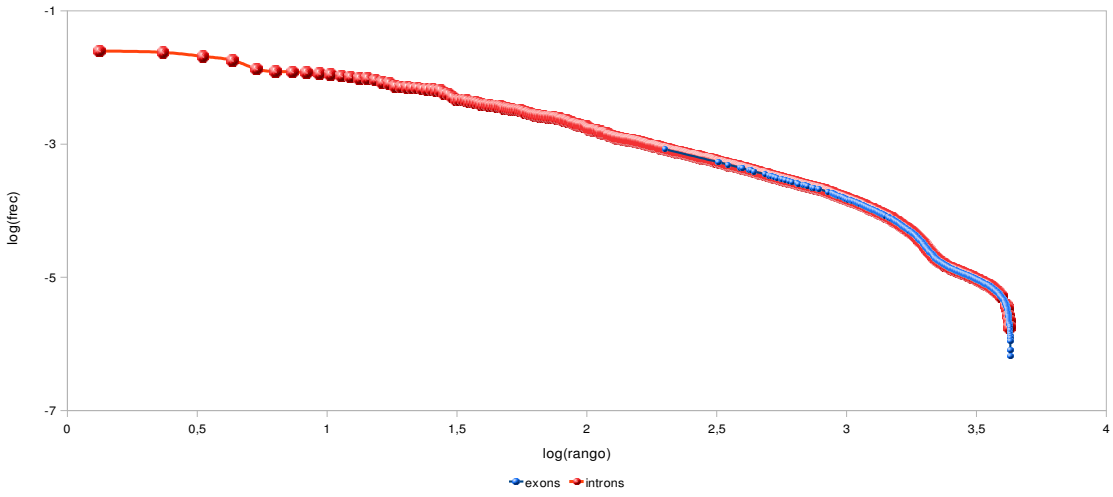
Exons/Gen			
Ranges		Qty genes	
0,95	1,00	21	
0,90	0,95	32	
0,85	0,90	30	
0,80	0,85	29	
0,75	0,80	24	
0,70	0,75	43	
0,65	0,70	26	
0,60	0,65	27	
0,55	0,60	21	0,60
0,50	0,55	15	
0,45	0,50	23	
0,40	0,45	20	
0,35	0,40	15	
0,30	0,35	13	
0,25	0,30	9	
0,20	0,25	7	
0,15	0,20	10	
0,10	0,15	6	

Introns/Gen			
Ranges		Qty genes	
1,15	1,20	2	
1,10	1,15	6	
1,05	1,10	6	
1,00	1,05	6	
0,95	1,00	11	
0,90	0,95	18	
0,85	0,90	20	
0,80	0,85	30	
0,75	0,80	36	
0,70	0,75	33	
0,65	0,70	44	
0,60	0,65	45	
0,55	0,60	54	0,91
0,50	0,55	46	
0,45	0,50	44	
0,40	0,45	36	
0,35	0,40	47	
0,30	0,35	22	
0,25	0,30	11	
0,20	0,25	11	
0,15	0,20	3	
0,10	0,15	1	

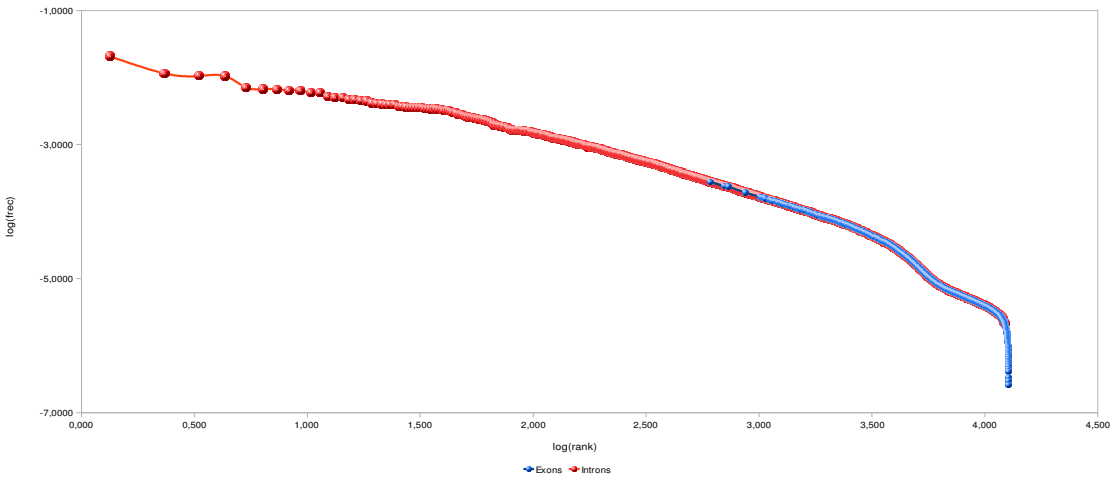
619 GI			
Ranges		Qty genes	
0,95	1,00	13	
0,90	0,95	10	
0,85	0,90	11	
0,80	0,85	23	
0,75	0,80	16	0,50
0,70	0,75	23	
0,65	0,70	45	
0,60	0,65	45	
0,55	0,60	64	
0,50	0,55	58	
0,45	0,50	69	
0,40	0,45	62	
0,35	0,40	77	
0,30	0,35	30	0,45
0,25	0,30	23	
0,20	0,25	11	
0,15	0,20	3	
0,05	0,10	1	

Relación de las dimensiones fractales por intervalos para exones e intrones del cromosoma 19 de *M. musculus*

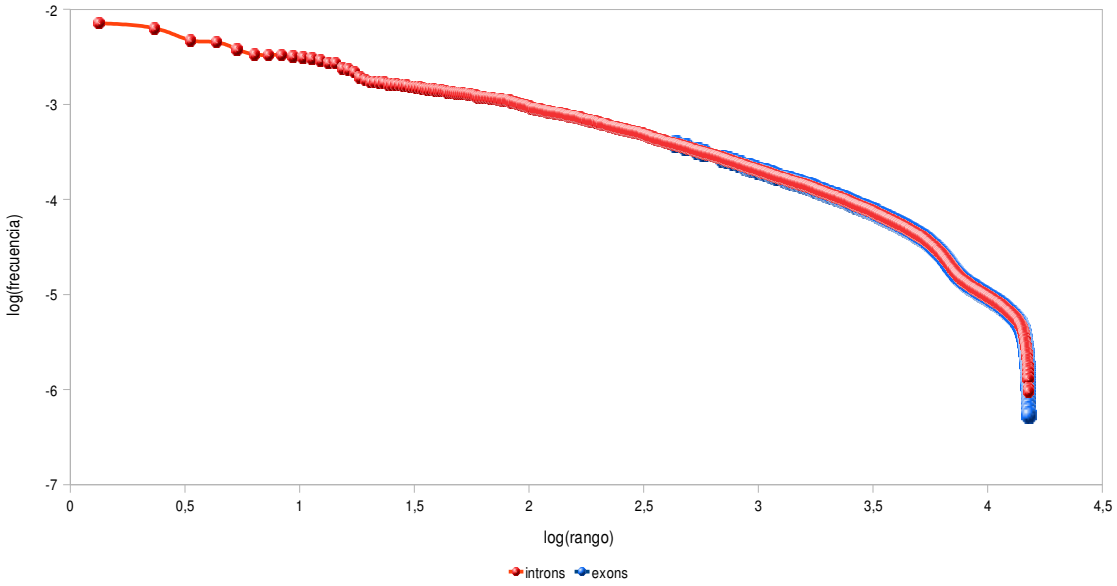
Anexo F: Distribución de exones e intrones en los 8 genomas



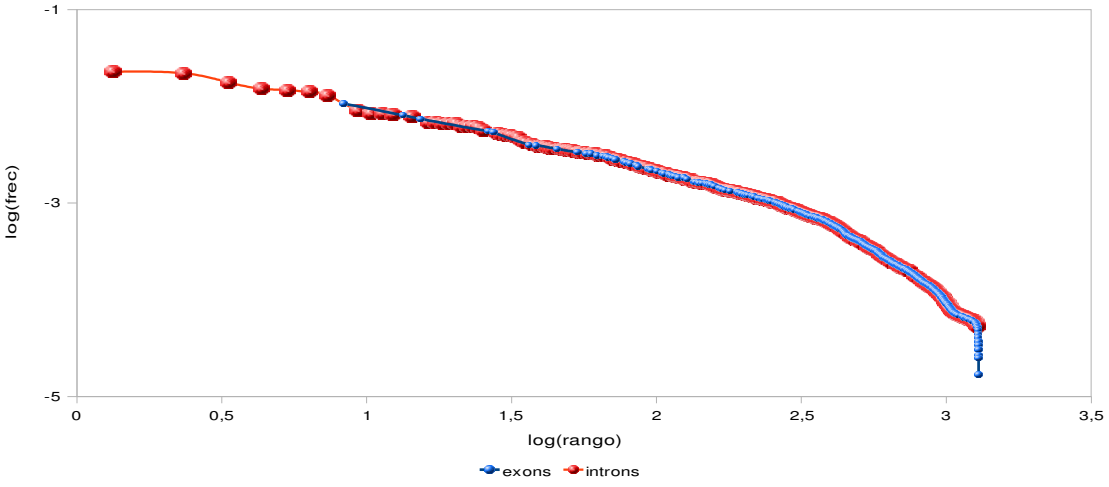
H. sapiens



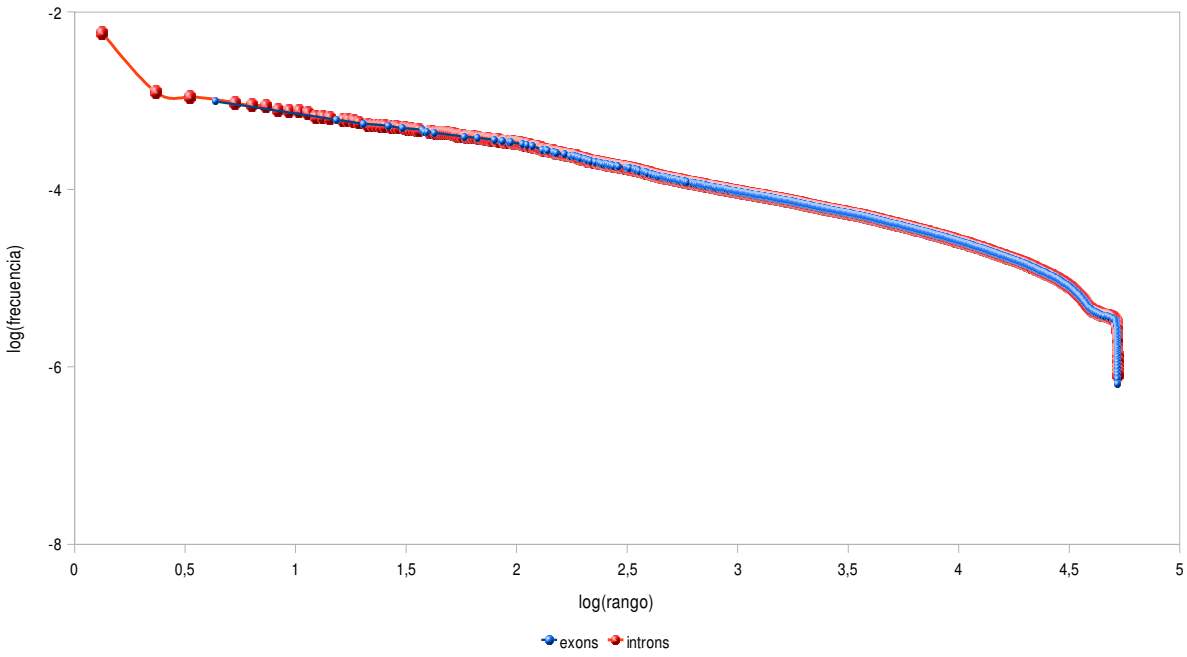
M. musculus



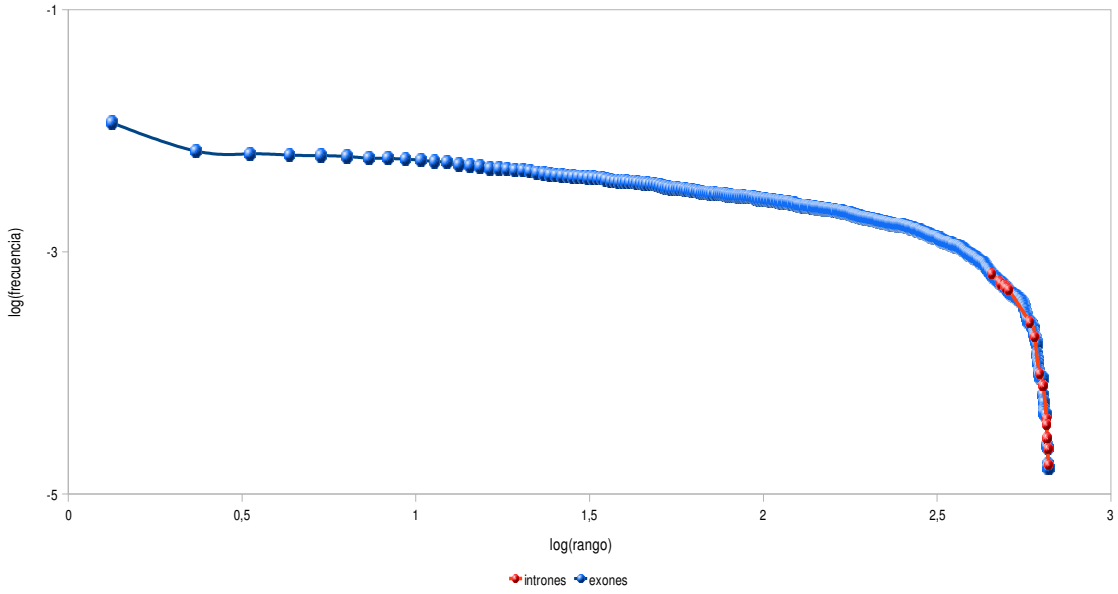
G.gallus



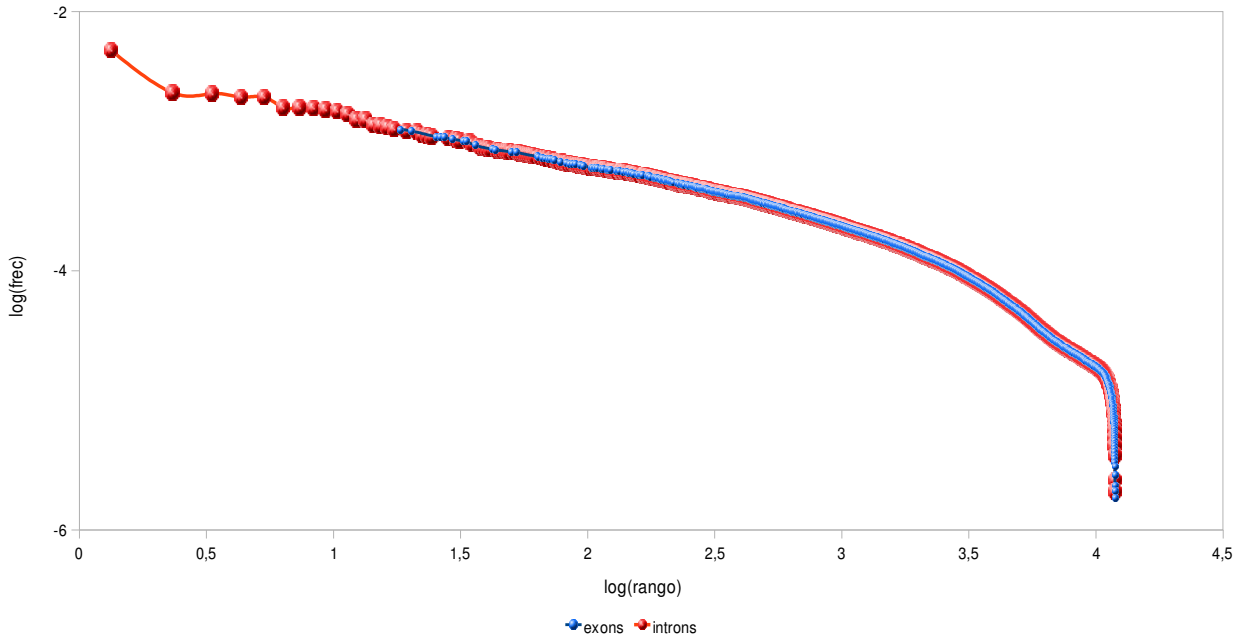
D. melanogaster



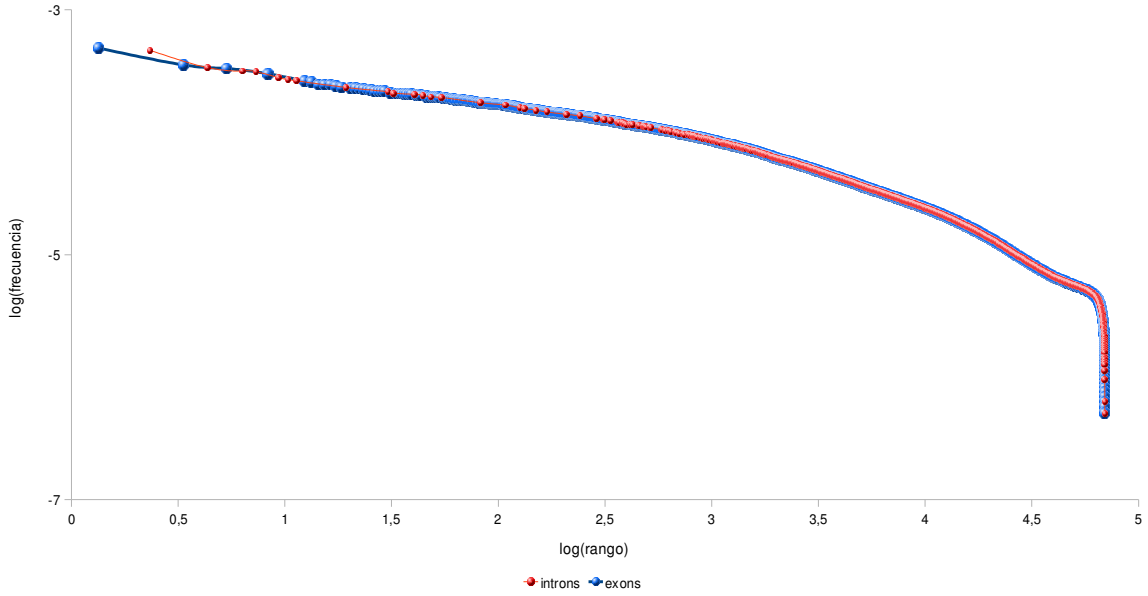
C. elegans



S. cerevisiae



O. sativa



A. thaliana

Distribución de exones e intrones a lo largo de los cromosomas de los organismos estudiados. En las especies superiores se ilustra que los exones están dispersos al final de la curva, mientras que los intrones están al inicio; este comportamiento cambia a medida que la especie se mueve en la escala evolutiva, donde los exones cambian al inicio de la curva y los intrones al final.

Anexo G: Sn y Sp para cromosomas y genomas con base en Ley de ZM

H.sapiens		Sn	Sp
	chr5	0,5875	0,8753
	chr6	0,6316	0,8128
	chr7	0,4956	0,8368
	chr8	0,2212	0,9651
	chr9	0,6242	0,8744
	chr17	0,7719	0,6364
	chr18	0,0000	1,0000
	chr21	0,0000	1,0000
	chr22	0,0000	1,0000
M.musculus	chr1	0,8804	0,6498
	chr2	0,9587	0,5264
	chr3	0,7163	0,8216
	chr4	0,8944	0,6347
	chr5	0,9385	0,6278
	chr6	0,7049	0,8058
	chr7	0,9167	0,5567
	chr8	0,7321	0,7674
	chr9	0,8604	0,6841
	chr13	0,6183	0,9436
	chr14	0,5621	0,9079
	chr15	0,6233	0,8537
	chr16	0,4970	0,9396
	chr17	0,6692	0,7619
	chr18	0,2035	0,9859
	chr19	0,5567	0,8767
G.gallus	chr1	0,6098	0,8913
	chr2	0,1485	0,9810
	chr3	0,1590	0,9768
	chr4	0,1489	0,9621
	chr5	0,1547	0,9651
	chr6	0,0000	1,0000

D.melanogaster		Sn	Sp
	chr1	0,1001	0,9198
	chr2	0,0468	0,9519
	chr3	0,1269	0,9170
	chr4	0,0000	1,0000
	chr5	0,0000	1,0000
	chr6	0,3323	0,7693
C.elegans	chr1	0,0005	0,9999
	chr2	0,0002	0,9999
	chr3	0,0002	1,0000
	chr4	0,0003	1,0000
	chr5	0,0008	0,9996
	chr6	0,0003	1,0000
S.cerevisiae	chr1	0,0000	1,0000
	chr2	0,0000	1,0000
	chr3	0,0000	1,0000
	chr4	0,0000	1,0000
	chr5	0,0000	1,0000
	chr6	0,0000	1,0000
	chr7	0,0000	1,0000
	chr8	0,0000	1,0000
	chr9	0,0000	1,0000
	chr10	0,0000	1,0000
	chr11	0,0000	1,0000
	chr12	0,0000	1,0000
	chr13	0,0000	1,0000
	chr14	0,0000	1,0000
	chr15	0,0000	1,0000
	chr16	0,0000	1,0000
O.sativa	chr1	0,0011	0,9994
	chr2	0,0013	0,9998
	chr3	0,0012	0,9996

	chr7	0,0000	1,0000

	chr4	0,0000	1,0000

G.gallus		Sn	Sp
	chr8	0,0000	1,0000
	chr9	0,0000	1,0000
	chr10	0,0000	1,0000
	chr11	0,0000	1,0000
	chr12	0,0000	1,0000
	chr13	0,0000	1,0000
	chr14	0,0000	1,0000
	chr15	0,0000	1,0000
	chr16	0,0000	1,0000
	chr17	0,0000	1,0000
	chr18	0,0000	1,0000
	chr19	0,0000	1,0000
	chr20	0,0000	1,0000
	chr21	0,0000	1,0000
	chr22	0,0000	1,0000
	chr23	0,0000	1,0000
	chr24	0,0000	1,0000
	chr25	0,0000	1,0000
	chr26	0,0000	1,0000
	chr27	0,0000	1,0000
	chr28	0,0000	1,0000

O.sativa		Sn	Sp
	chr5	0,0000	1,0000
	chr6	0,0000	1,0000
	chr7	0,0000	1,0000
	chr8	0,0000	1,0000
	chr9	0,0000	1,0000
	chr10	0,0000	1,0000
	chr11	0,0000	1,0000
	chr12	0,0000	1,0000
A.thaliana	chr1	0,0009	0,9999
	chr2	0,0000	1,0000
	chr3	0,0005	1,0000
	chr4	0,0004	1,0000
	chr5	0,0001	1,0000

Validez de la Ley de Zipf-Mandelbrot por cromosoma y genoma mediante las medidas de Sensibilidad (Sn) y Especificidad (Sp).

Anexo H: Resultados Uso de codón

homo	Exons	Introns
Chr21	+	-
Chr18	+	-
Chr22	+	-
Chr8	+	-
Chr9	+	+
Chr5	-	-
Chr6	+	+
Chr7	-	-
chr17	+	-
mus		
Chr7	+	-
Chr2	+	-
Chr1	+	-
Chr4	+	-
Chr5	+	+
Chr9	+	+
Chr6	-	-
Chr8	+	+
Chr17	-	-
Chr3	-	+
Chr13	-	-
Chr14	+	-
Chr15	+	-
Chr19	+	-
Chr16	+	-
Chr18	-	-
Gallus		
Chr16	+	+
Chr25	+	-
Chr22	-	+
Chr24	-	-
Chr27	+	+
Chr23	+	+
Chr28	-	-
Chr21	-	-
Chr26	+	+
Chr11	+	-
Chr17	+	-
Chr12	+	+
Chr13	+	+
Chr20	-	-
Chr19	-	+
Chr14	+	+
Chr15	+	-
Chr10	+	-
Chr9	+	+
Chr6	+	+
Chr7	+	-
Chr18	+	-

drosophyla	Exons	Introns
Chr5	+	-
Chr6	+	+
Chr1	-	-
Chr3	-	+
Chr2	-	-
Chr4	-	+
celegans		
Chr3	+	+
Chr1	+	-
Chr2	+	-
Chr6	+	-
Chr4	-	-
Chr5	+	-
scerevisiae		
Chr1	+	+
Chr6	-	+
Chr3	-	+
Chr9	-	+
Chr8	-	-
Chr5	-	+
Chr11	+	+
Chr10	-	+
Chr14	-	+
Chr2	-	+
Chr16	-	+
Chr13	-	+
Chr15	+	+
Chr12	-	+
Chr7	-	+
Chr4	-	+
Oriza		
Chr10	-	+
Chr9	+	+
Chr11	+	+
Chr12	+	+
Chr8	+	+
Chr7	-	+
Chr6	-	-
Chr5	+	+
Chr4	+	-
Chr2	-	-
Chr3	-	+
Chr1	+	-
arabidopsis		
Chr1	+	+
Chr2	+	+
Chr3	+	+
Chr4	+	+
Chr5	+	+

Resultados de uso de codón para cada cromosoma de los genomas estudiados

Anexo I: Sn y Sp para cromosomas y genomas con base en el contenido GC

H.sapiens	Sn	Sp
chr5	0,929	0,391
chr6	0,927	0,363
chr7	0,933	0,356
chr8	0,928	0,405
chr9	0,921	0,348
chr17	0,931	0,256
chr18	0,941	0,427
chr21	0,932	0,288
chr22	0,933	0,265
M.musculus		
chr1	0,941	0,340
chr2	0,932	0,311
chr3	0,932	0,313
chr4	0,931	0,296
chr5	0,937	0,288
chr6	0,928	0,343
chr7	0,923	0,299
chr8	0,934	0,289
chr9	0,930	0,303
chr13	0,934	0,337
chr14	0,936	0,319
chr15	0,935	0,292
chr16	0,938	0,348
chr17	0,930	0,262
chr18	0,939	0,339
chr19	0,937	0,272

D.melanogaster	Sn	Sp
chr1	0,741	0,359
chr2	0,752	0,320
chr3	0,748	0,349
chr4	0,762	0,333
chr5	0,676	0,638
chr6	0,728	0,325
C.elegans		
chr1	0,754	0,494
chr2	0,775	0,419
chr3	0,759	0,505
chr4	0,741	0,474
chr5	0,771	0,416
chr6	0,807	0,410
S.cerevisiae		
chr1	0,682	0,000
chr2	0,758	0,333
chr3	0,750	0,333
chr4	0,754	0,455
chr5	0,773	0,462
chr6	0,704	0,545
chr7	0,761	0,360
chr8	0,751	0,176
chr9	0,693	0,625
chr10	0,728	0,400
chr11	0,723	0,333
chr12	0,729	0,154
chr13	0,712	0,379
chr14	0,757	0,300
chr15	0,728	0,375

chr16	0,742	0,538

G.gallus		
chr1	0,942	0,219
chr2	0,947	0,230
chr3	0,947	0,216
chr4	0,941	0,187
chr5	0,940	0,209
chr6	0,949	0,215
chr7	0,941	0,192
chr8	0,944	0,200
chr9	0,943	0,171
chr10	0,943	0,185
chr11	0,947	0,192
chr12	0,946	0,196
chr13	0,938	0,202
chr14	0,938	0,195
chr15	0,938	0,199
chr16	0,864	0,199
chr17	0,943	0,192
chr18	0,935	0,211
chr19	0,941	0,181
chr20	0,944	0,188
chr21	0,941	0,170
chr22	0,925	0,163
chr23	0,937	0,163
chr24	0,930	0,157
chr25	0,941	0,108
chr26	0,957	0,174
chr27	0,948	0,166
chr28	0,939	0,160

O.sativa		
chr1	0,888	0,112
chr2	0,890	0,109
chr3	0,807	0,196
chr4	0,882	0,116
chr5	0,888	0,105
chr6	0,800	0,207
chr7	0,810	0,183
chr8	0,811	0,193
chr9	0,881	0,127
chr10	0,879	0,131
chr11	0,864	0,148
chr12	0,897	0,118
A.thaliana		
chr1	0,889	0,180
chr2	0,887	0,180
chr3	0,884	0,179
chr4	0,883	0,186
chr5	0,885	0,185

Validez del contenido de GC para cromosomas y genomas estudiados mediante Sensibilidad (Sn) y Especificidad (SP)

Anexo J: Sn y Sp para cromosomas y genomas con base en el análisis Bidimensional

G.gallus	Sn	Sp
chr1	0,984	0,109
chr6	0,989	0,095
chr7	0,987	0,082
chr8	0,989	0,072
chr9	0,991	0,082
chr10	0,990	0,083
chr11	0,987	0,078
chr12	0,990	0,101
chr13	0,985	0,070
chr14	0,995	0,066
chr15	0,986	0,062
chr16	0,942	0,050
chr17	0,993	0,051
chr18	0,991	0,082
chr19	0,991	0,061
chr20	0,992	0,080
chr21	0,991	0,055
chr22	0,985	0,052
chr25	0,982	0,033
chr26	0,990	0,041
chr27	0,990	0,035
chr28	0,990	0,038
S.cerevisiae	Sn	Sp
chr1	0,9545	0,0000
chr2	0,9456	0,0000
chr3	0,9712	0,0000
chr4	0,9360	0,0000
chr5	0,9426	0,0769
chr6	0,9691	0,0000

O.sativa	Sn	Sp
chr1	0,975	0,036
chr2	0,976	0,034
chr3	0,977	0,031
chr4	0,978	0,036
chr5	0,975	0,036
chr6	0,974	0,028
chr7	0,975	0,032
chr8	0,974	0,035
chr9	0,970	0,034
chr10	0,972	0,025
chr11	0,987	0,034
chr12	0,962	0,035
A.thaliana	Sn	Sp
chr1	0,9779	0,0115
chr2	0,9769	0,0117
chr3	0,9785	0,0106
chr4	0,9784	0,0109
chr5	0,9762	0,0104
C.elegans	Sn	Sp
chr1	0,9820	0,0092
chr2	0,9866	0,0074
chr3	0,9834	0,0075
chr4	0,9834	0,0066
chr5	0,9881	0,0053
chr6	0,9877	0,0058

chr7	0,9602	0,0400
chr8	0,9486	0,0000

S.cerevisiae	Sn	Sp
chr9	0,9432	0,0000
chr10	0,9569	0,0000
chr11	0,9243	0,0000
chr12	0,9349	0,0000
chr13	0,9455	0,0000
chr14	0,9473	0,0000
chr15	0,9443	0,0000
chr16	0,9431	0,0000

Validez del análisis bidimensional mediante Sensibilidad (Sn) y Especificidad (Sp) para cromosomas y genomas estudiados

Anexo K: Análisis R/S de los 8 genomas estudiados

Distribuciones de exones e intrones con base en el análisis R/S para los ocho genomas estudiados

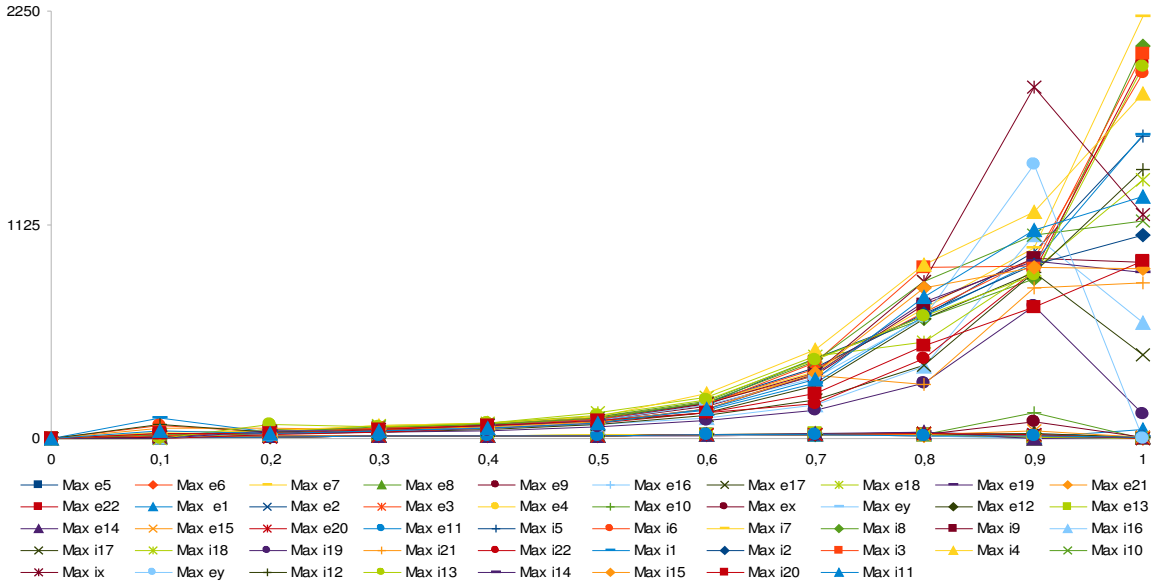


Imagen Máximos H.sapiens

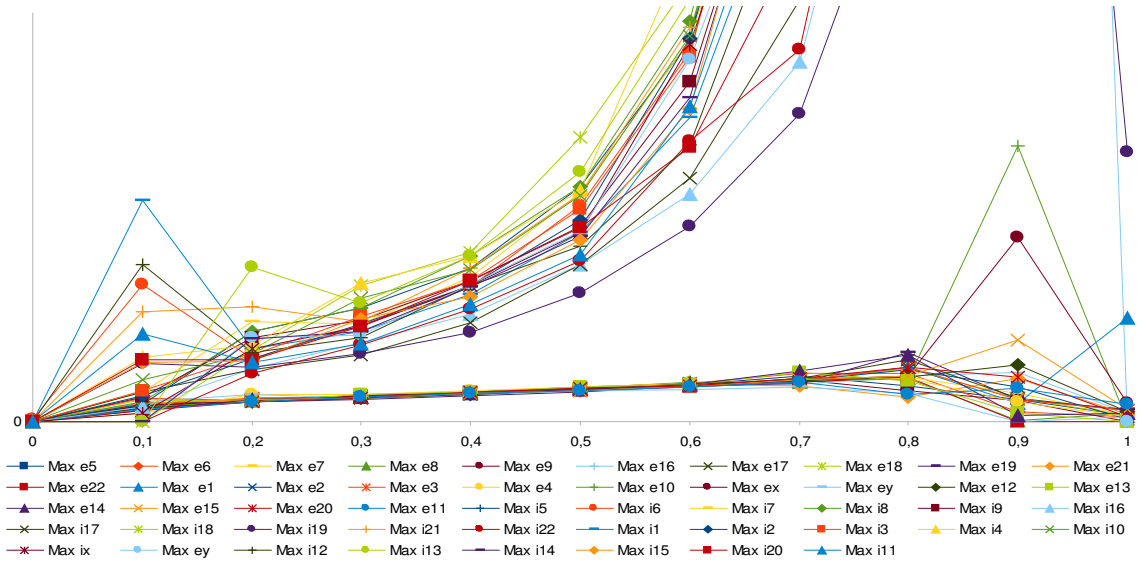
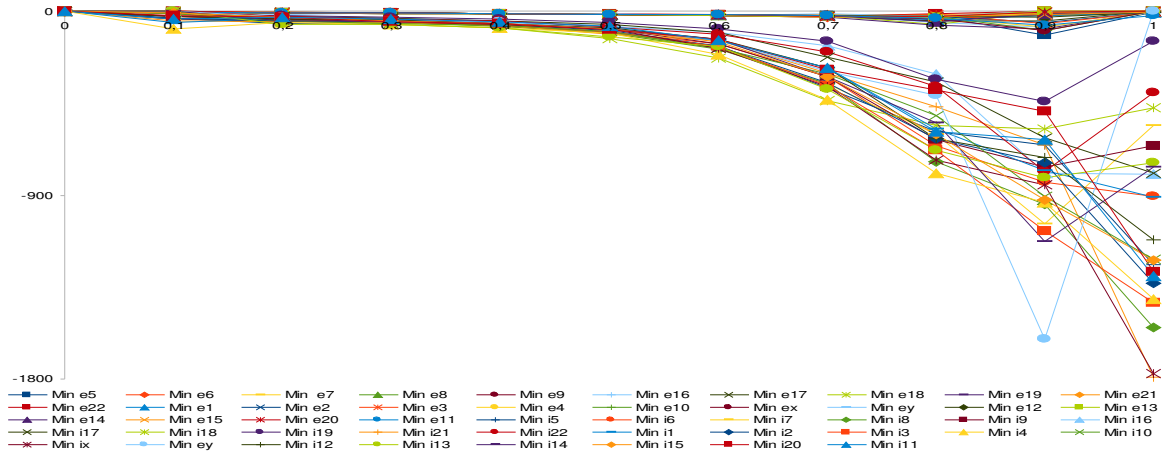
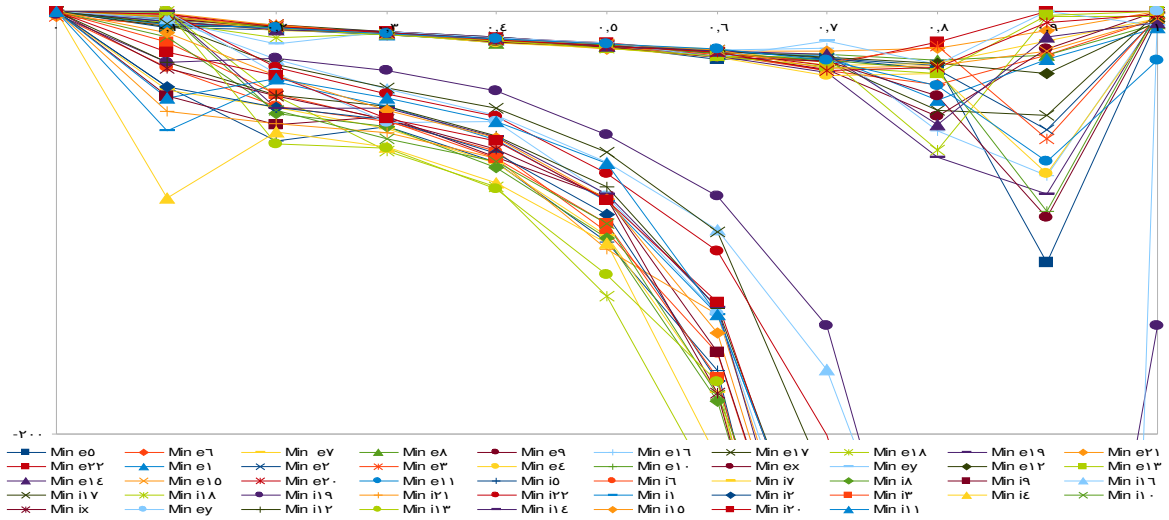


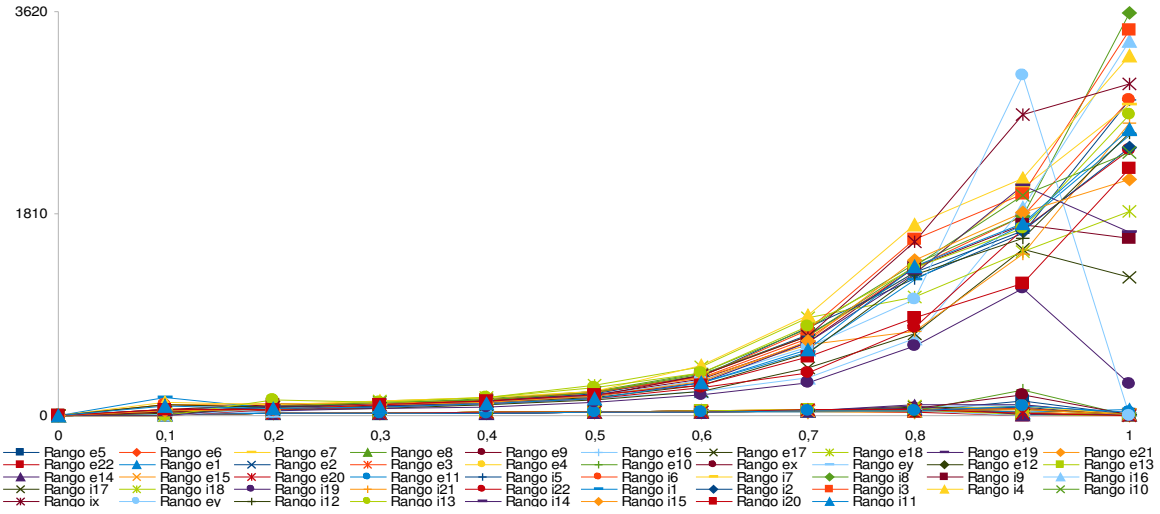
Imagen zoom máximos H.sapiens



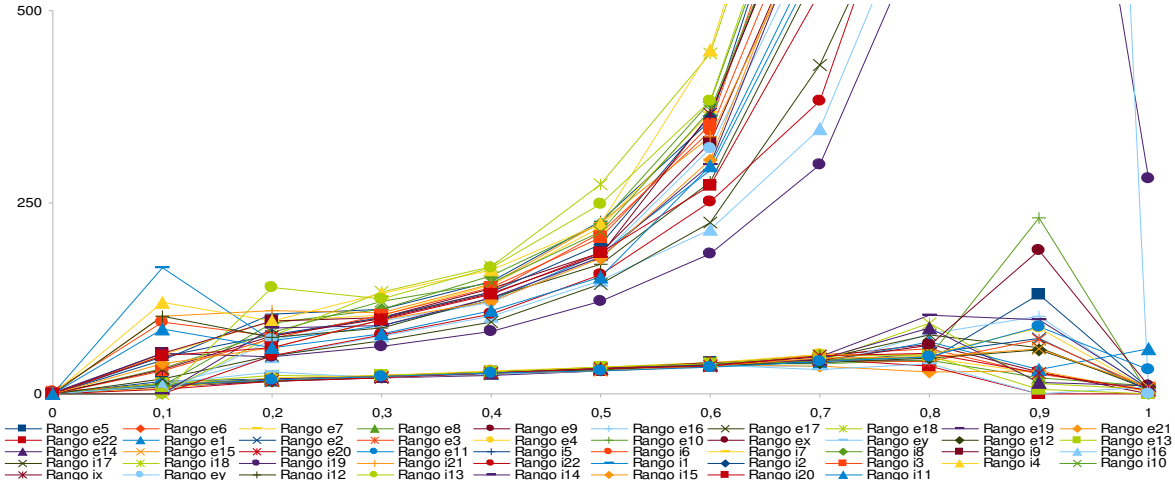
Grafica Mínimos H.sapiens



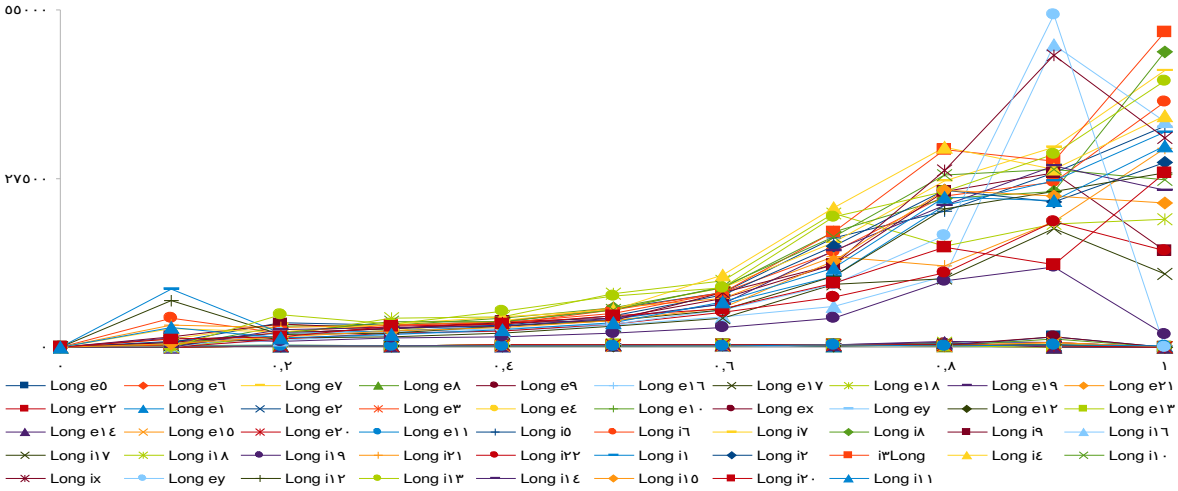
Grafica ampliada (Zoom) de los mínimos del H.sapiens



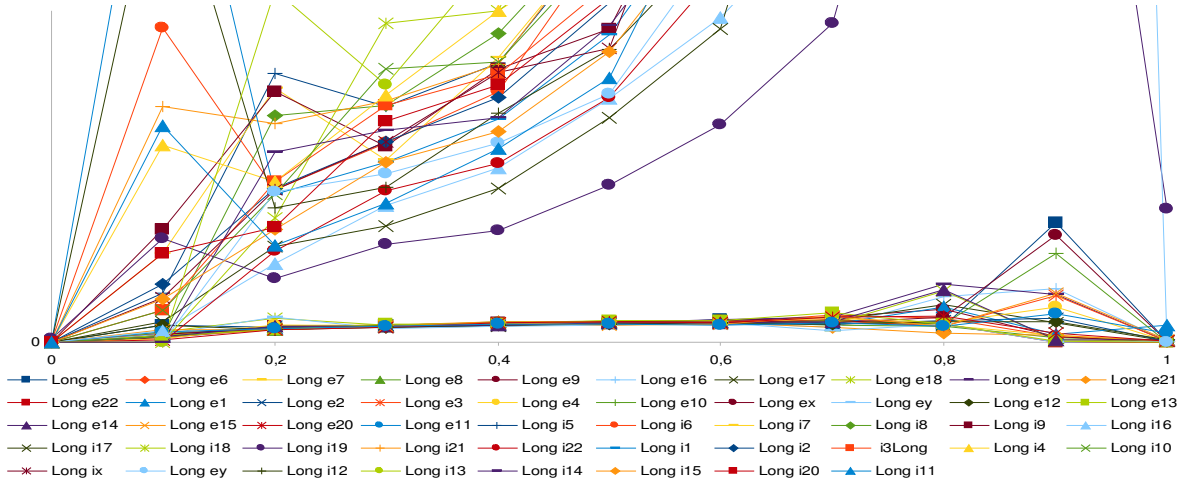
Grafica Rangos H.sapiens



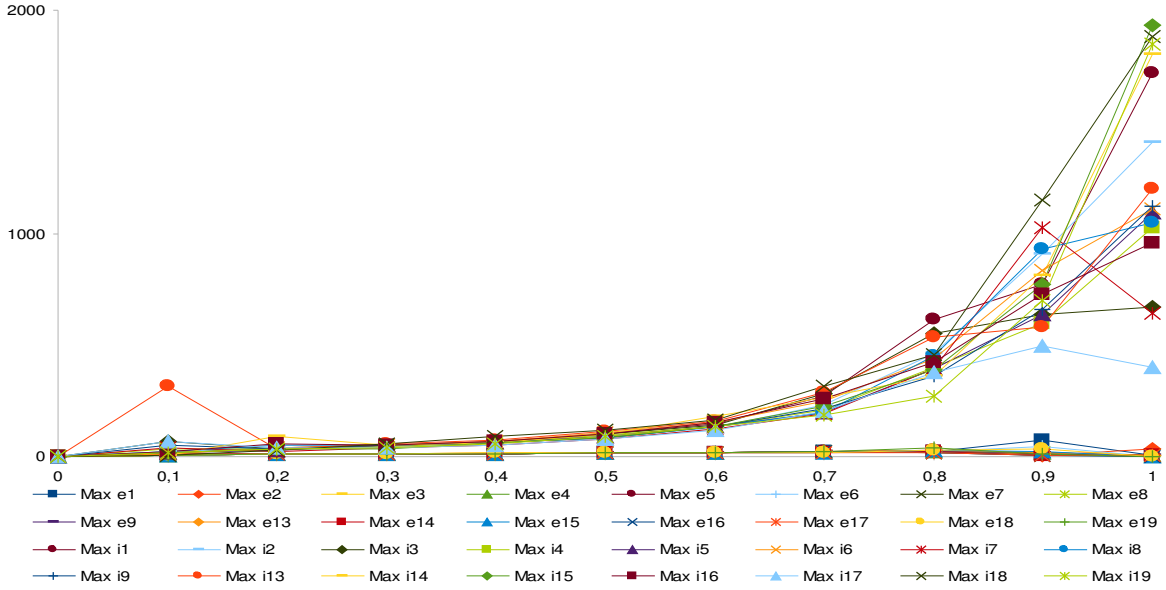
Grafica Zoom de Rangos H.sapiens



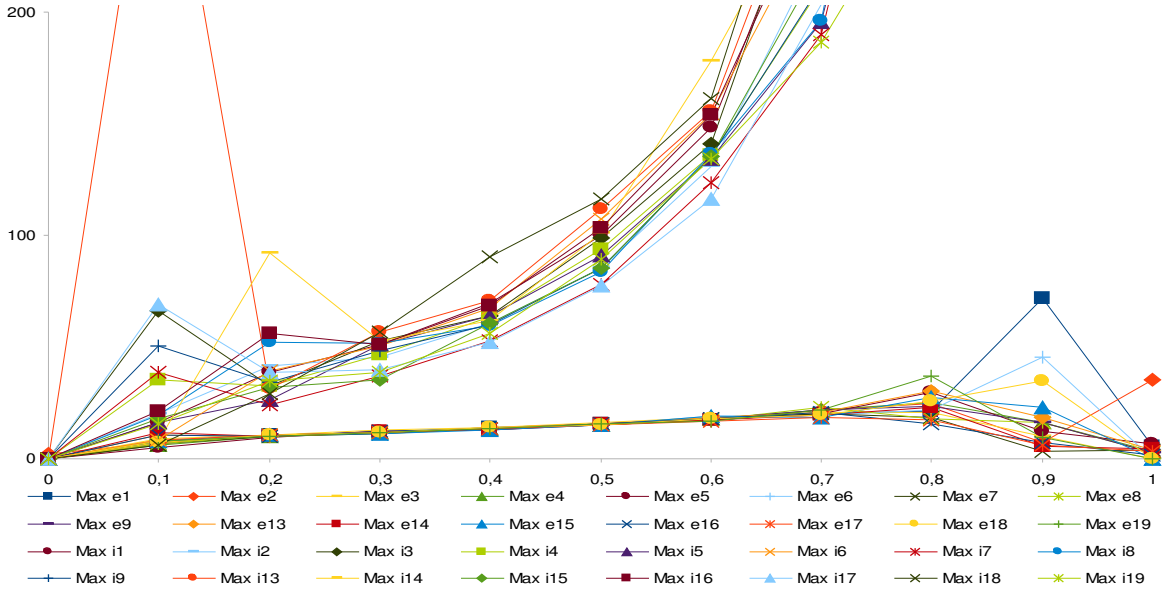
Grafica longitud máximo H.sapiens



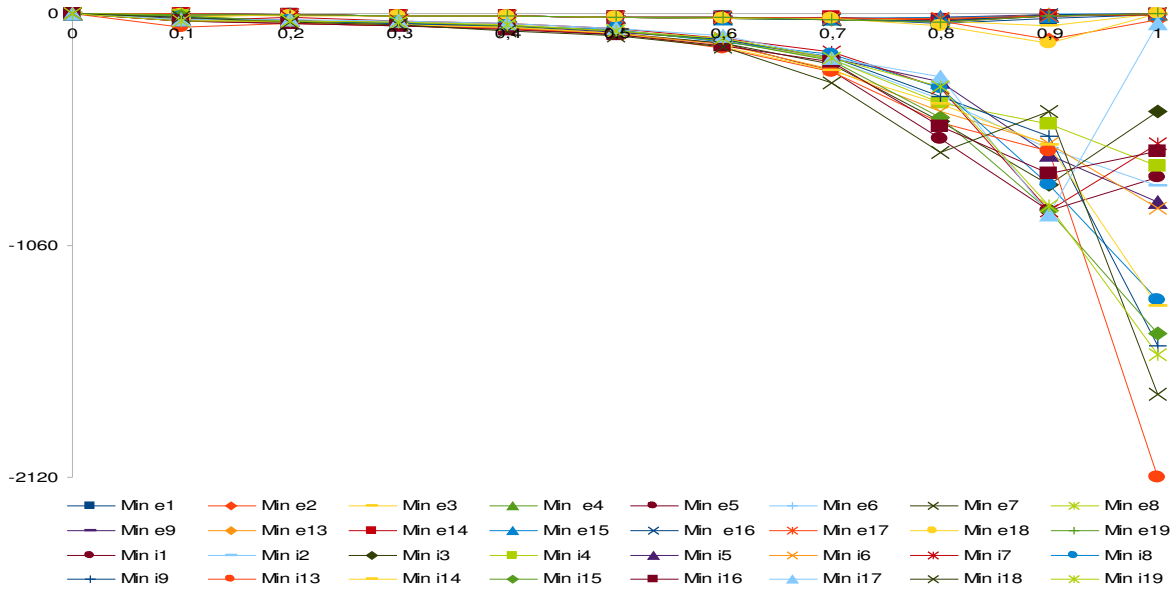
Zoom longitud máximos de H.sapiens



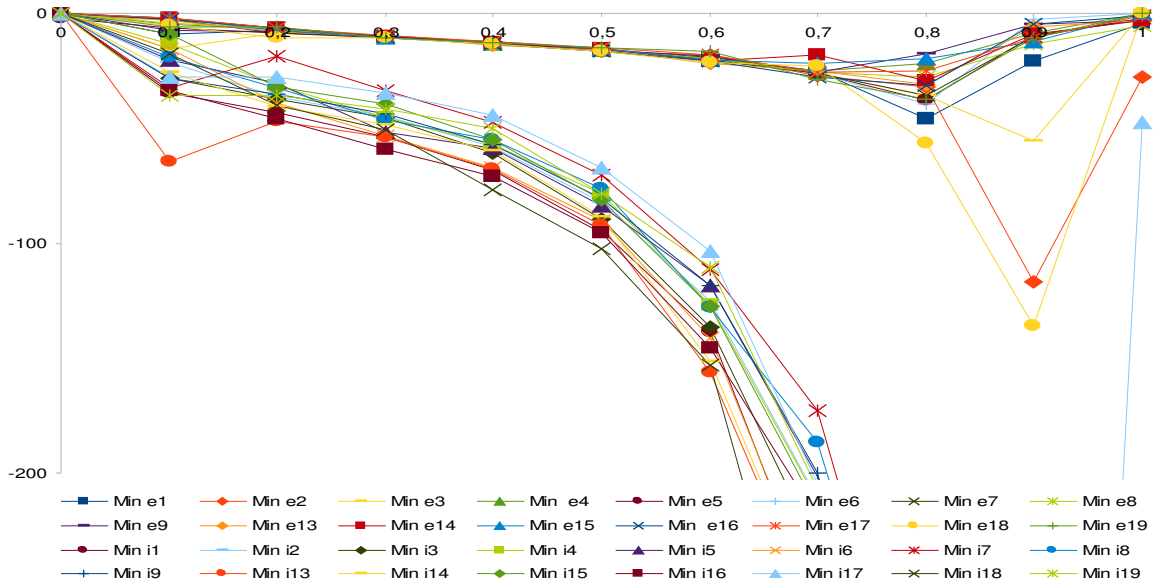
Máximos M.musculus



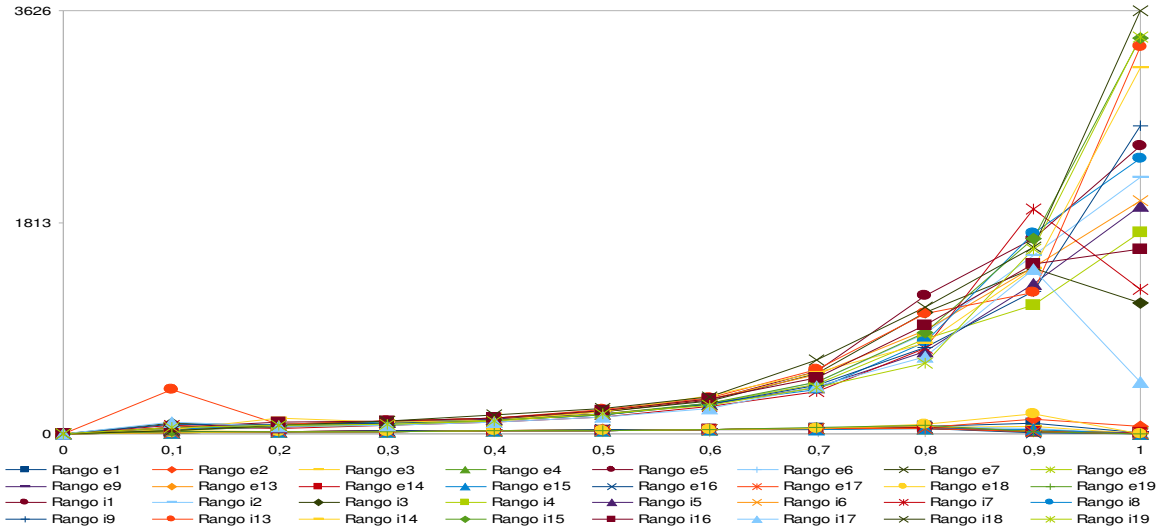
Zoom máximos M.musculus



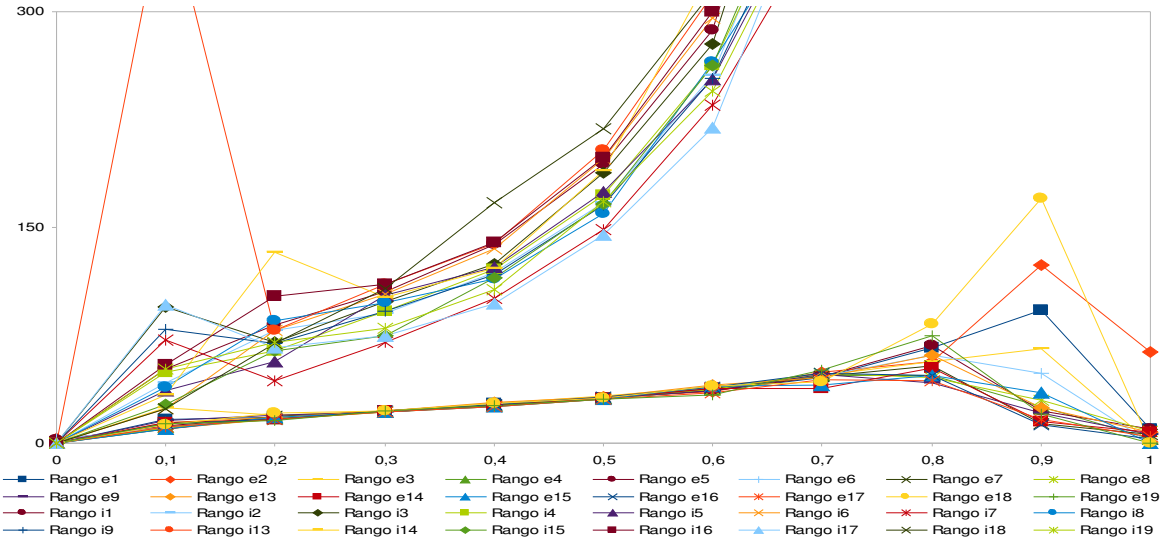
Mínimo M.musculus



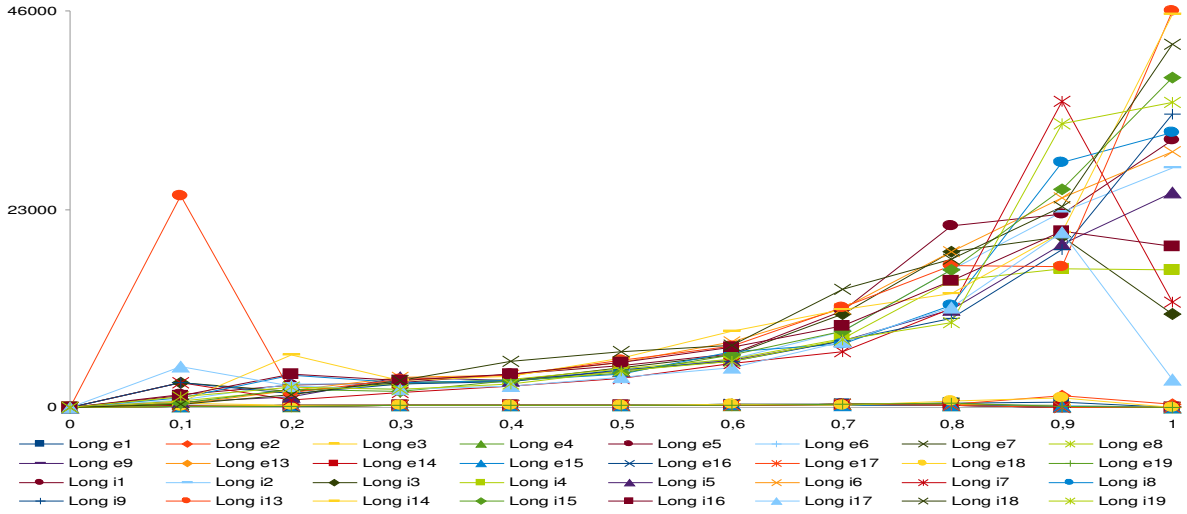
Zoom mínimo M.musculus



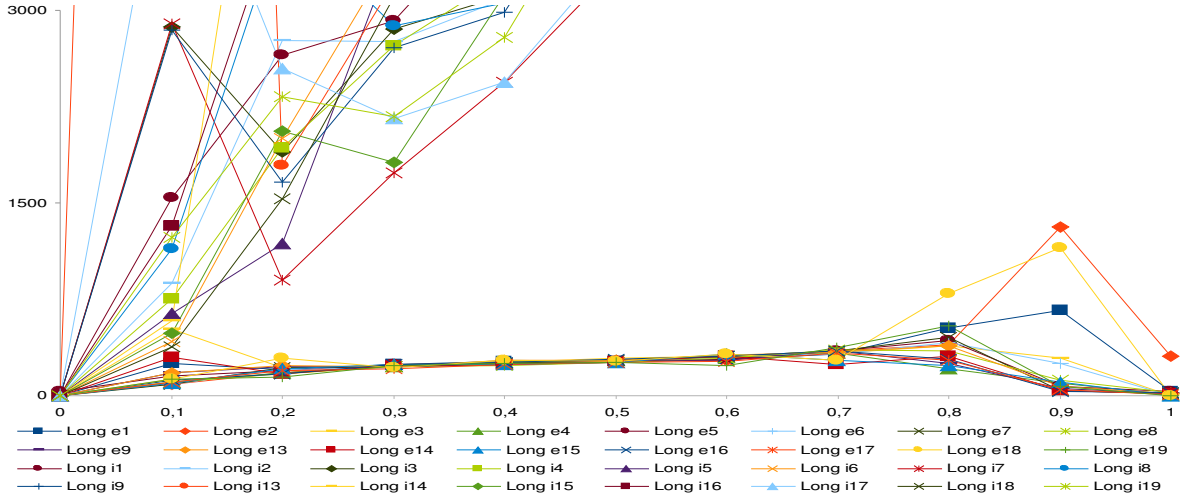
Rango M.musculus



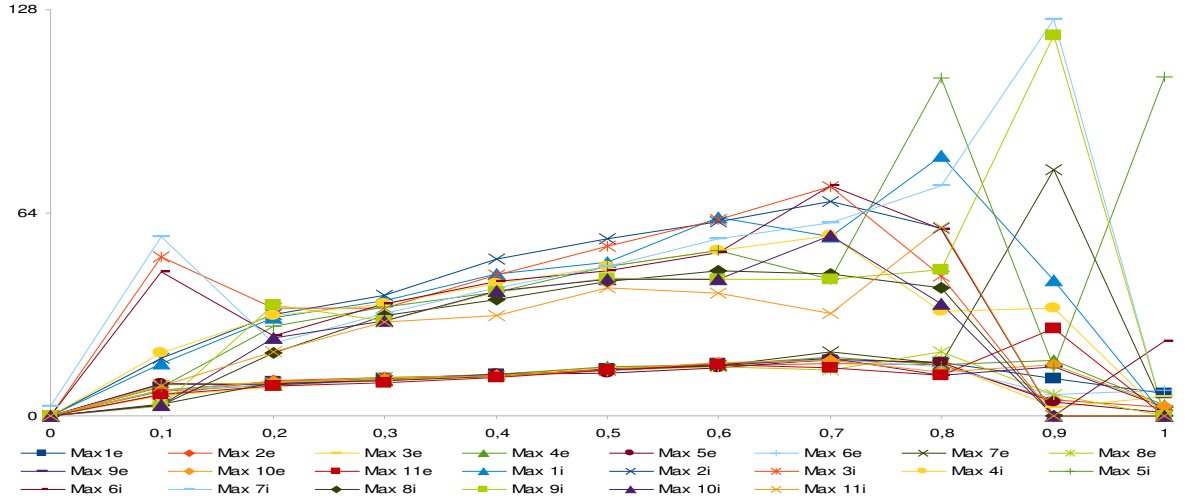
Zoom Rango M.musculus



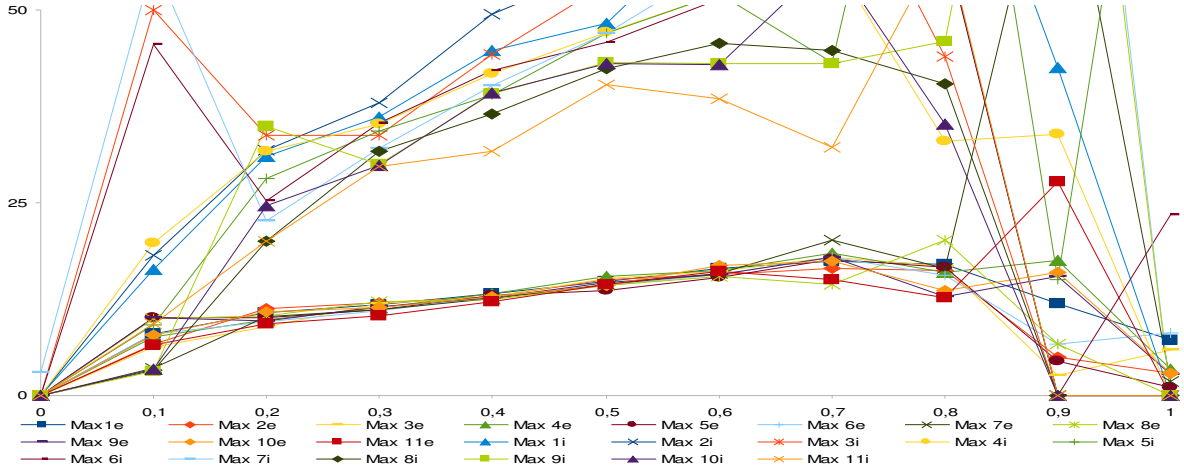
Longitud M.musculus



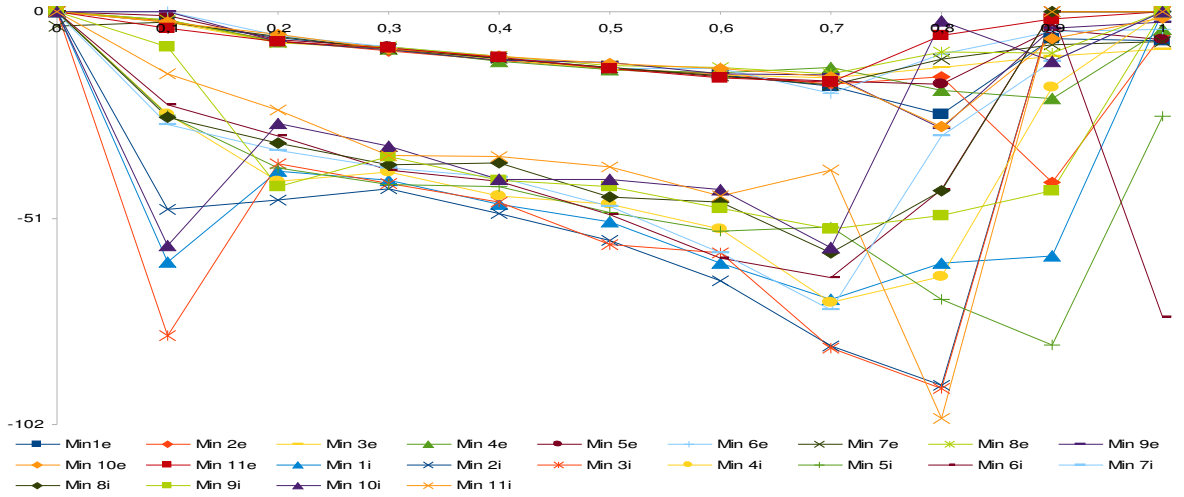
Zoom Longitud M.musculus



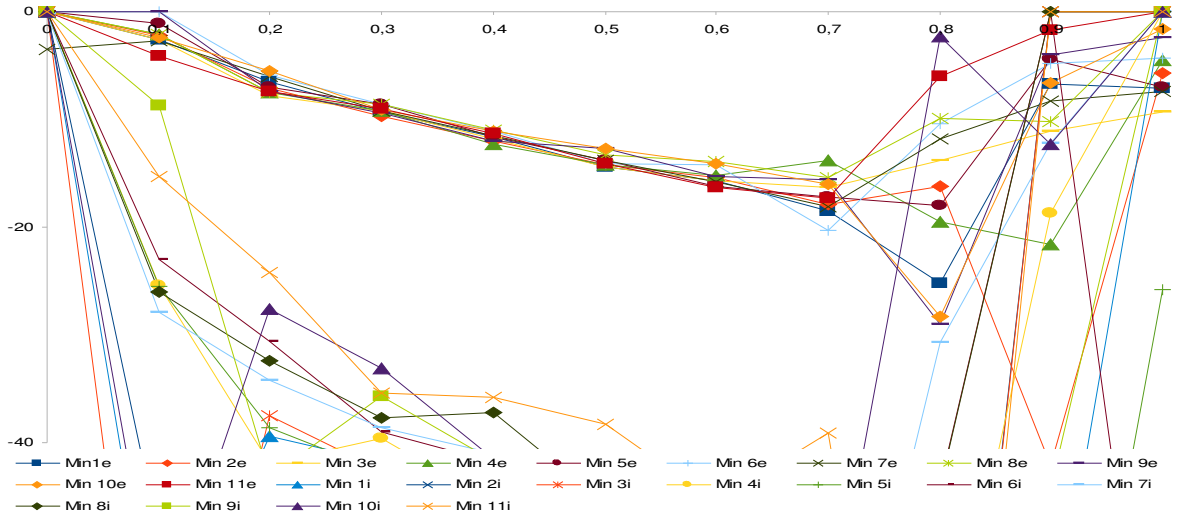
Máximos G.gallus (cromosomas 1-11)



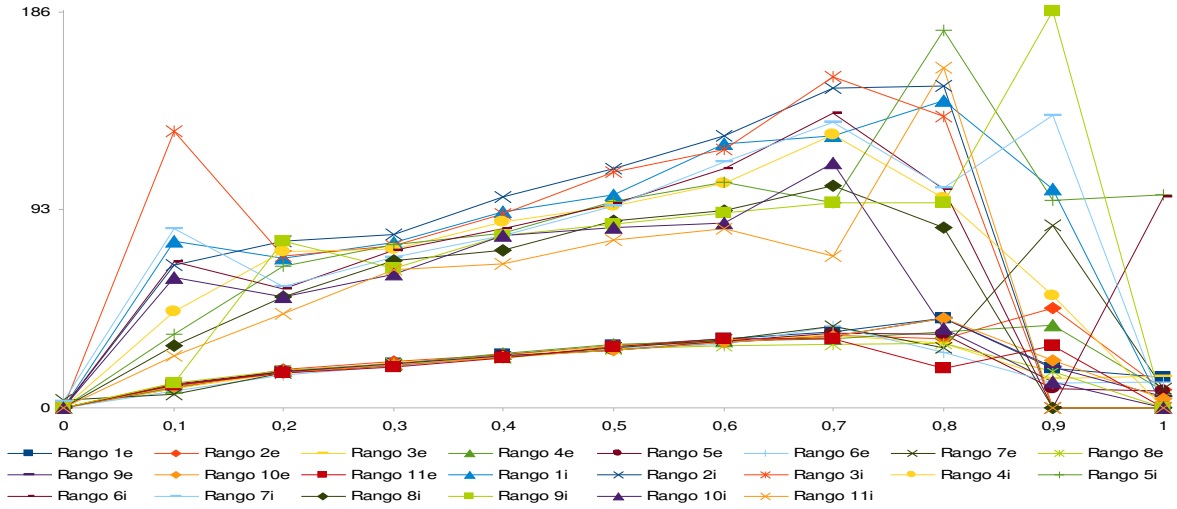
Zoom Máximos G.gallus (cromosomas 1-11)



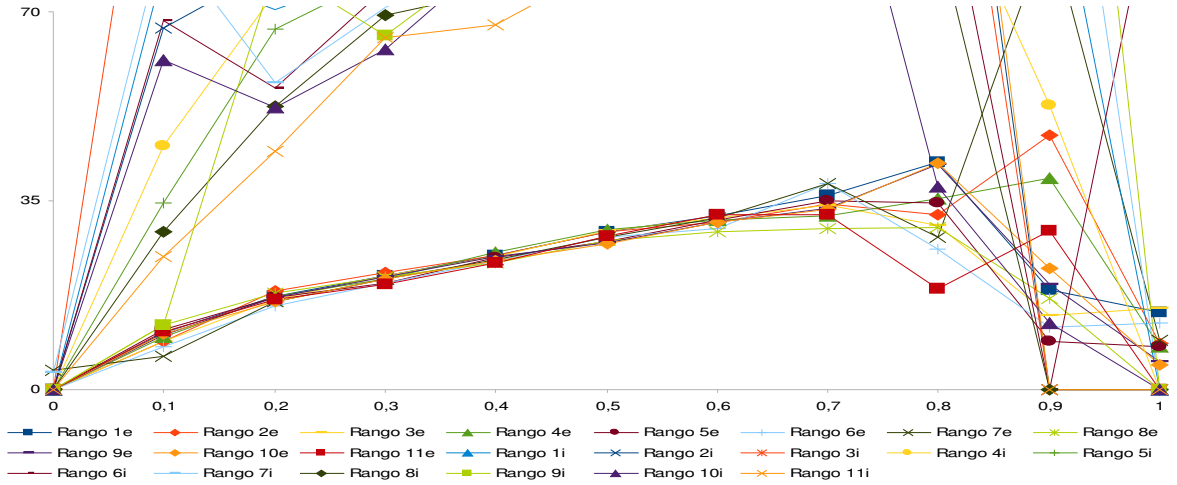
Minimos G.gallus (cromosomas 1-11)



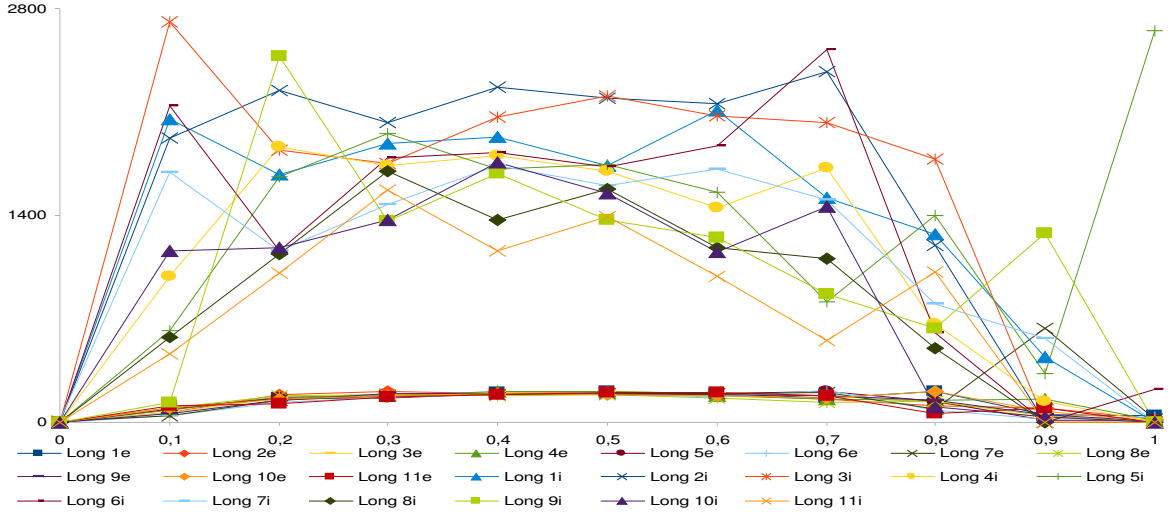
Zoom Minimos G.gallus (cromosomas 1-11)



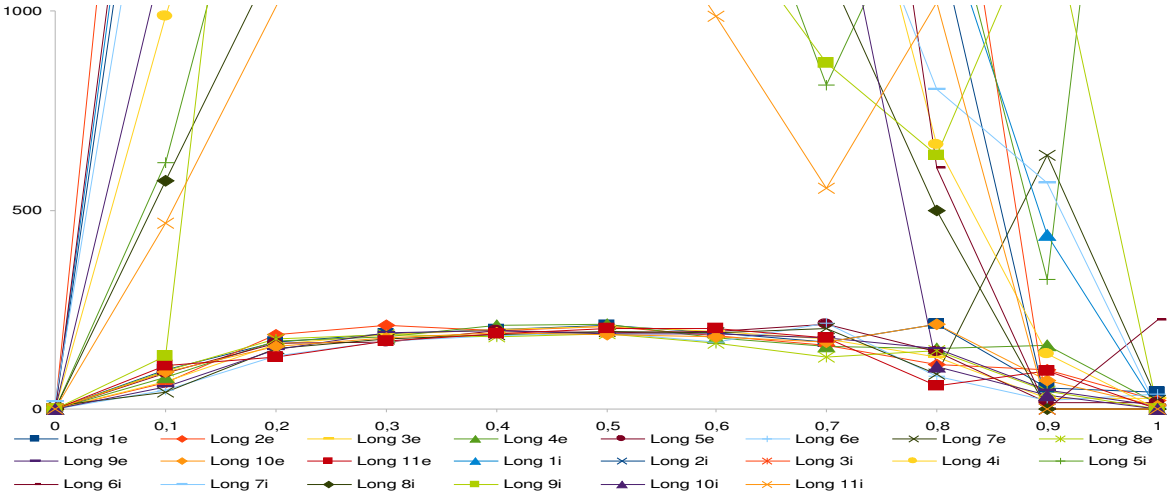
Rangos G.gallus (cromosomas 1-11)



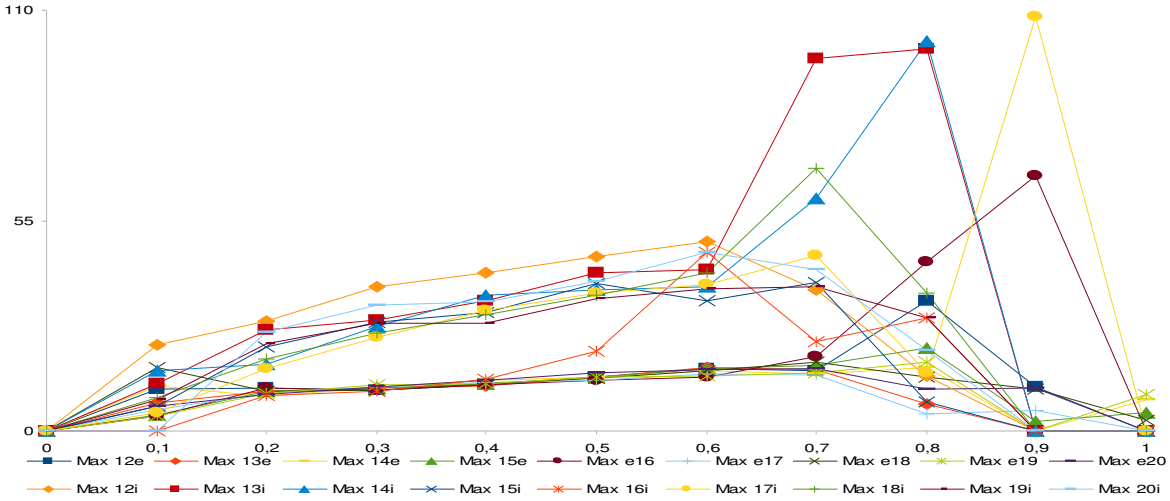
Zoom Rangos G.gallus (cromosomas 1-11)



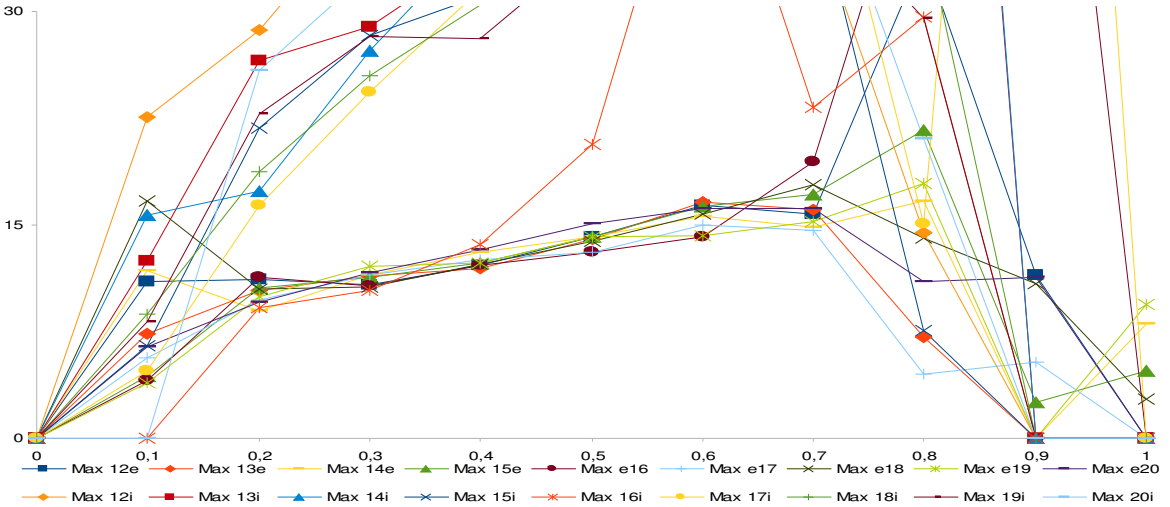
Longitudes G.gallus (cromosomas 1-11)



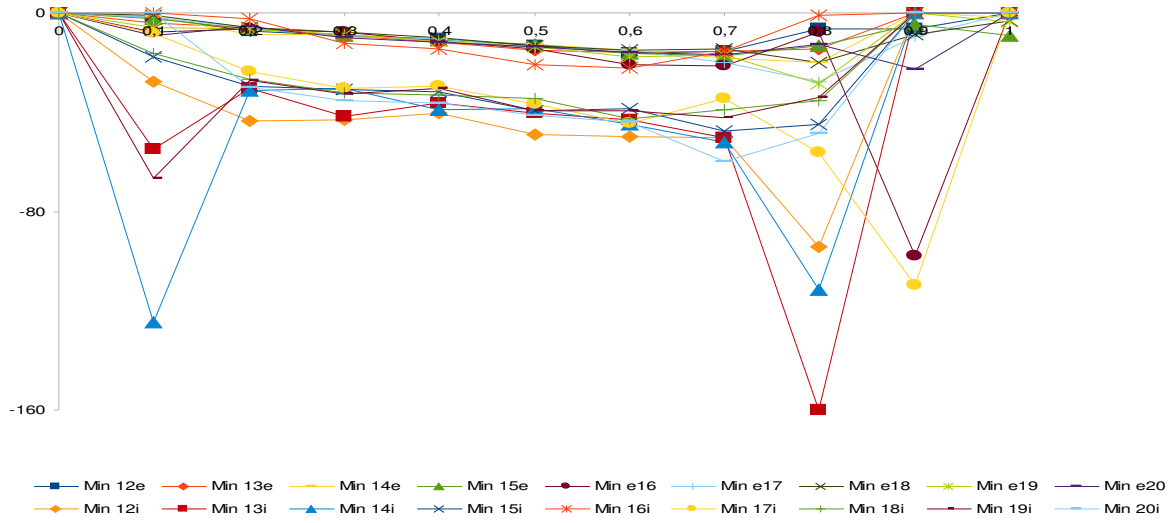
Zoom Longitudes G.gallus (cromosomas 1-11)



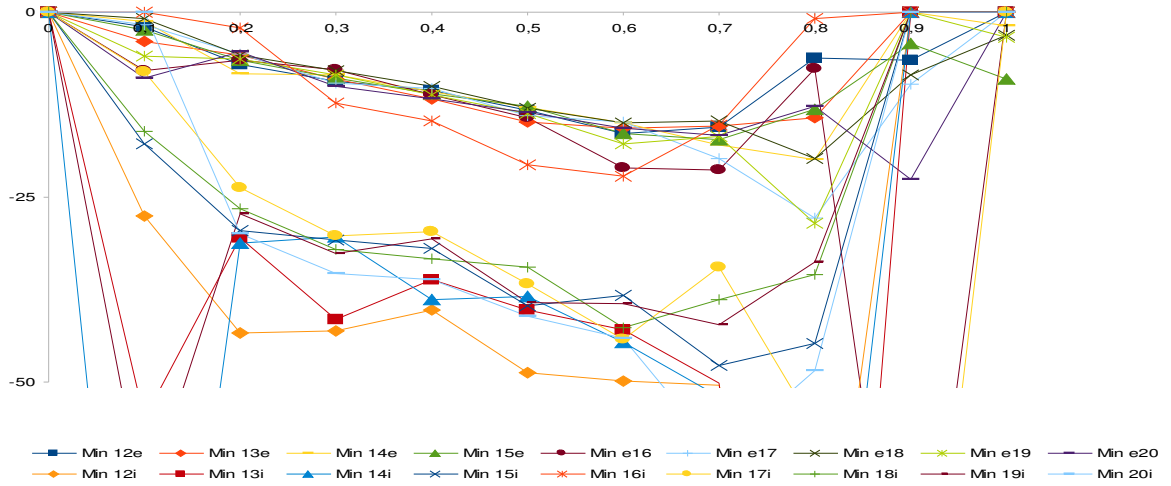
Maximo G.gallus (cromosomas 12-20)



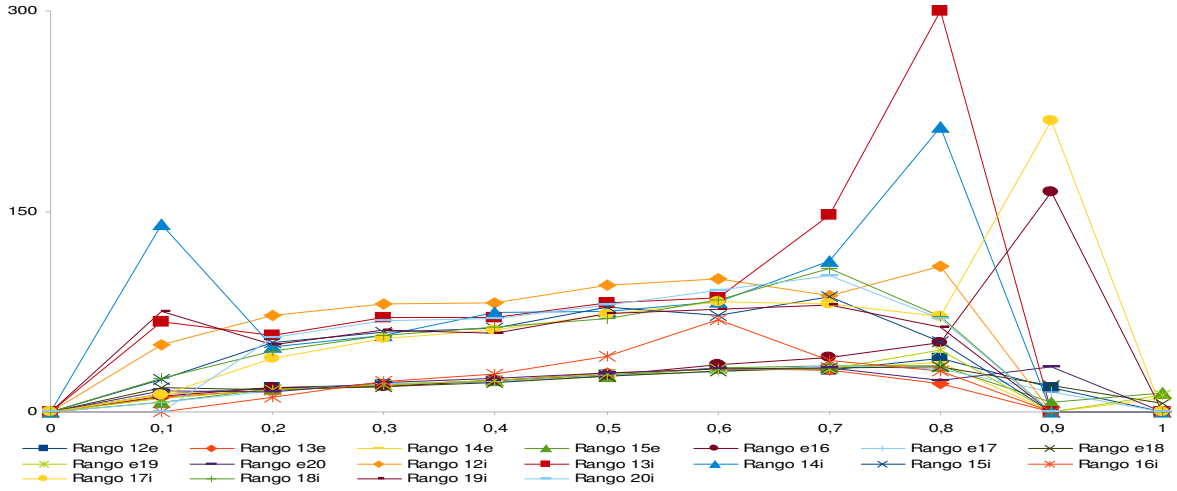
Zoom Maximo G.gallus (cromosomas 12-20)



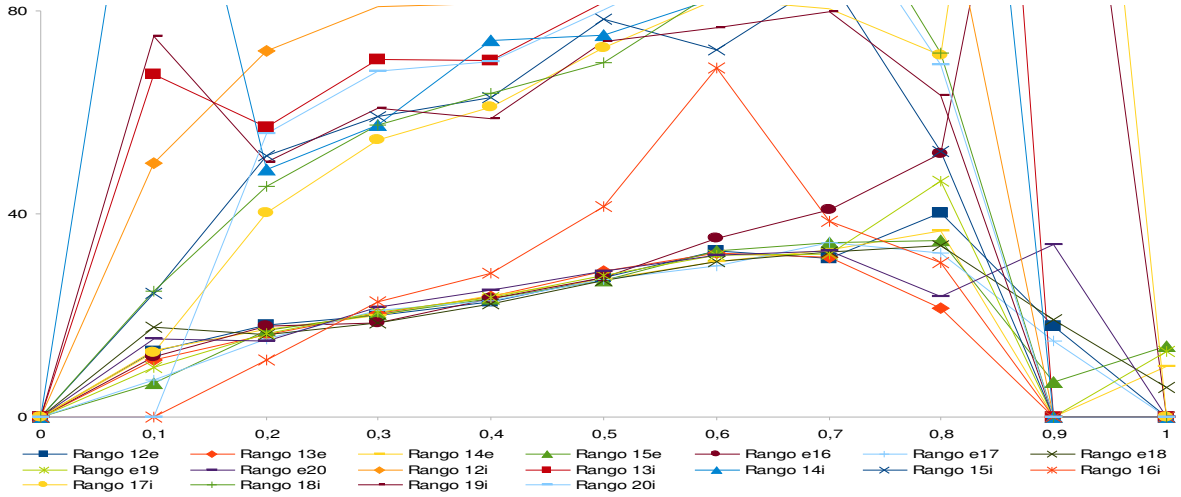
Minimo G.gallus (cromosomas 12-20)



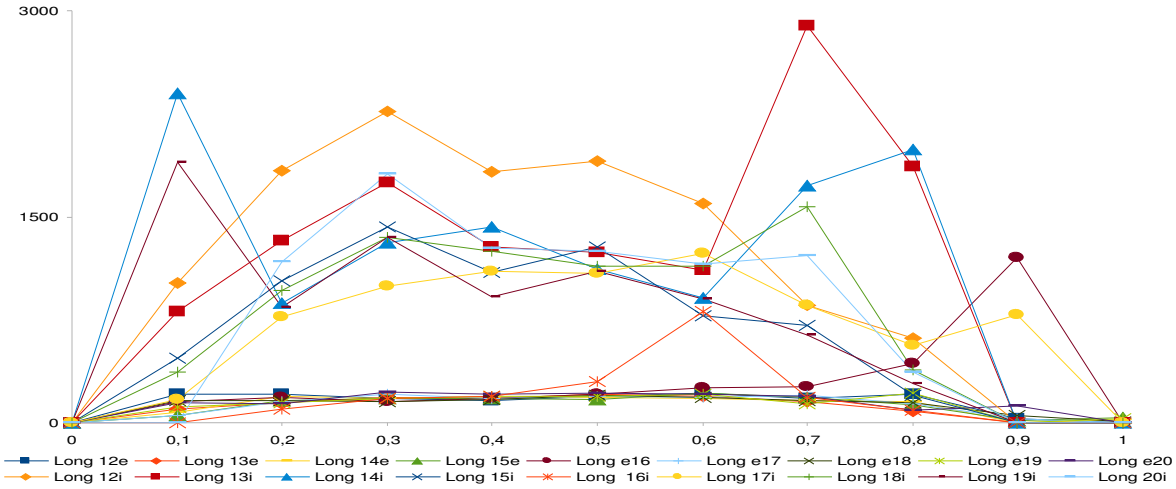
Zoom Minimo G.gallus (cromosomas 12-20)



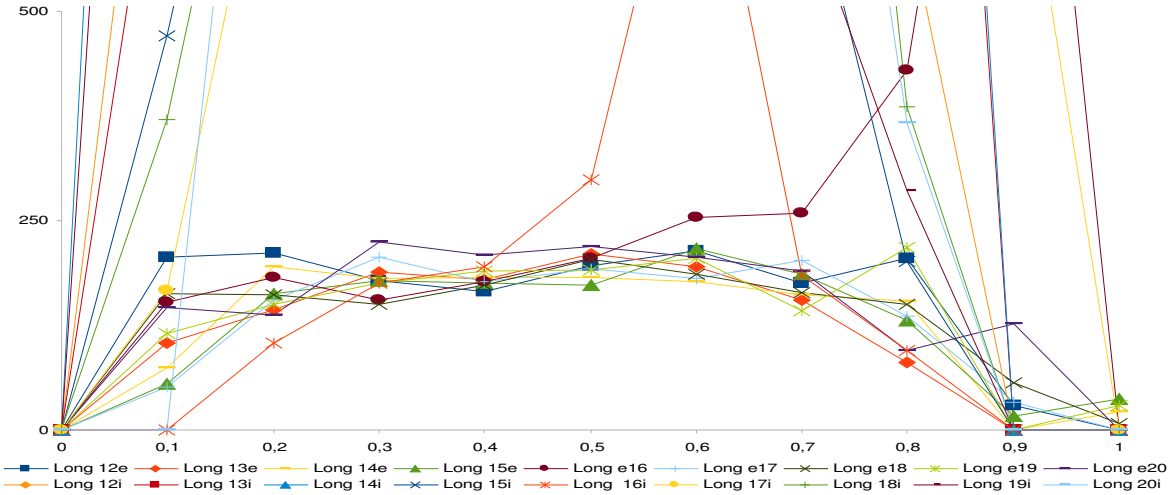
Rangos G.gallus (cromosomas 12-20)



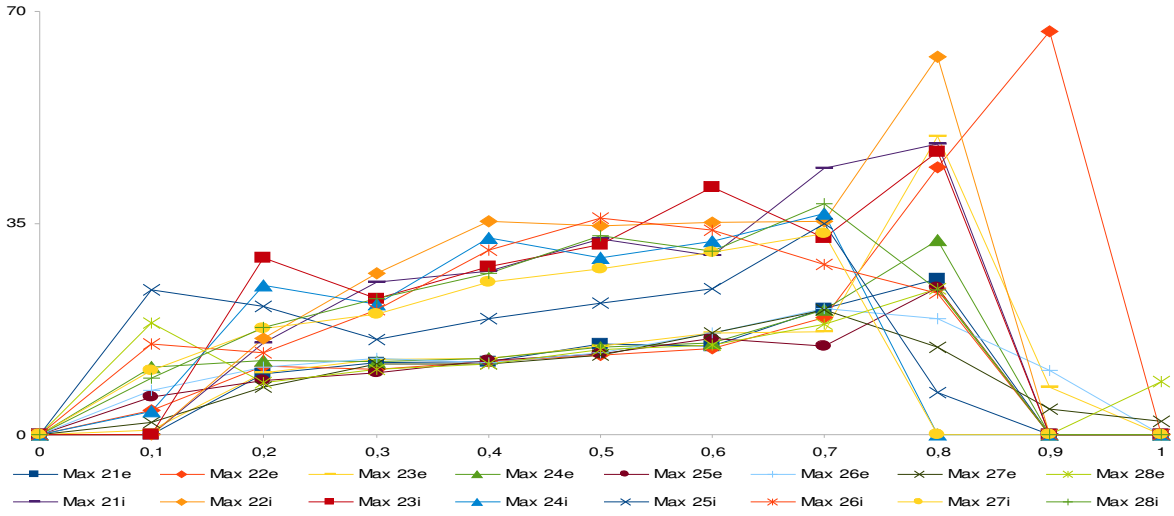
Zoom Rangos G.gallus (cromosomas 12-20)



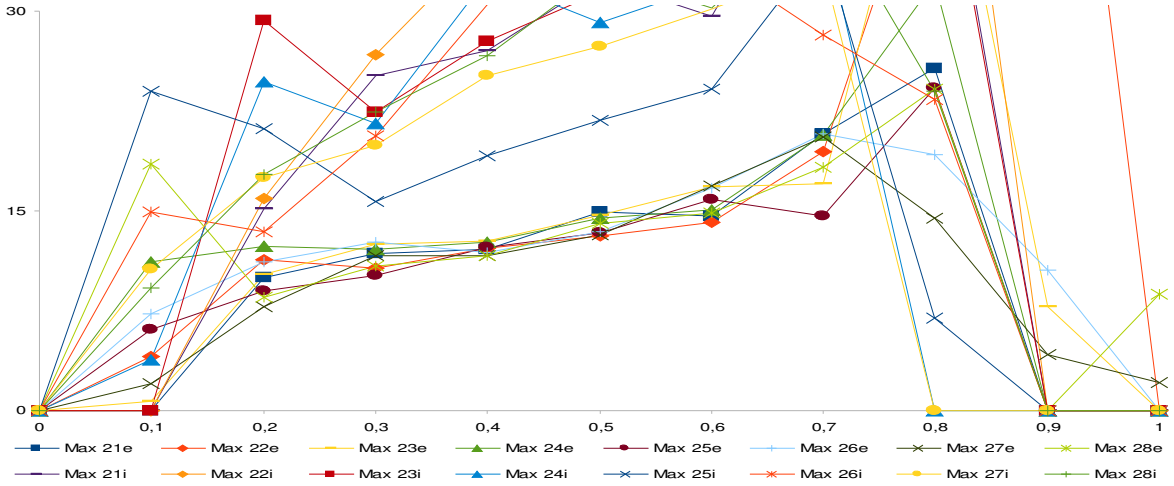
Longitud G.gallus (cromosomas 12-20)



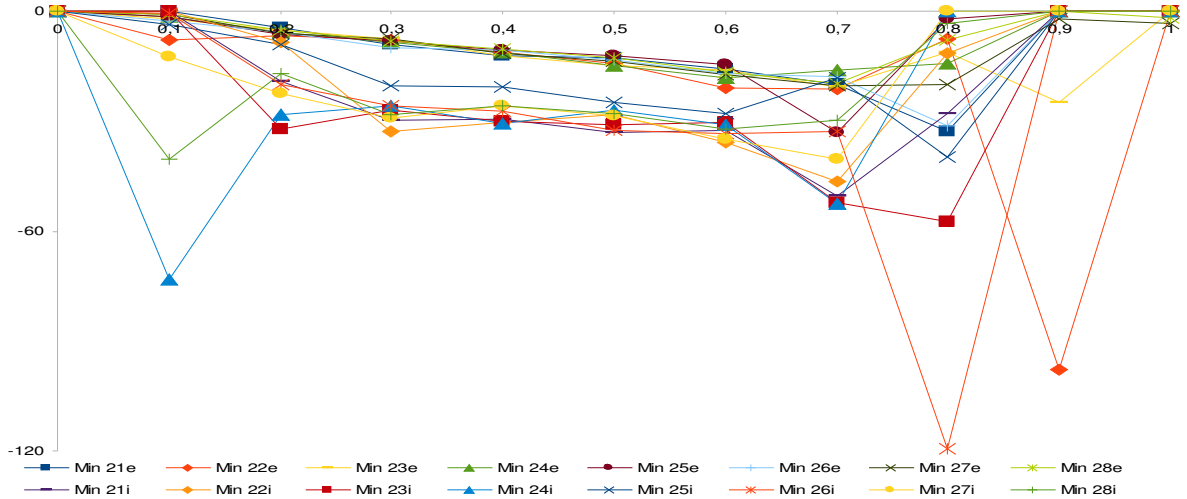
Zoom Longitud G.gallus (cromosomas 12-20)



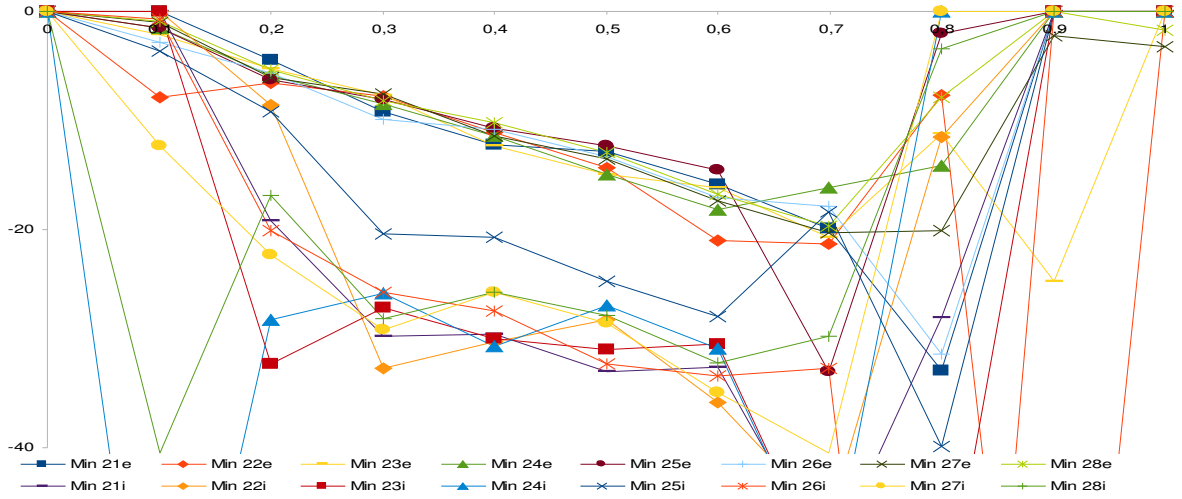
Maximos G.gallus (cromosomas 21-28)



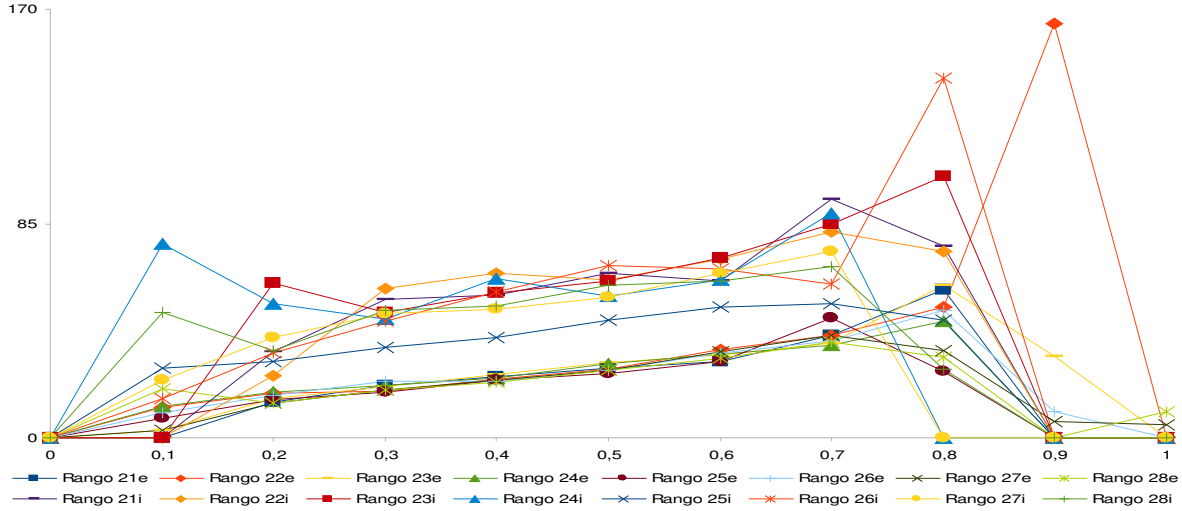
Zoom Maximos G.gallus (cromosomas 21-28)



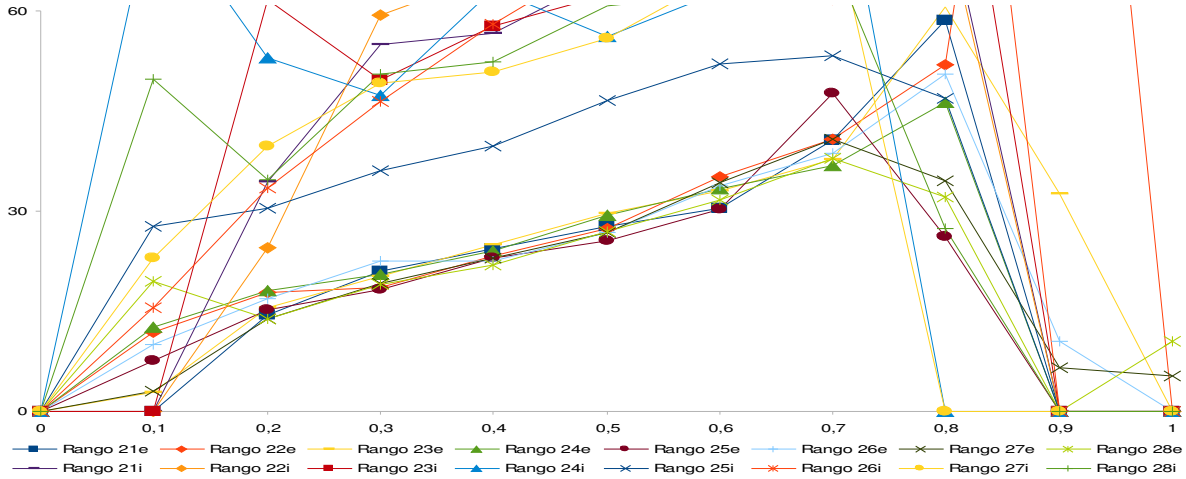
Minimos G.gallus (cromosomas 21-28)



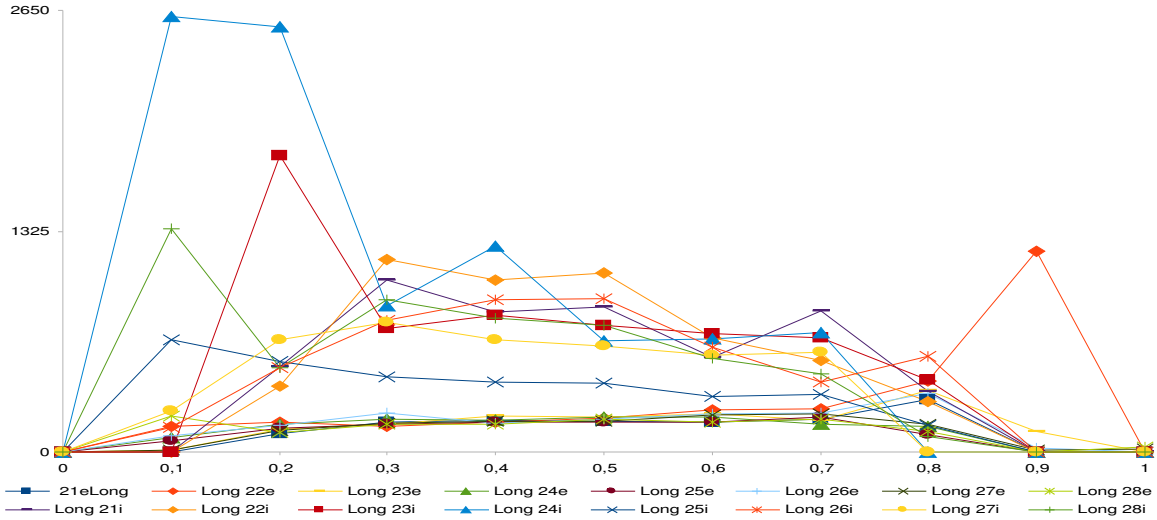
Zoom Minimos G.gallus (cromosomas 21-28)



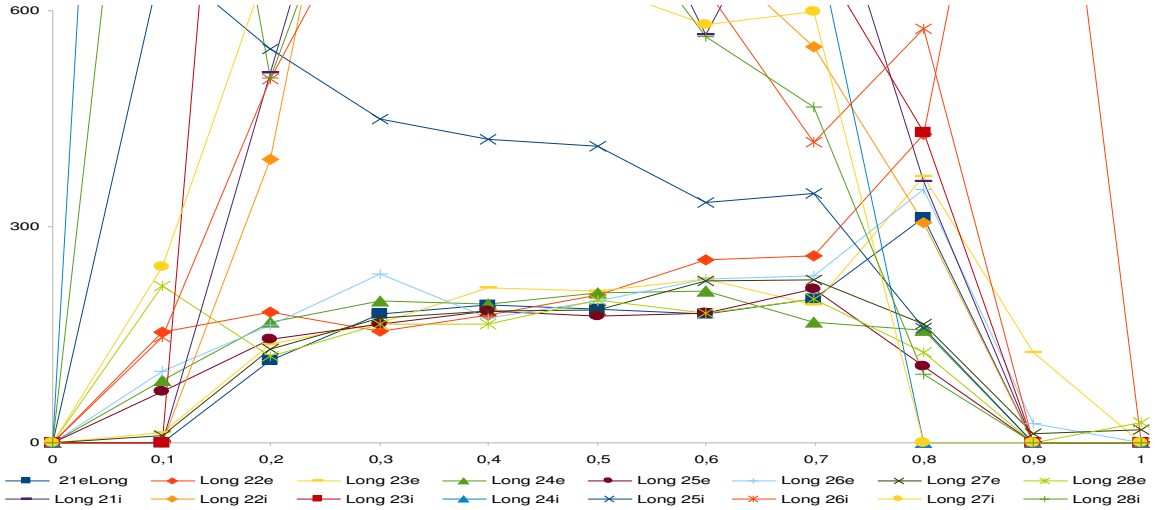
Rango G.gallus (cromosomas 21-28)



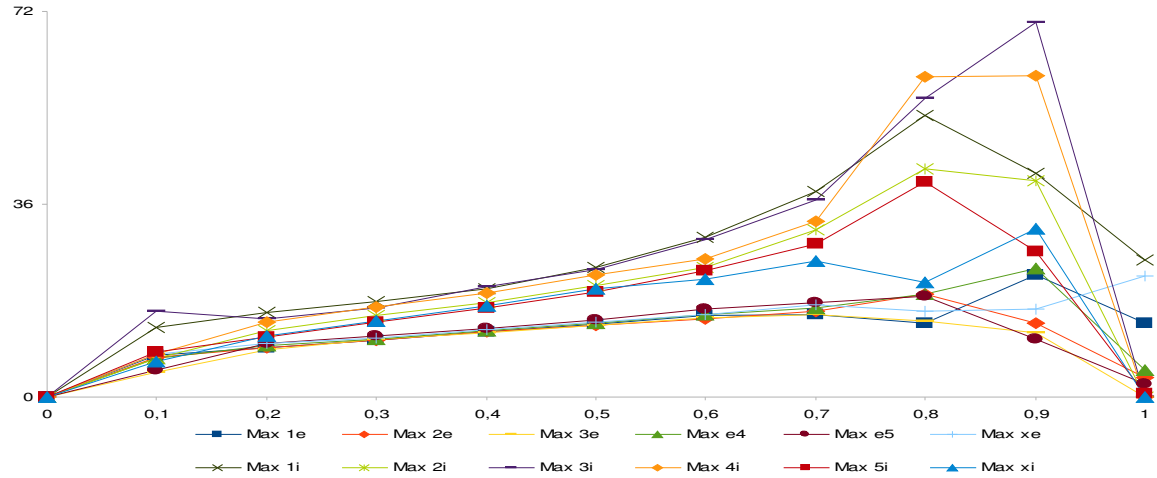
Zoom Rango G.gallus (cromosomas 21-28)



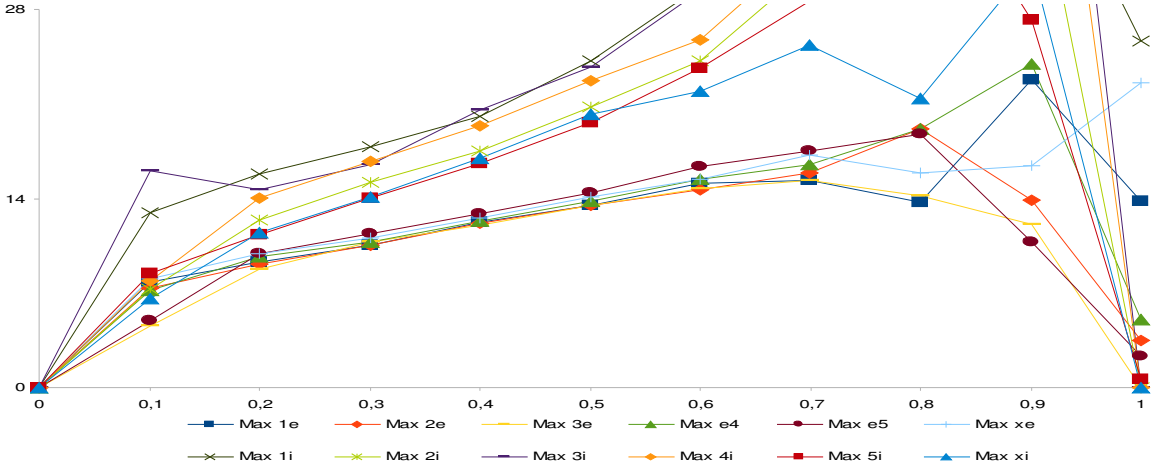
Longitud G.gallus (cromosomas 21-28)



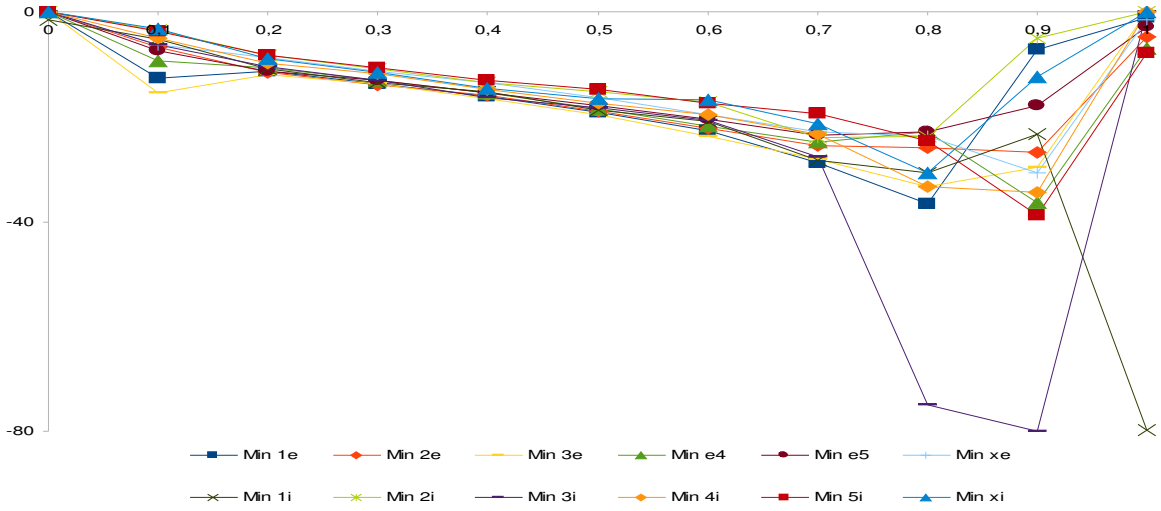
Zoom Longitud G.gallus (cromosomas 21-28)



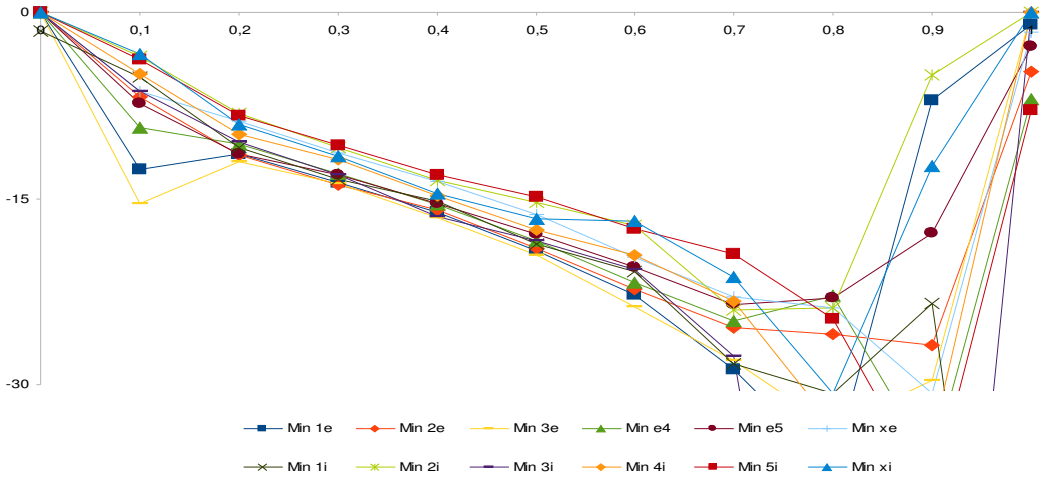
Máximos C.elegans



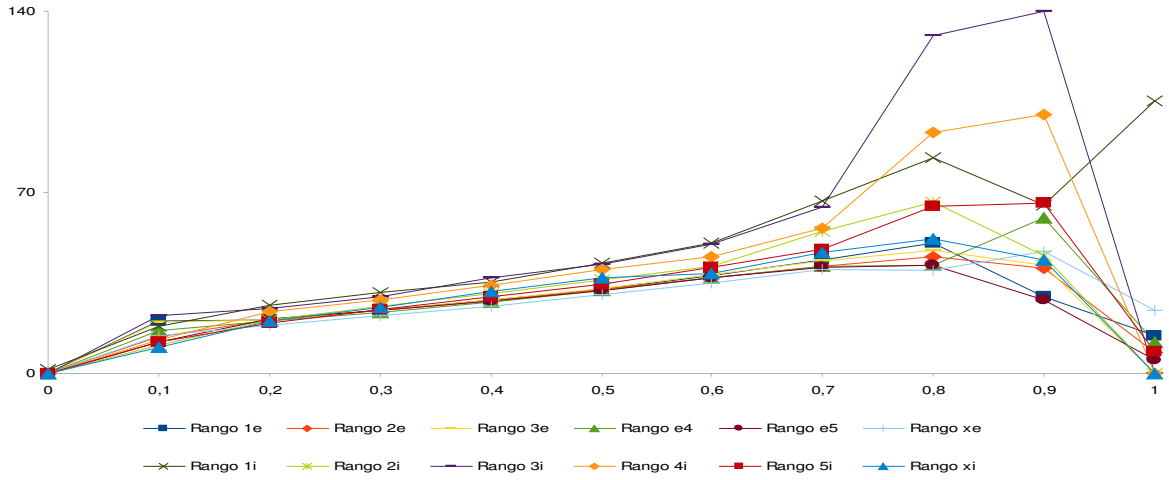
Zoom Máximos C.elegans



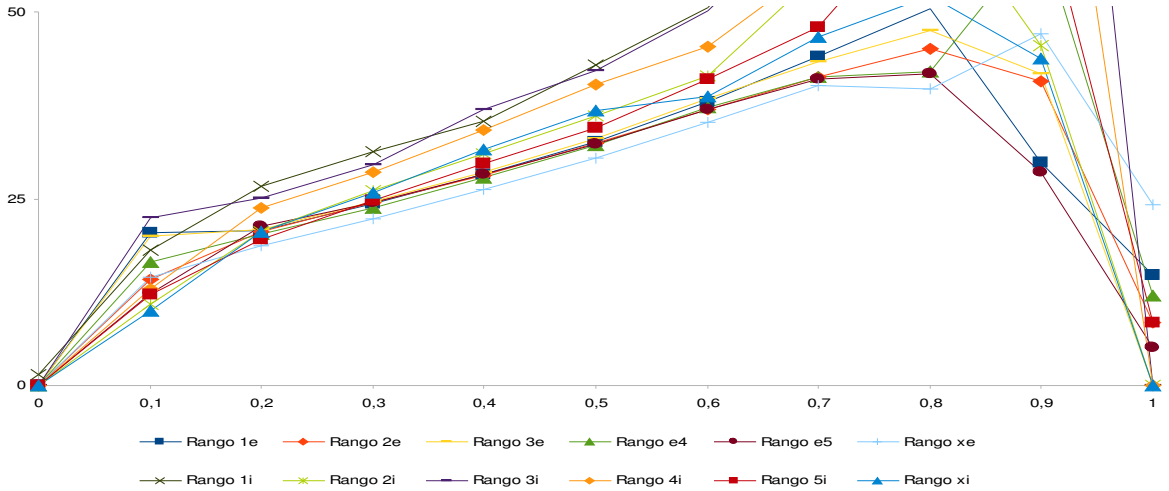
Mínimos C.elegans



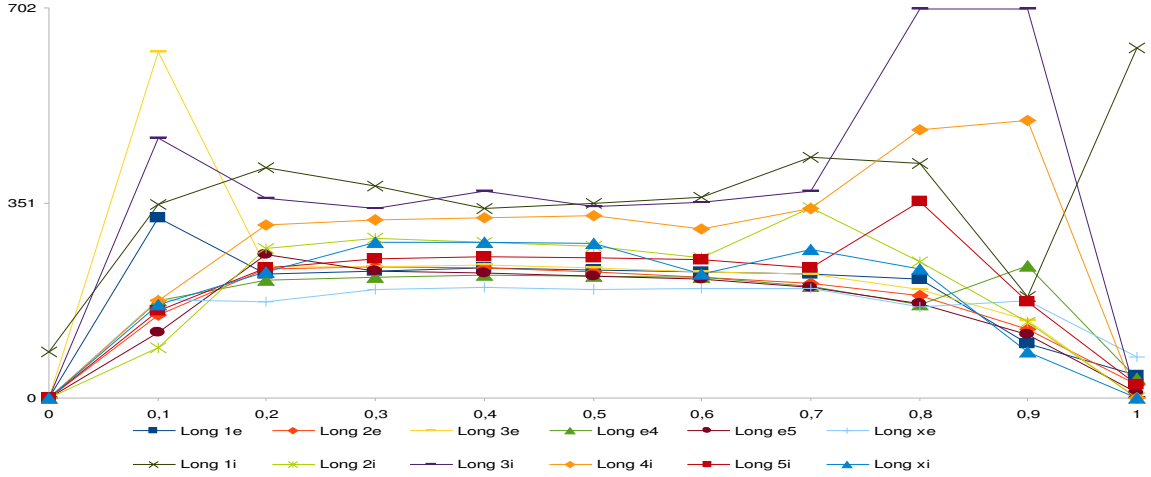
Zoom Mínimos C.elegans



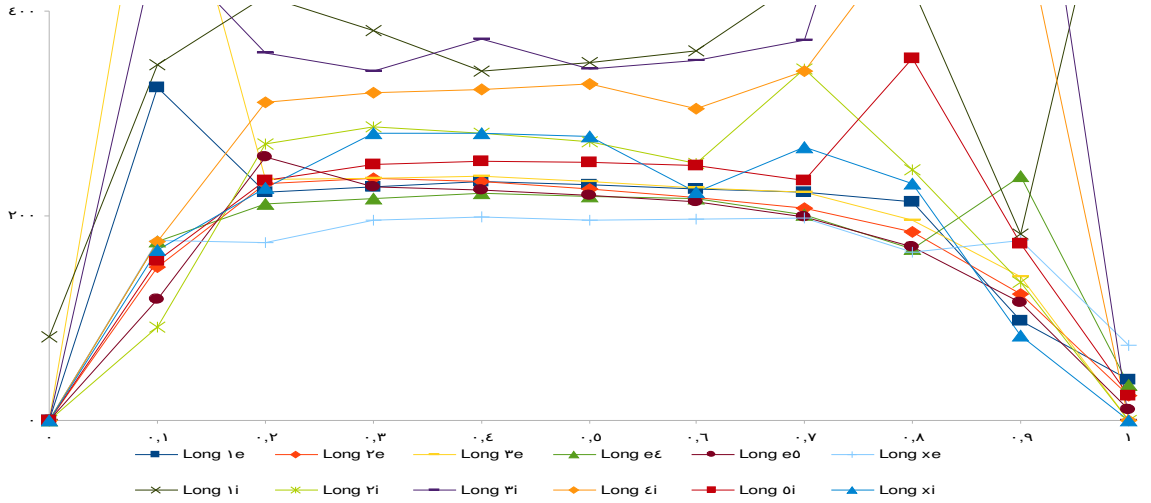
Rango C.elegans



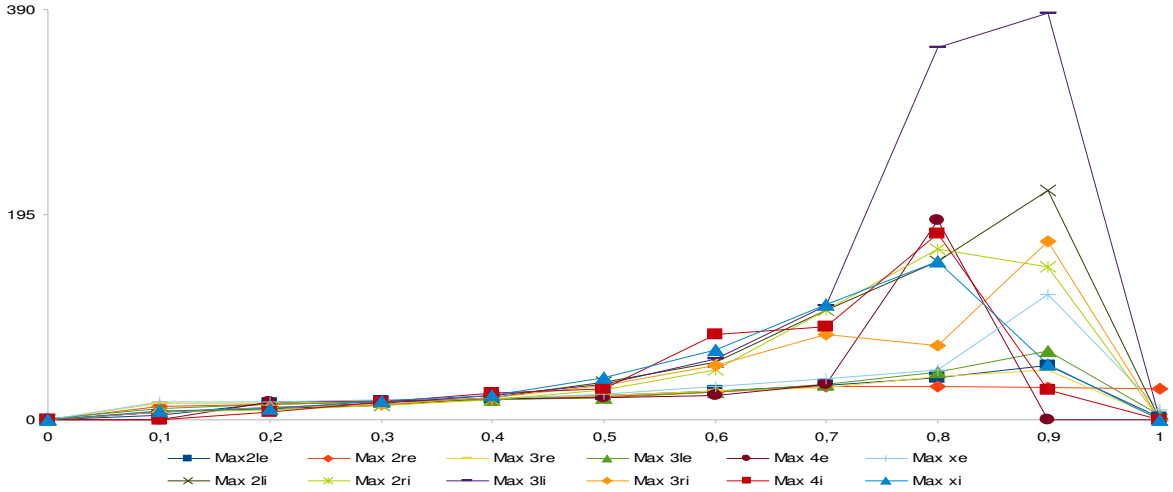
Zomm Rango C.elegans



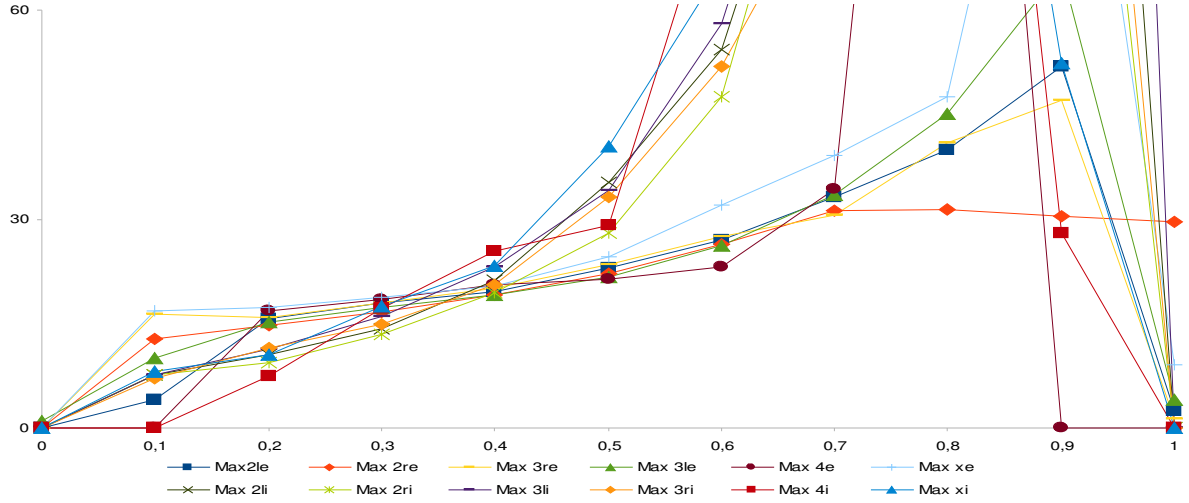
Longitud C.elegans



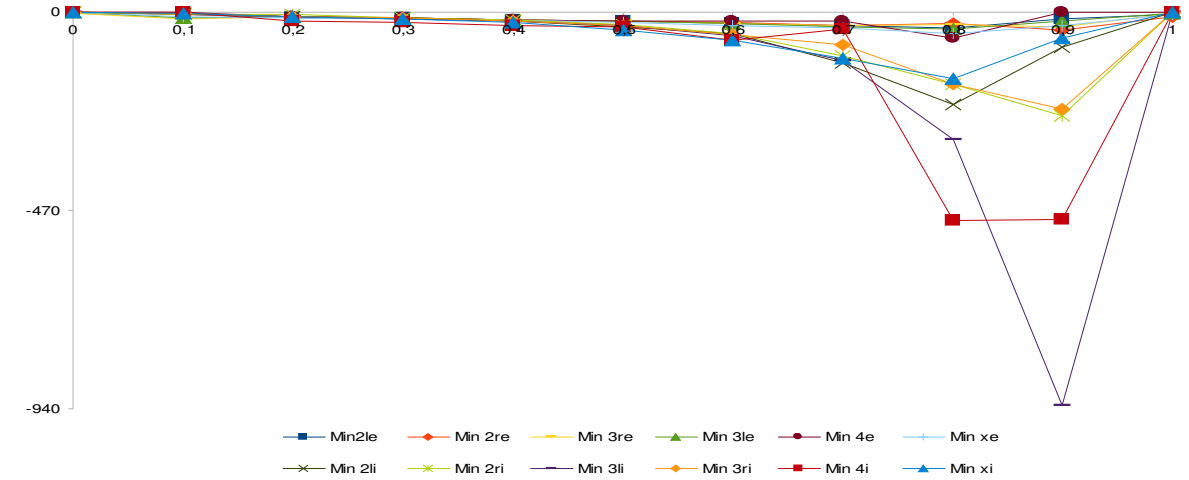
Zoom Longitud C.elegans



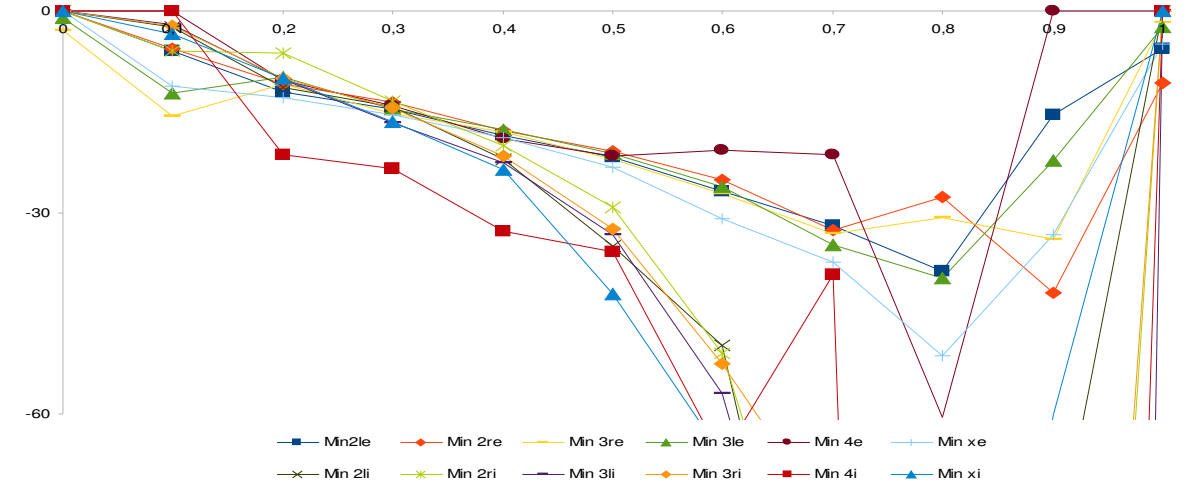
Máximos D.Melanogaster



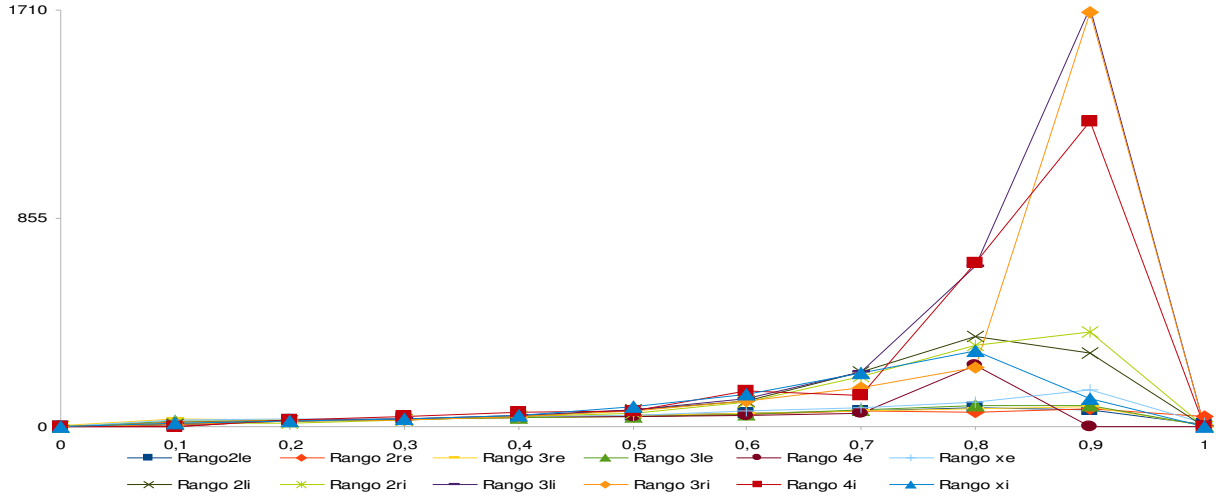
Zoom Máximos D.Melanogaster



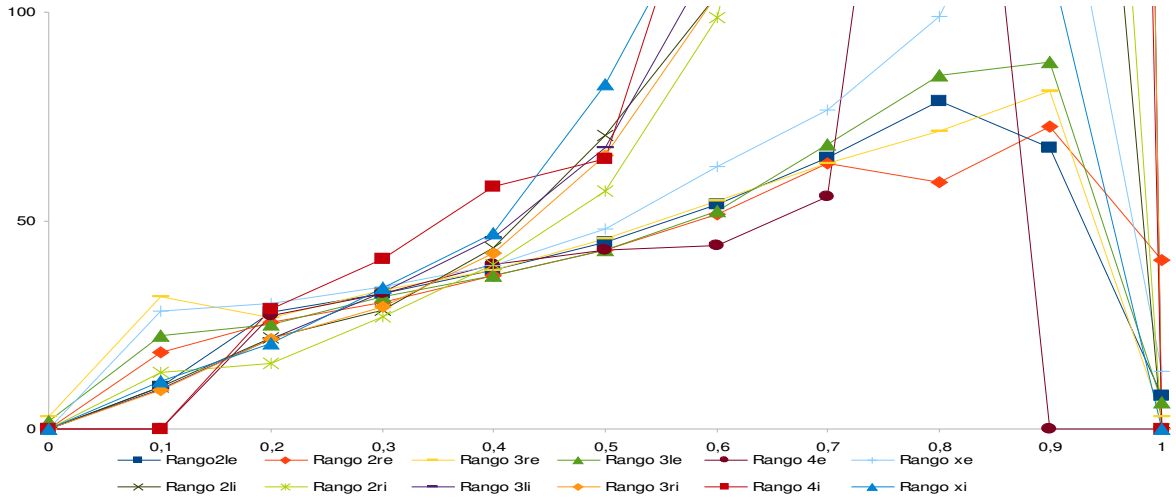
Mínimo D.Melanogaster



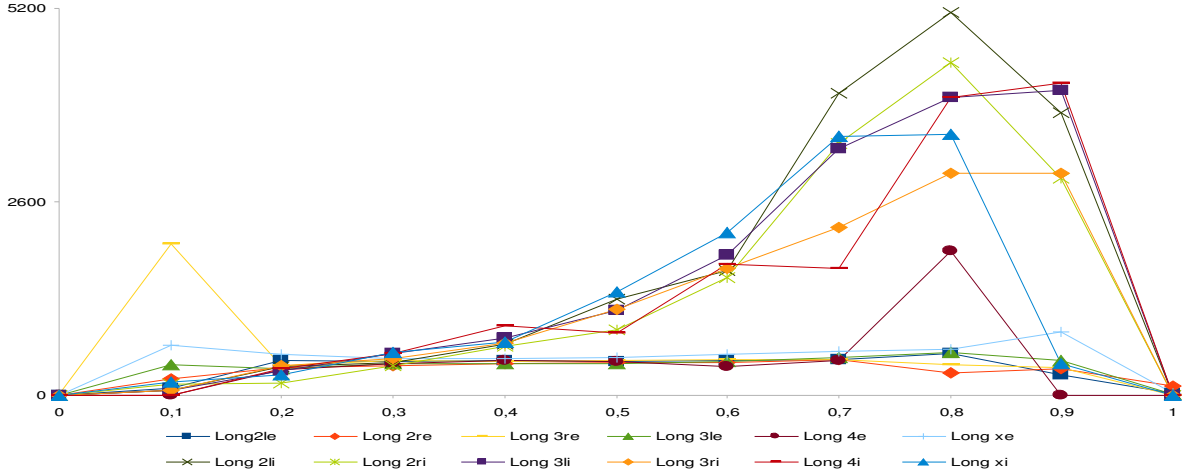
Zoom Mínimo D.Melanogaster



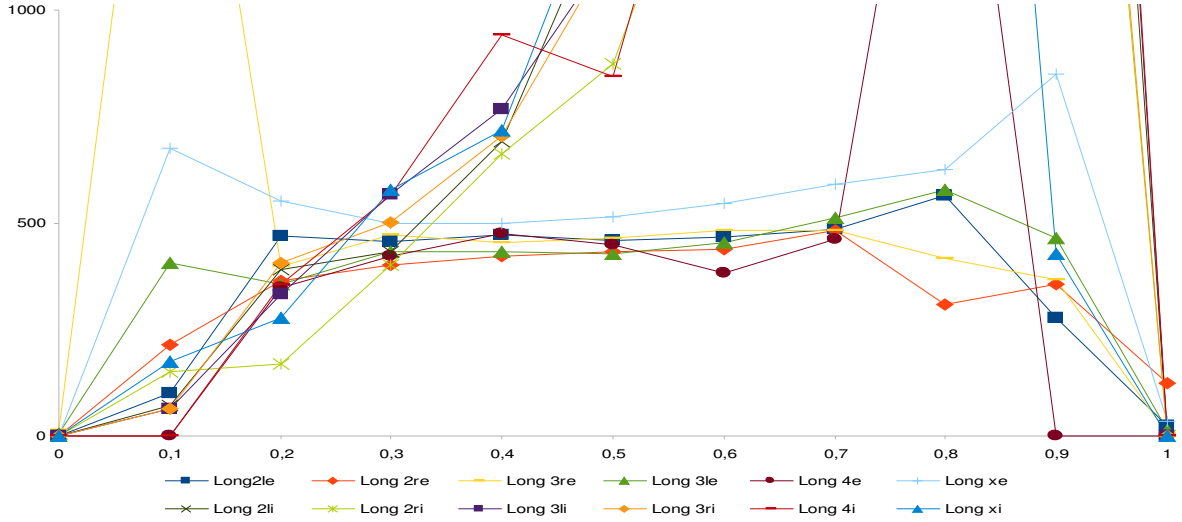
Rangos D.Melanogaster



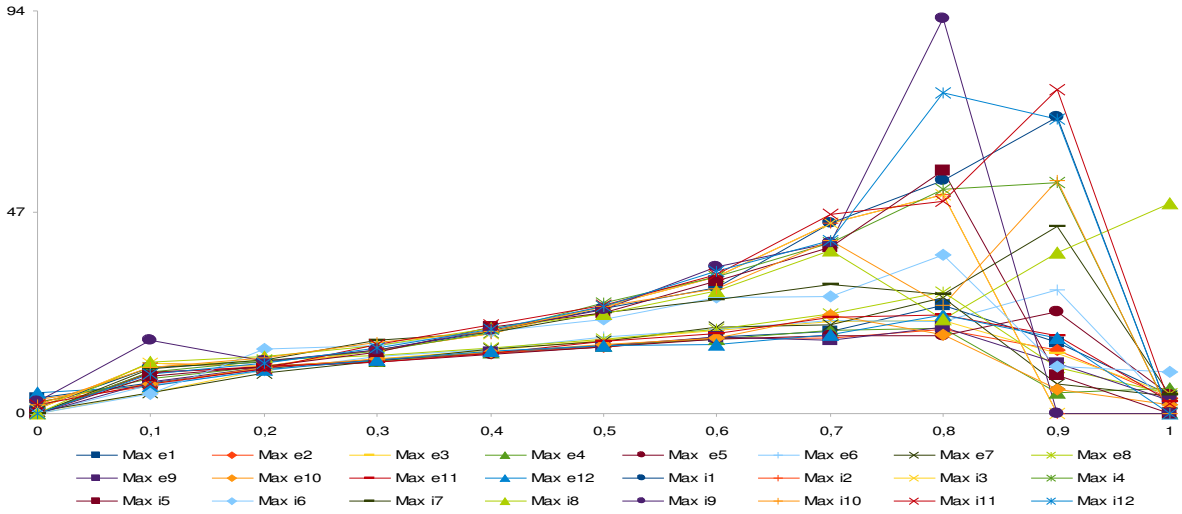
Zoom Rangos D.Melanogaster



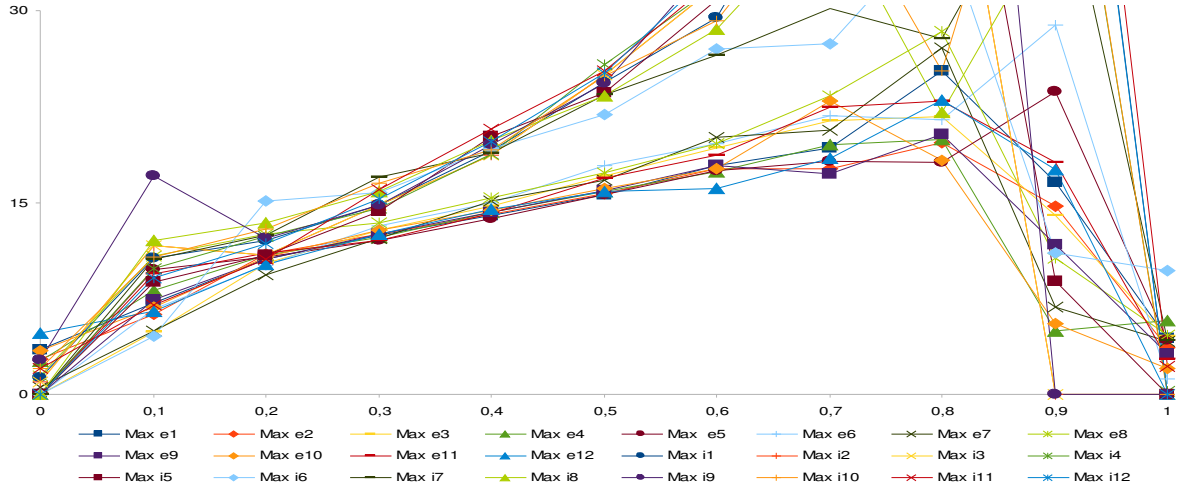
Longitud D.Melanogaster



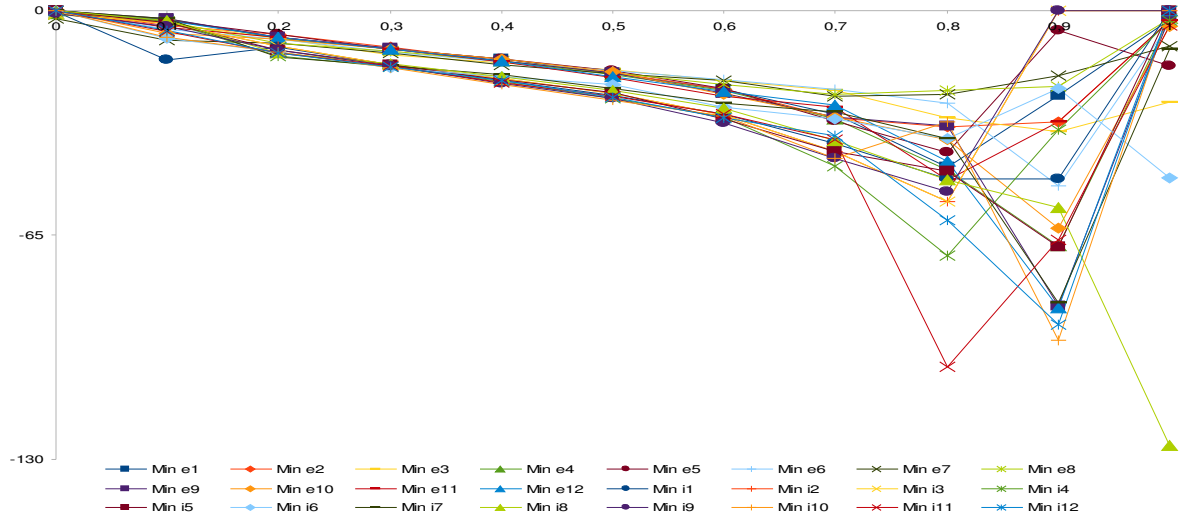
Zoom Longitud D.Melanogaster



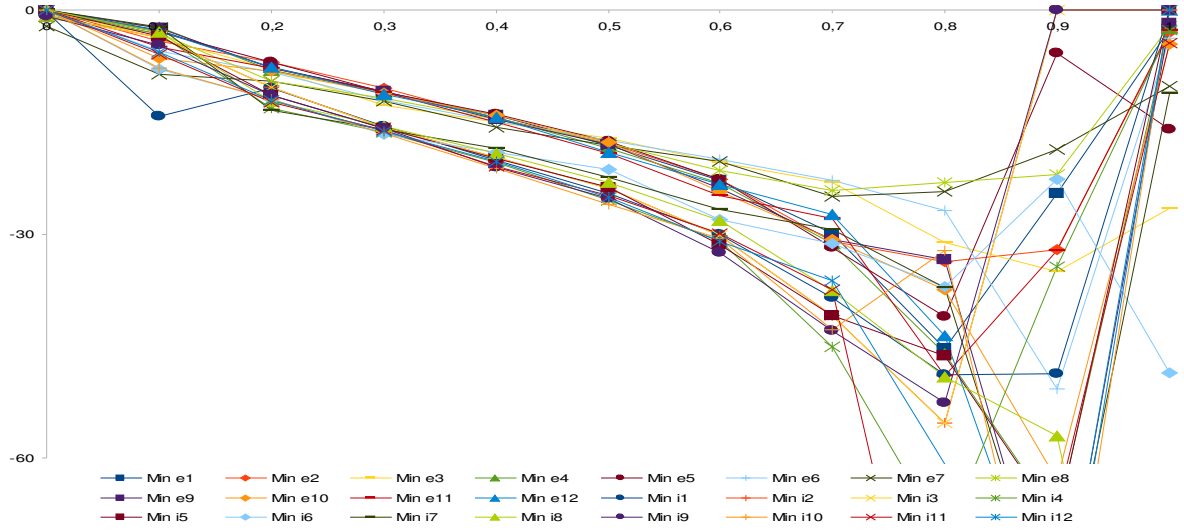
Máximos O.sativa



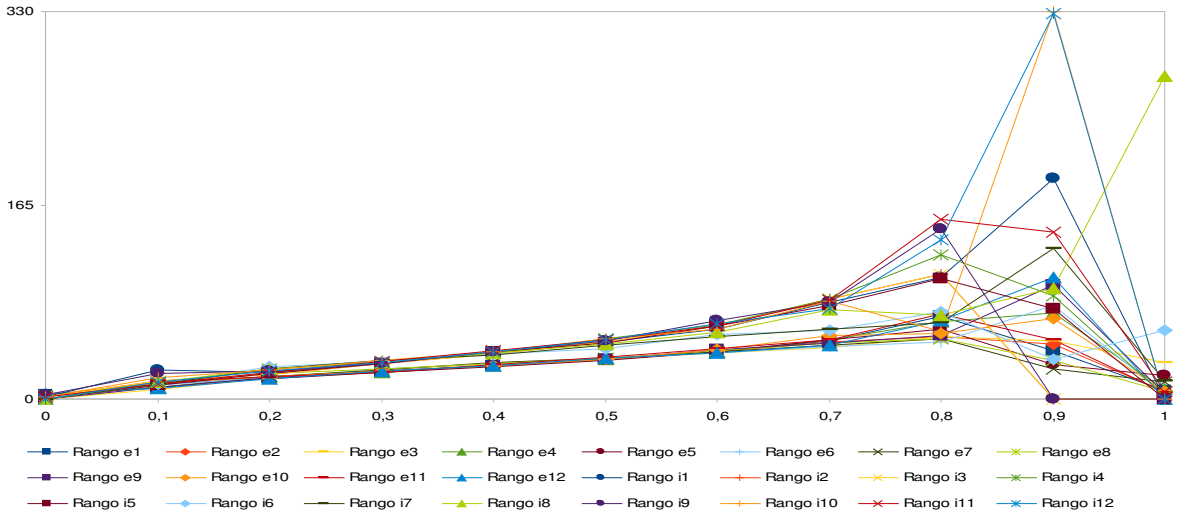
Zoom Máximos O.sativa



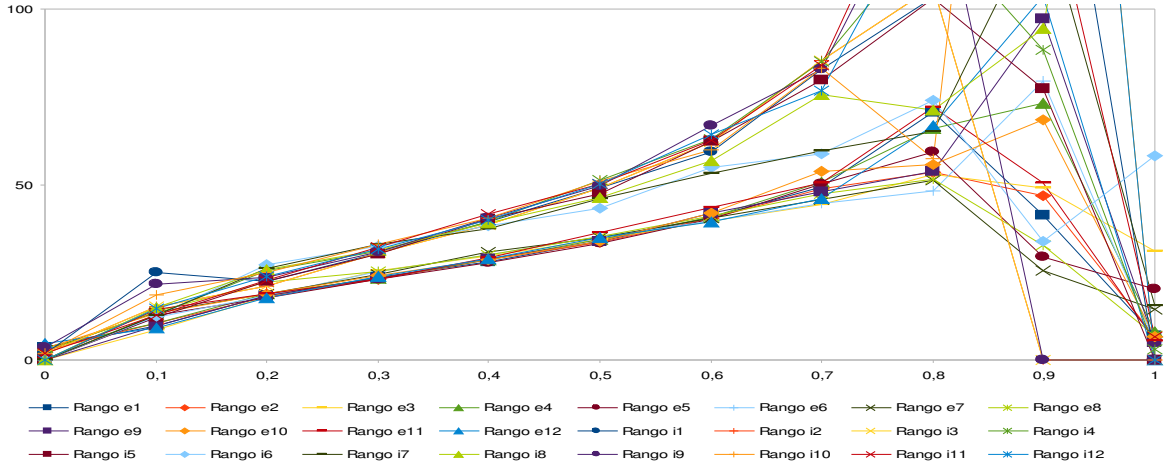
Mínimo O.sativa



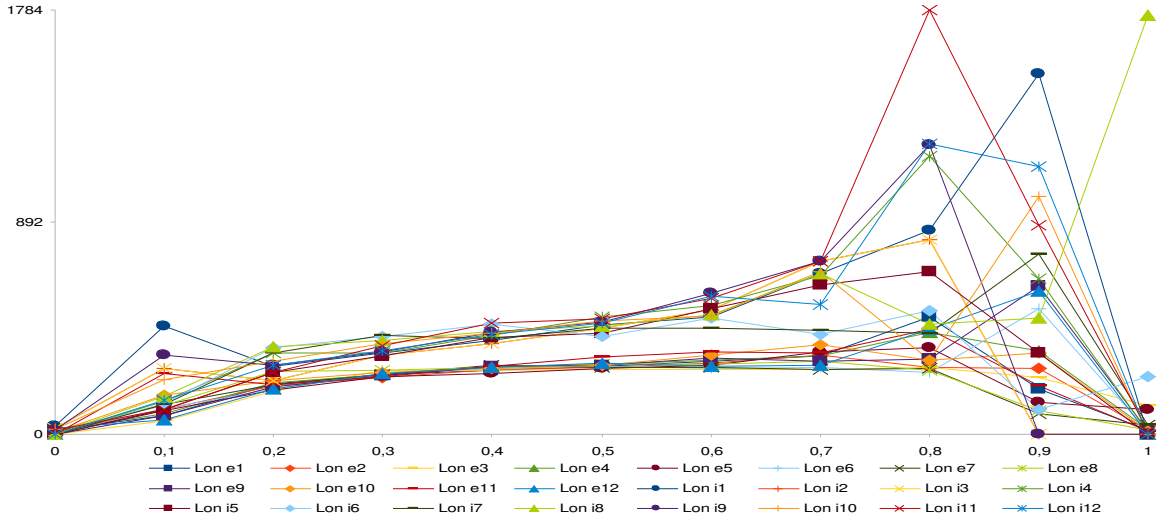
Zoom Mínimo O.sativa



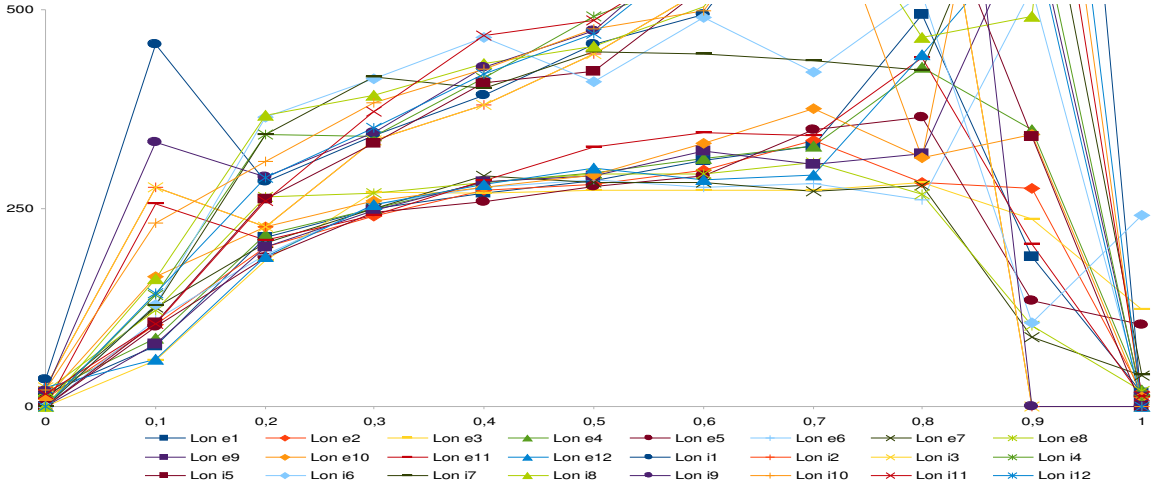
Rango O.sativa



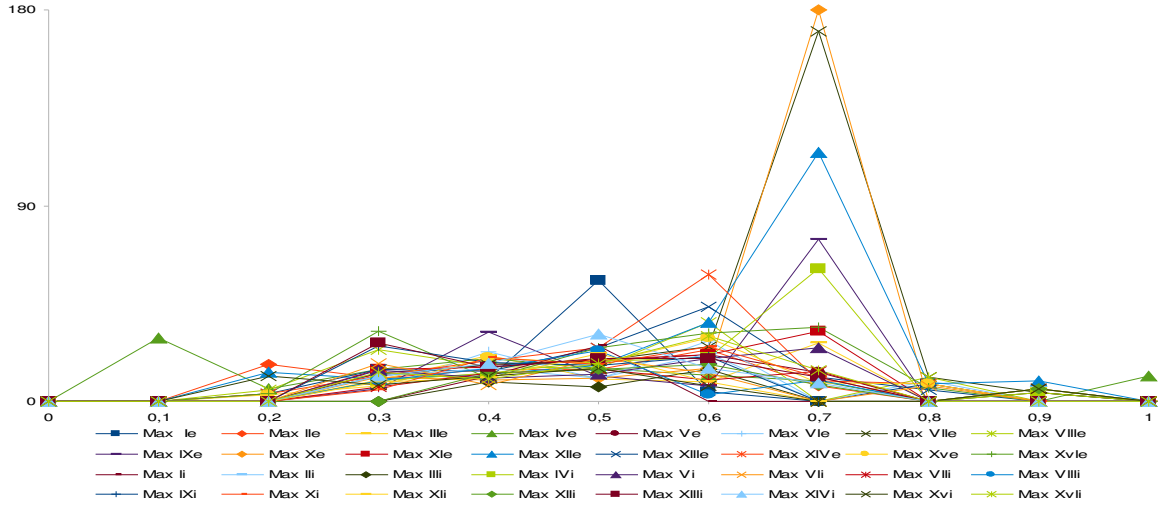
Zoom Rango O.sativa



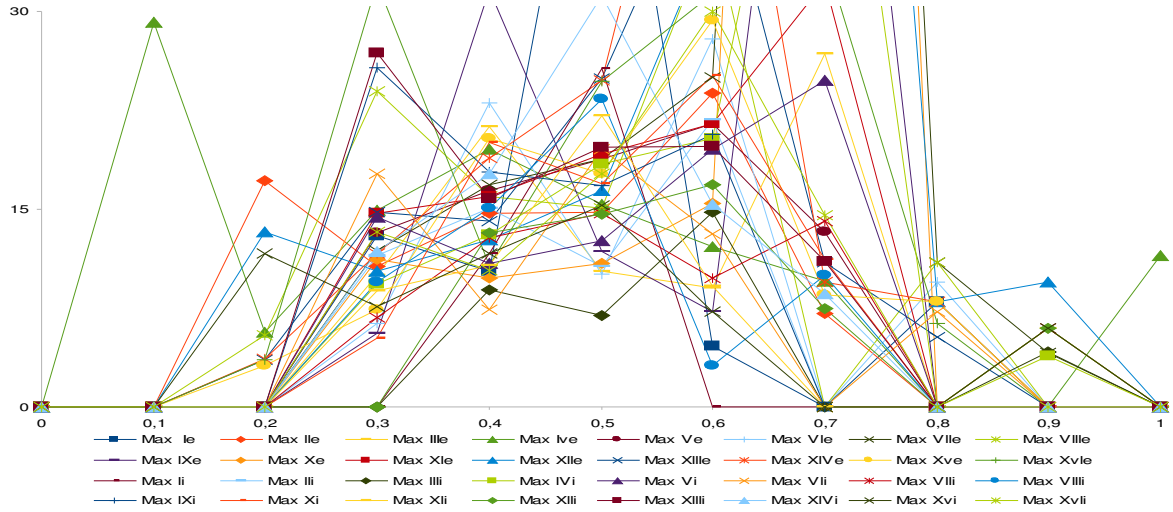
Longitud O.sativa



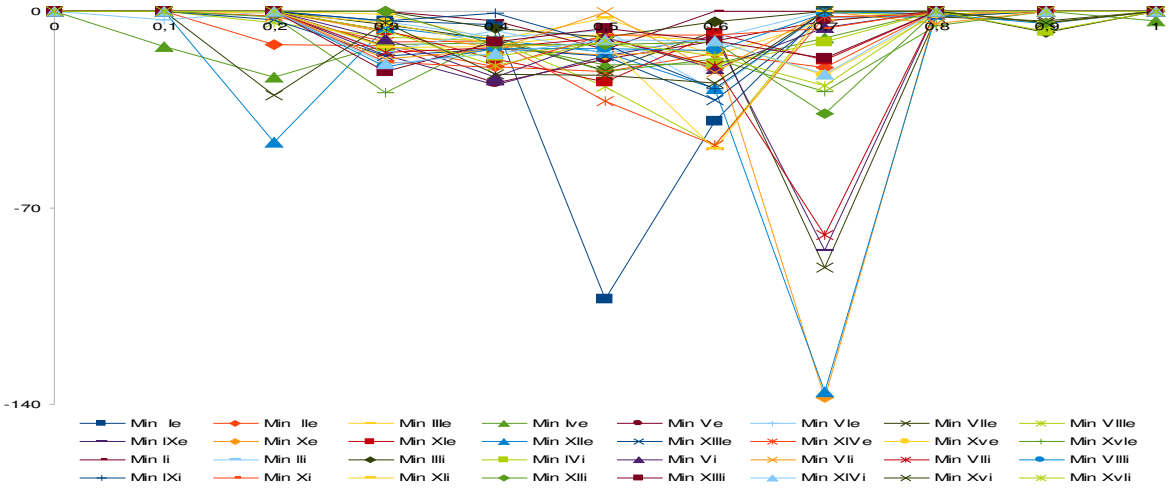
Zoom Longitud O.sativa



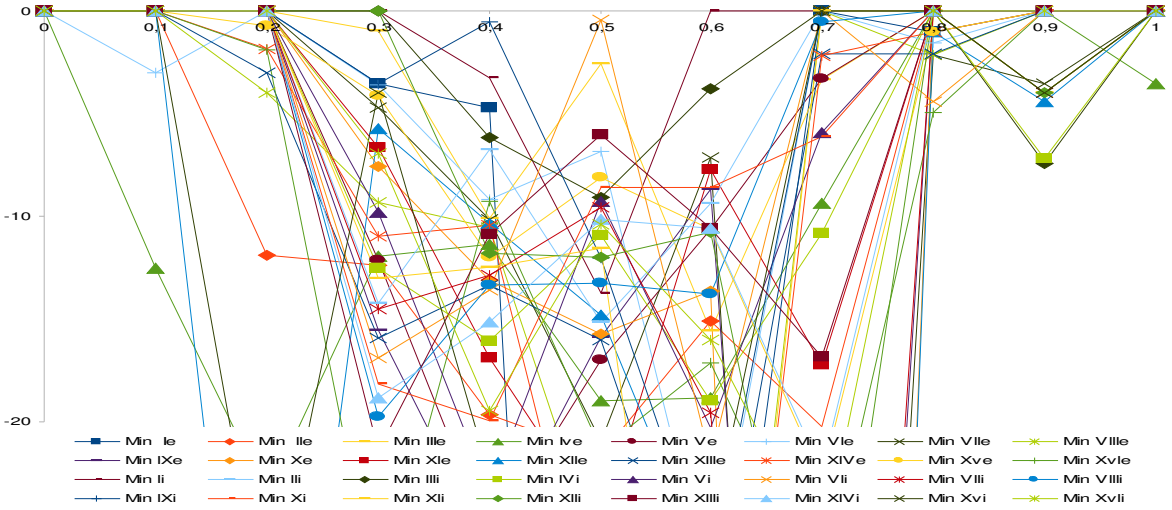
Máximo S.cerevisiae



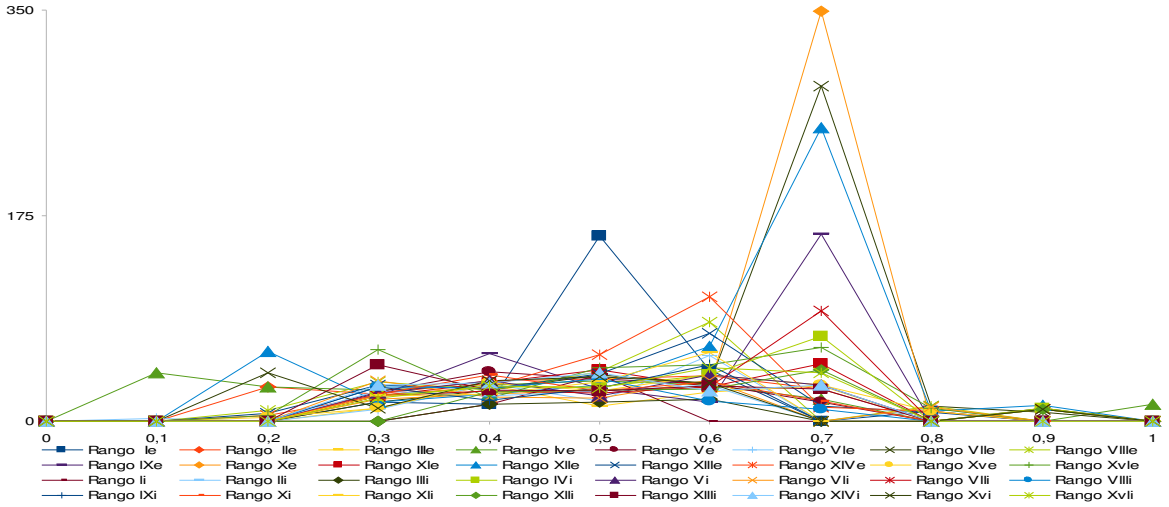
Zoom Máximo S.cerevisiae



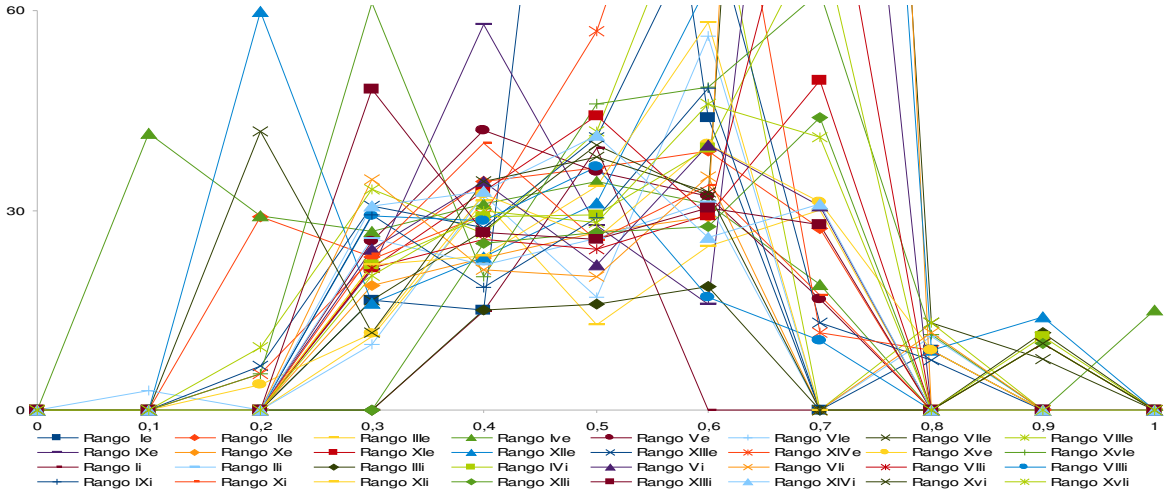
Mínimo S.cerevisiae



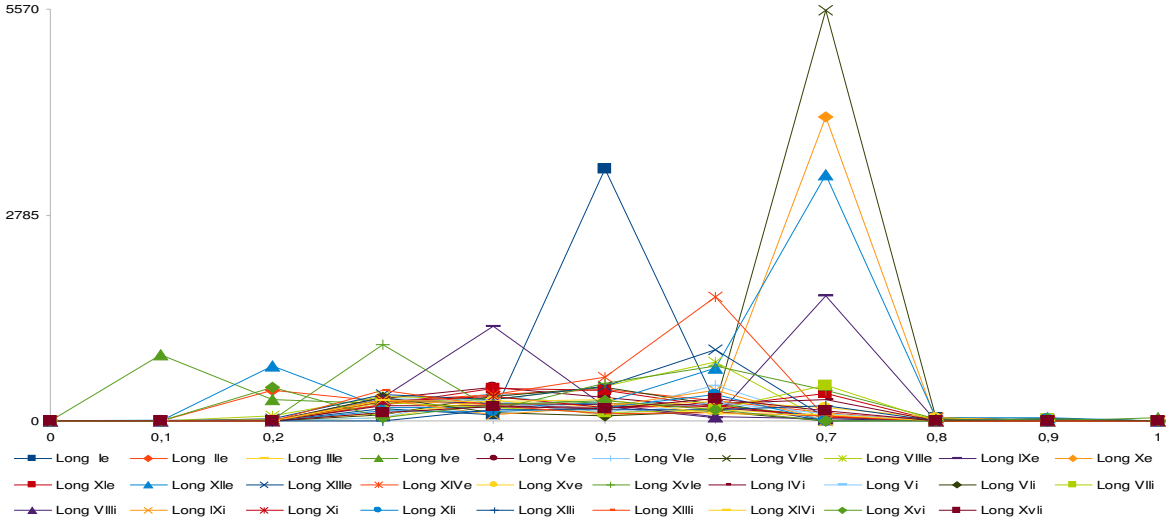
Zoom Mínimo S.cerevisiae



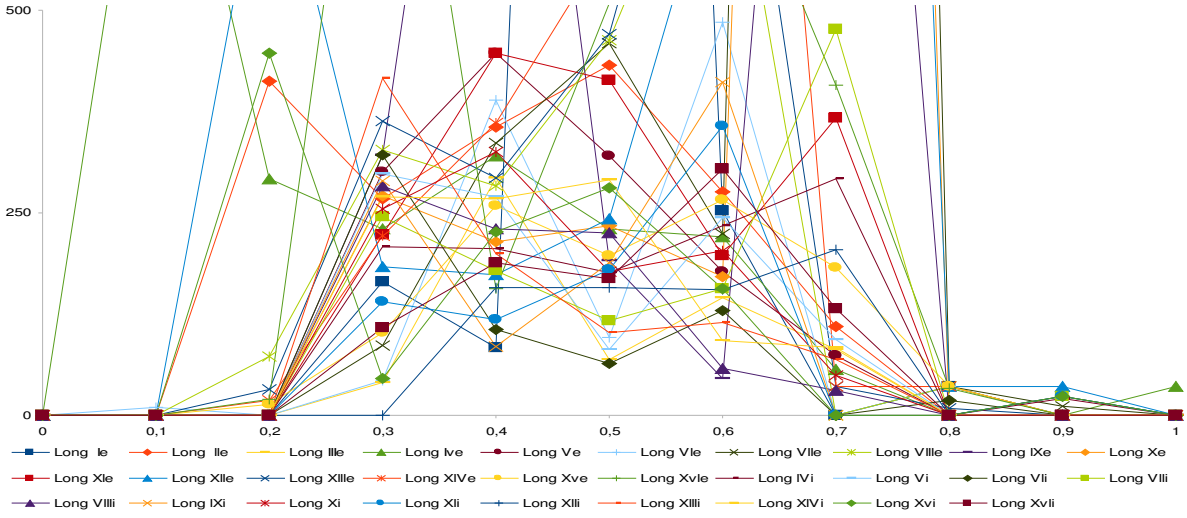
Rango S.cerevisiae



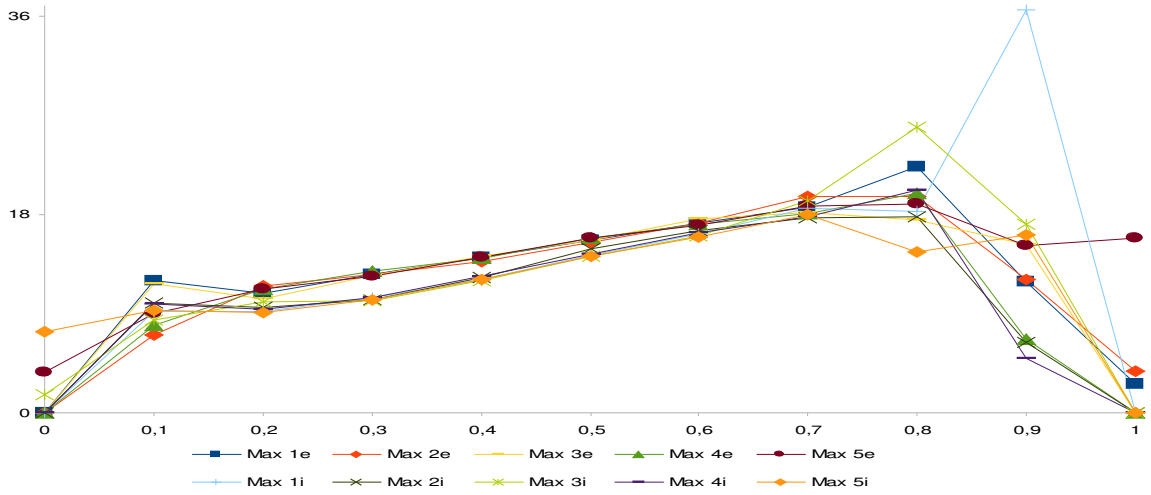
Zoom Rango S.cerevisiae



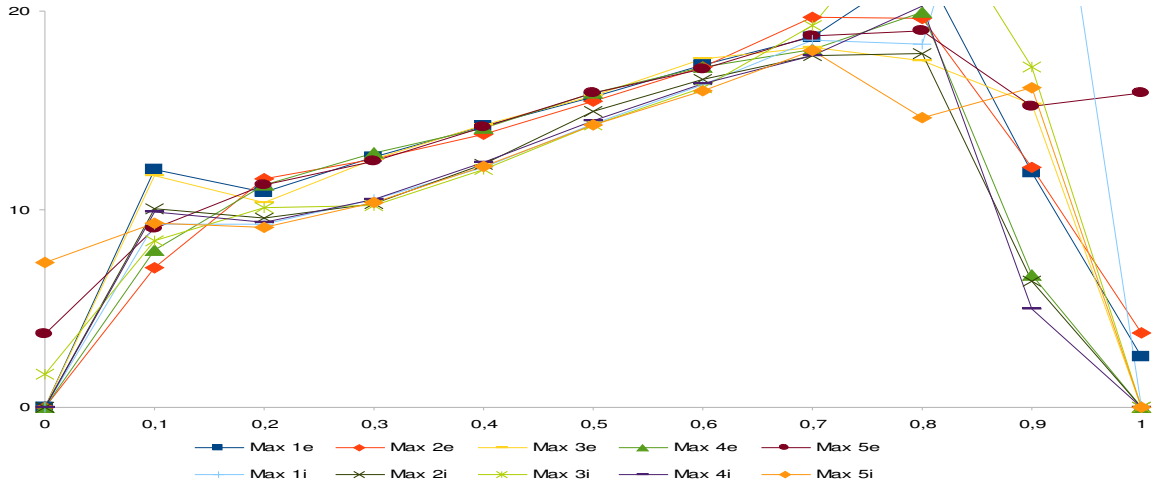
Longitud S.cerevisiae



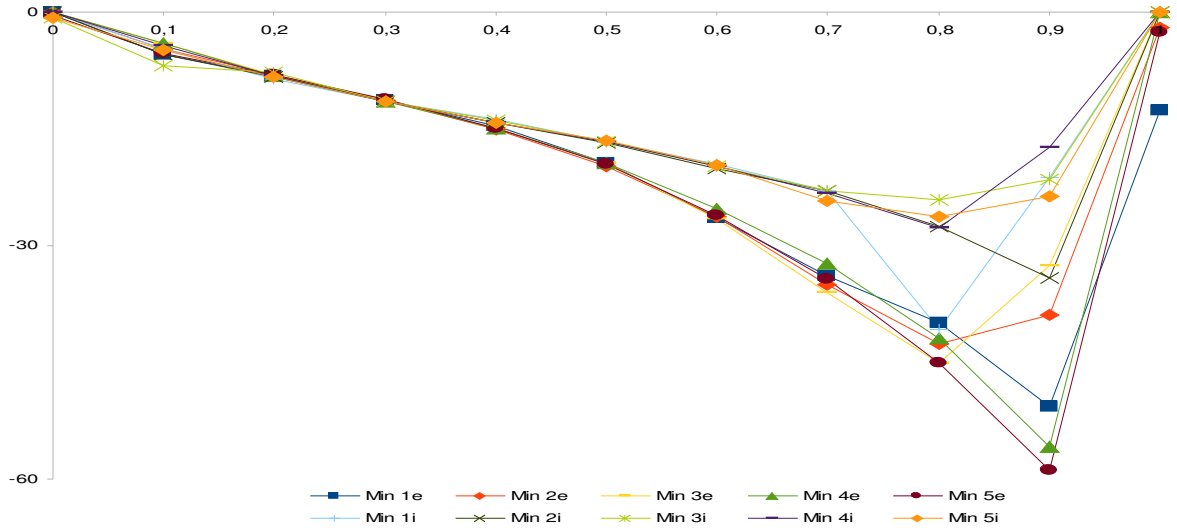
Zoom Longitud S.cerevisiae



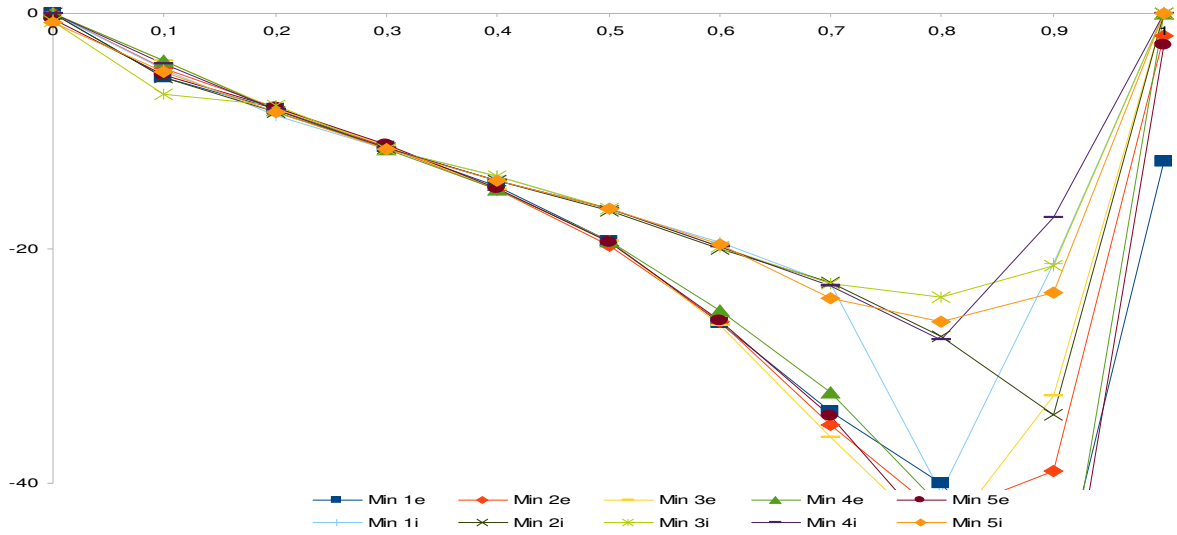
Máximo A.thaliana



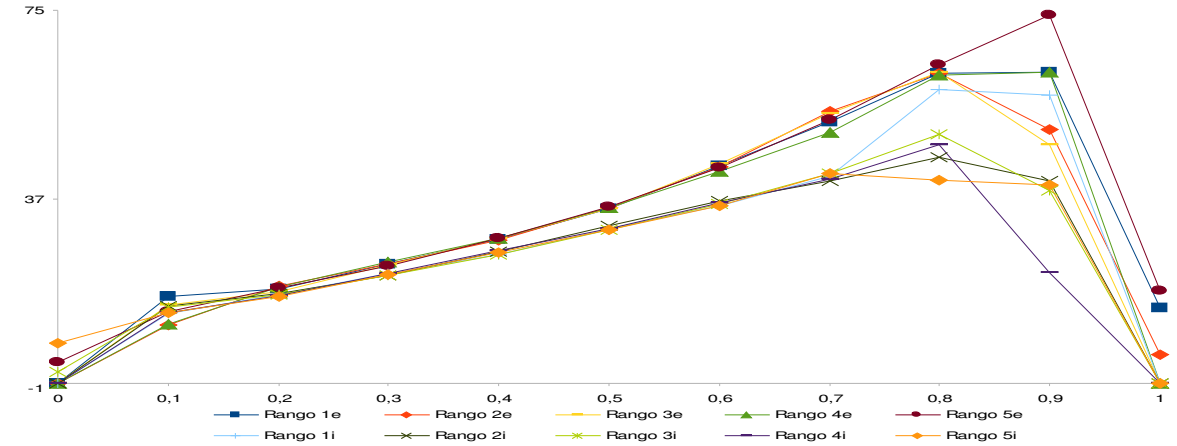
Zoom Máximo A.thaliana



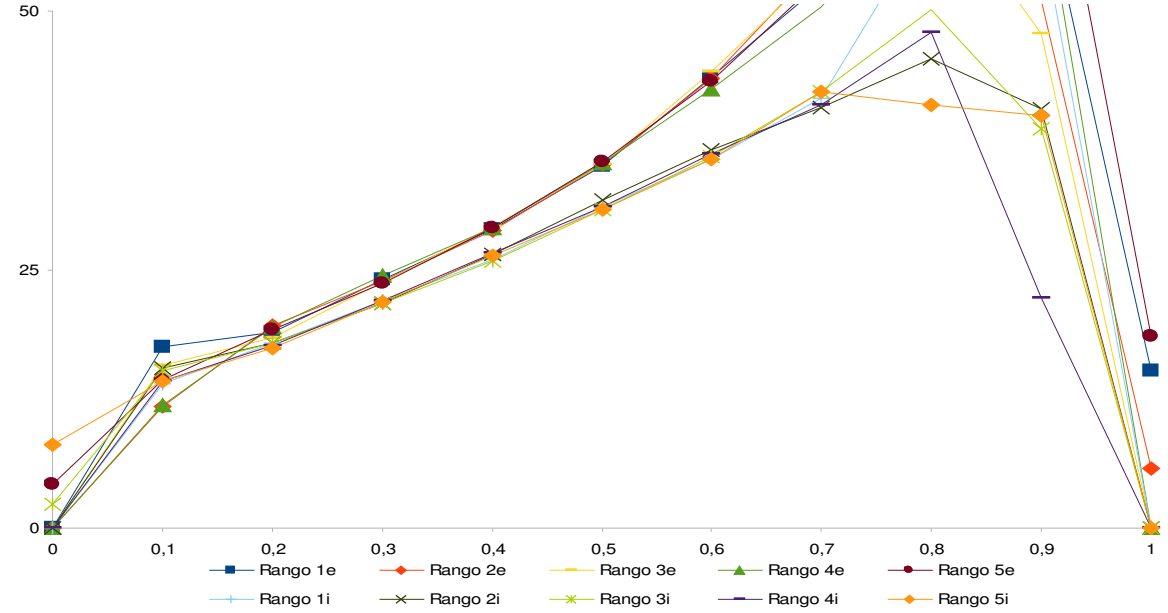
Mínimo A.thaliana



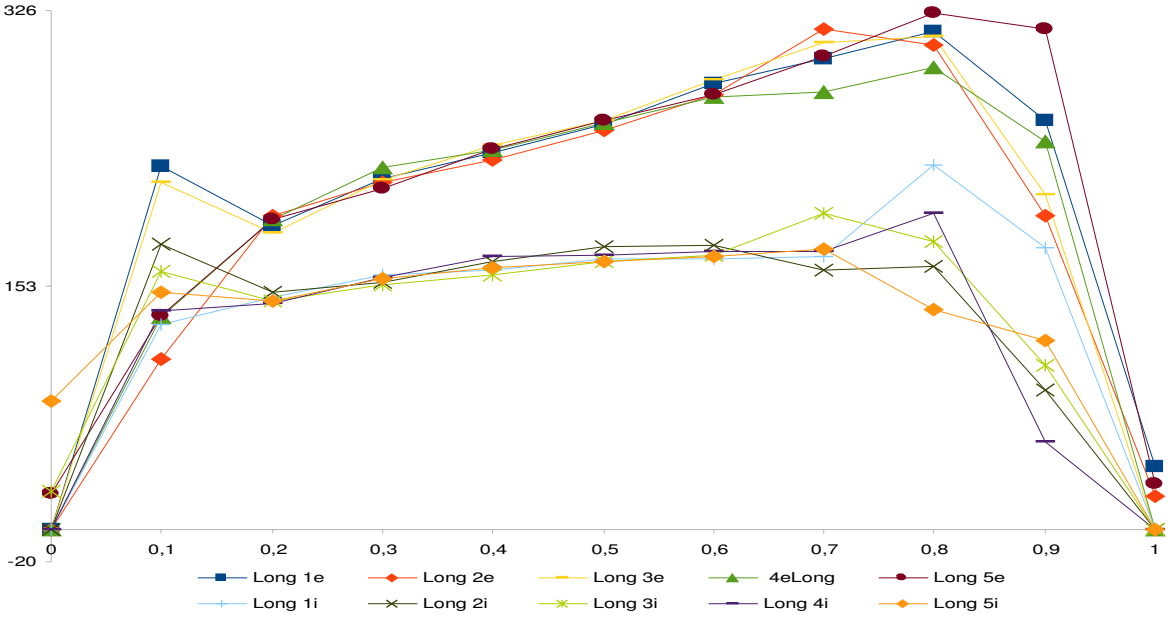
Zoom Mínimo A.thaliana



Rangos A.thaliana



Zoom Rangos A.thaliana



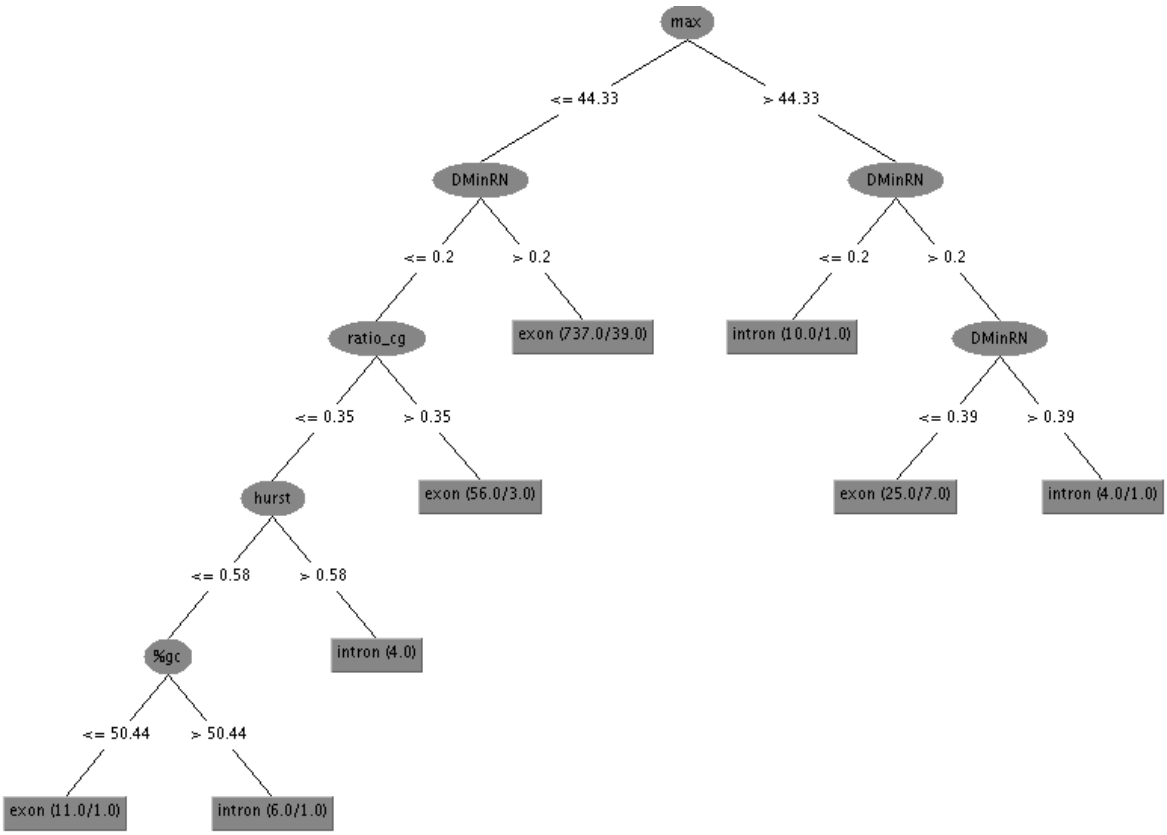
Longitud A.thaliana

Anexo L: Tabla guía de hexámeros para genomas eucariotes

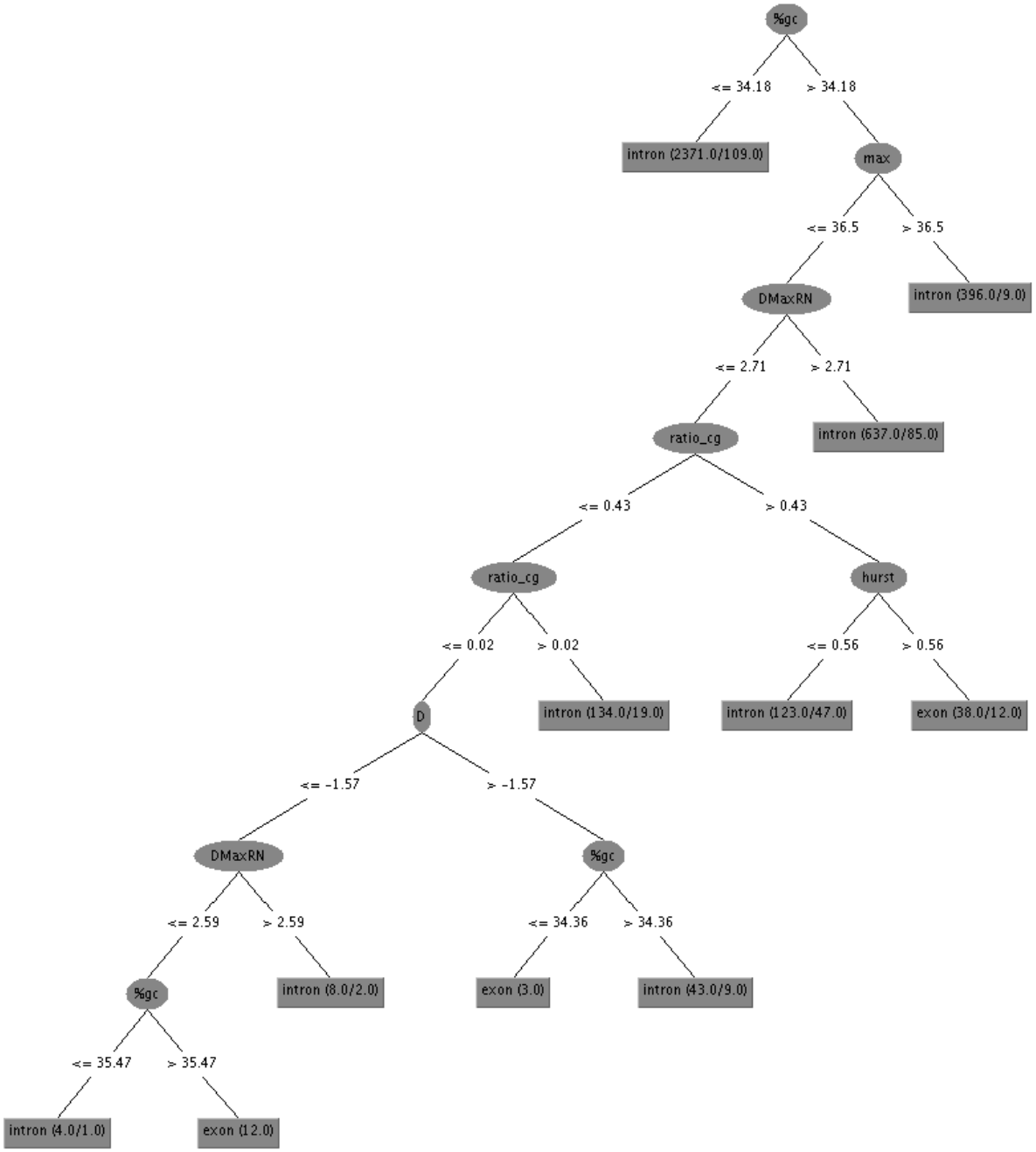
AAAAAA	AAAAAC	AAAAAT		**AAACAG**		AAAGGA	AAAGGC	AAAGGT	AAAGGG
AACAAA	AACAAC	AACAAT		**AACCAG**		AACGGA	AACGGC	AACGGT	AACGGG
AATAAA	AATAAC	AATAAT		**AATCAG**		AATGGA	AATGGC	AATGGT	AATGGG
AAGAAA	AAGAAC	AAGAAT		**AAGCAG**		AAGGGA	AAGGGC	AAGGGT	AAGGGG
ACAAAA	ACAAAC	ACAAAT		**ACACAG**		ACAGGA	ACAGGC	ACAGGT	ACAGGG
ACCAAA	ACCAAC	ACCAAT		**ACCCAG**		ACCGGA	ACCGGC	ACCGGT	ACCGGG
ACTAAA	ACTAAC	ACTAAT		**ACTCAG**		ACTGGA	ACTGGC	ACTGGT	ACTGGG
ACGAAA	ACGAAC	ACGAAT		**ACGCAG**		ACGGGA	ACGGGC	ACGGGT	ACGGGG
ATAAAA	ATAAAC	ATAAAT		**ATACAG**		ATAGGA	ATAGGC	ATAGGT	ATAGGG
ATCAAA	ATCAAC	ATCAAT		**ATCCAG**		ATCGGA	ATCGGC	ATCGGT	ATCGGG
ATGAAA	ATGAAC	ATGAAT		**ATGCAG**		ATGGGA	ATGGGC	ATGGGT	ATGGGG
AGAAAA	AGAAAC	AGAAAT		**AGACAG**		AGAGGA	AGAGGC	AGAGGT	AGAGGG
AGCAAA	AGCAAC	AGCAAT		**AGCCAG**		AGCGGA	AGCGGC	AGCGGT	AGCGGG
AGTAAA	AGTAAC	AGTAAT		**AGTCAG**		AGTGGA	AGTGGC	AGTGGT	AGTGGG
AGGAAA	AGGAAC	AGGAAT		**AGGCAG**		AGGGGA	AGGGGC	AGGGGT	AGGGGG
CAAAAA	CAAAAC	CAAAAT		**CAACAG**		CAAGGA	CAAGGC	CAAGGT	CAAGGG
CACAAA	CACAAC	CACAAT		**CACCAG**		CACGGA	CACGGC	CACGGT	CACGGG
CATAAA	CATAAC	CATAAT		**CATCAG**		CATGGA	CATGGC	CATGGT	CATGGG
CAGAAA	**CAGAAC**	**CAGAAT**		**CAGCAG**		**CAGGGA**	**CAGGGC**	**CAGGGT**	**CAGGGG**
CCAAAA	CCAAAC	CCAAAT		**CCACAG**		CCAGGA	CCAGGC	CCAGGT	CCAGGG
CCCAAA	CCCAAC	CCCAAT		**CCCCAG**		CCCGGA	CCCGGC	CCCGGT	CCCGGG
CCTAAA	CCTAAC	CCTAAT		**CCTCAG**		CCTGGA	CCTGGC	CCTGGT	CCTGGG
CCGAAA	CCGAAC	CCGAAT		**CCGCAG**		CCGGGA	CCGGGC	CCGGGT	CCGGGG
CTAAAA	CTAAAC	CTAAAT	...	**CTACAG**	...	CTAGGA	CTAGGC	CTAGGT	CTAGGG
CTCAAA	CTCAAC	CTCAAT		**CTCCAG**		CTCGGA	CTCGGC	CTCGGT	CTCGGG
CTTAAA	CTTAAC	CTTAAT		**CTTCAG**		CTTGGA	CTTGGC	CTTGGT	CTTGGG
CTGAAA	**CTGAAC**	**CTGAAT**		**CTGCAG**		**CTGGGA**	**CTGGGC**	**CTGGGT**	**CTGGGG**
CGAAAA	CGAAAC	CGAAAT		**CGACAG**		CGAGGA	CGAGGC	CGAGGT	CGAGGG
CGCAAA	CGCAAC	CGCAAT		**CGCCAG**		CGCGGA	CGCGGC	CGCGGT	CGCGGG
CGTAAA	CGTAAC	CGTAAT		**CGTCAG**		CGTGGA	CGTGGC	CGTGGT	CGTGGG
GCTAAA	GCTAAC	GCTAAT		**GCTCAG**		GCTGGA	GCTGGC	GCTGGT	GCTGGG
					. ..				
GCGAAA	GCGAAC	GCGAAT		**GCGCAG**		GCGGGA	GCGGGC	GCGGGT	GCGGGG
GTCAAA	GTCAAC	GTCAAT		**GTCCAG**		GTCGGA	GTCGGC	GTCGGT	GTCGGG
GTTAAA	GTTAAC	GTTAAT		**GTTCAG**		GTTGGA	GTTGGC	GTTGGT	GTTGGG
GTGAAA	GTGAAC	GTGAAT		**GTGCAG**		GTGGGA	GTGGGC	GTGGGT	GTGGGG
GGCAAA	GGCAAC	GGCAAT		**GGCCAG**		GGCGGA	GGCGGC	GGCGGT	GGCGGG
GGTAAA	GGTAAC	GGTAAT		**GGTCAG**		GGTGGA	GGTGGC	GGTGGT	GGTGGG
GGGAAA	GGGAAC	GGGAAT		**GGGCAG**		GGGGGA	GGGGGC	GGGGGT	GGGGGG

(Extracto) Organización de la Matriz 64x64 de hexámeros en genoma eucariote

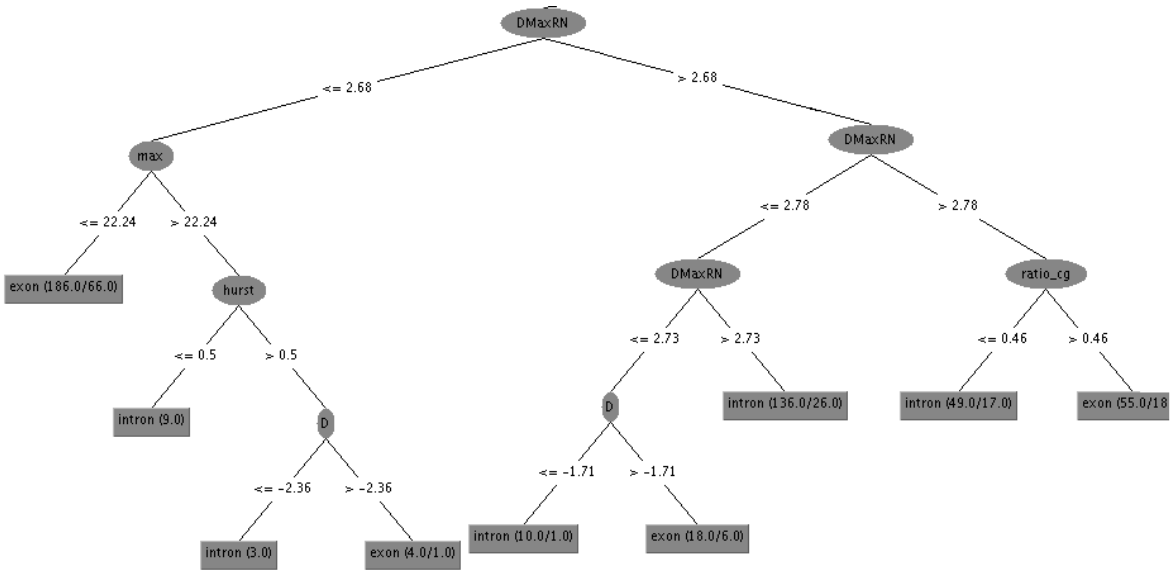
Anexo M: Ramas del árbol del modelo 13



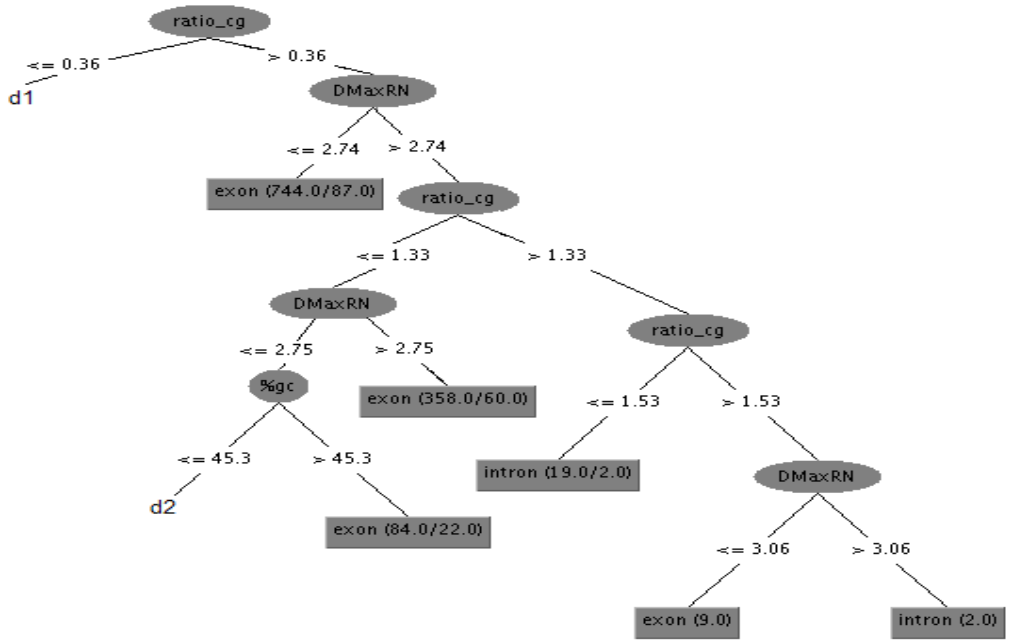
Gráfica 34: **Rama "a" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de Hurst (hurst), Hurst máximo (max)



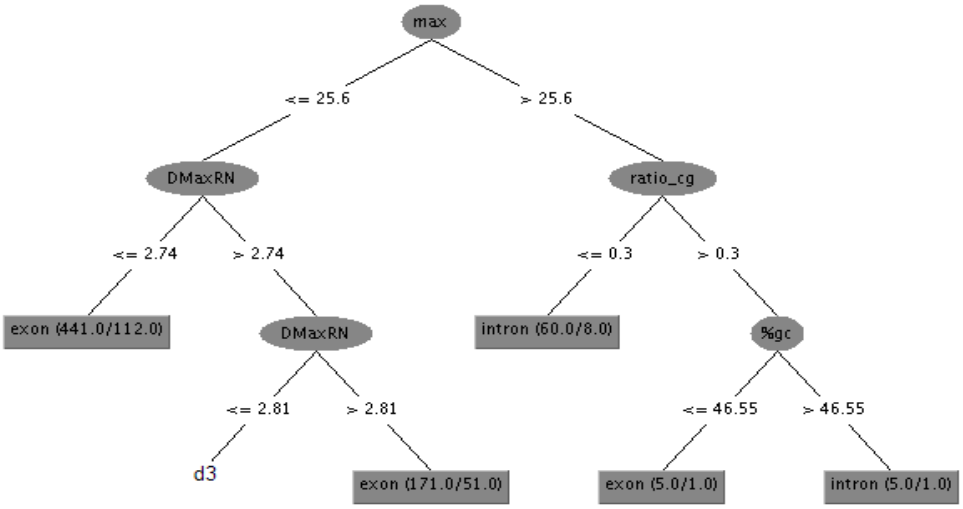
Gráfica 35: **Rama "b" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



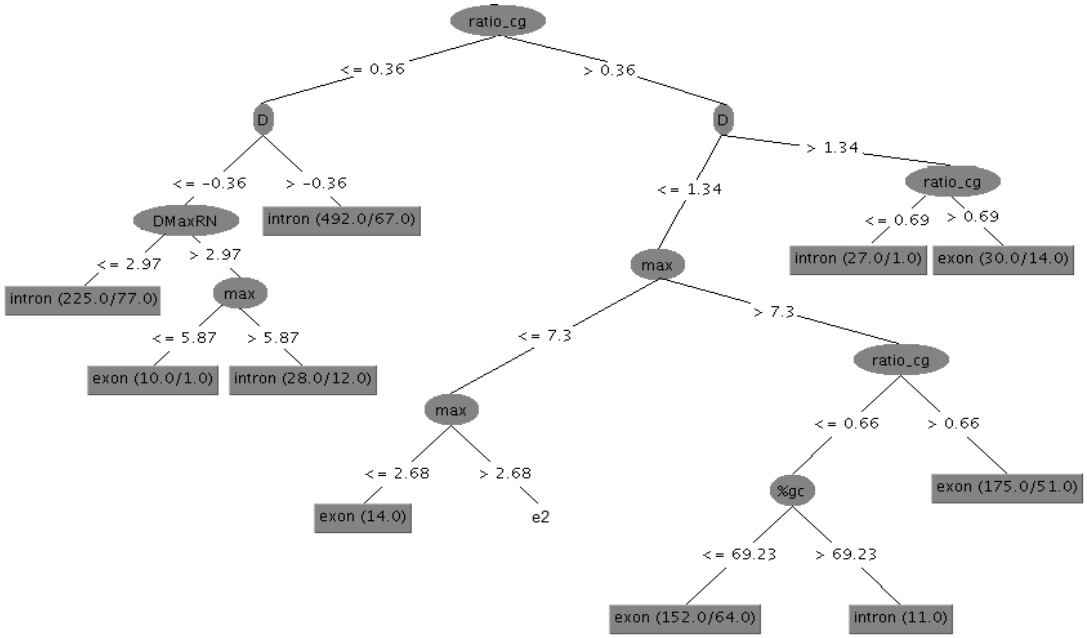
Gráfica 36: **Rama "c" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



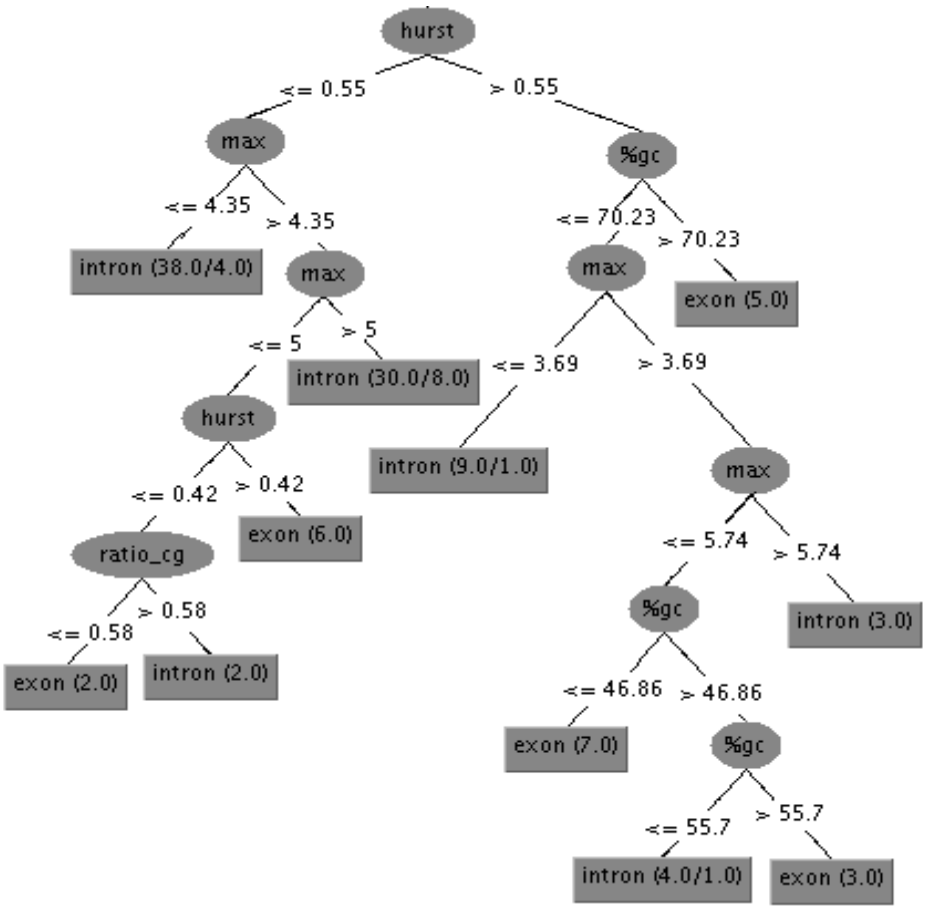
Gráfica 37: **Rama "c" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



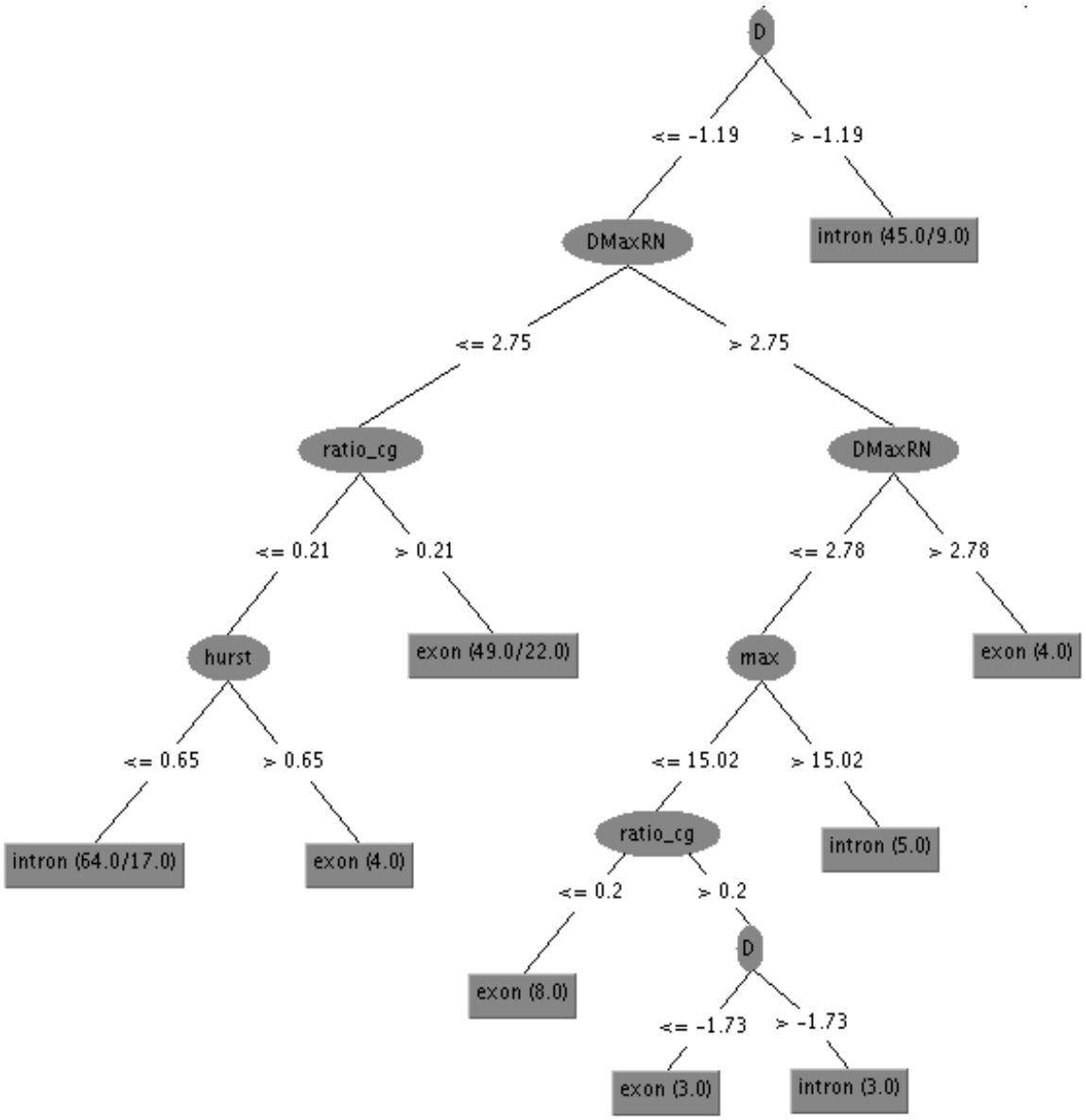
Gráfica 38: **Rama "d" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



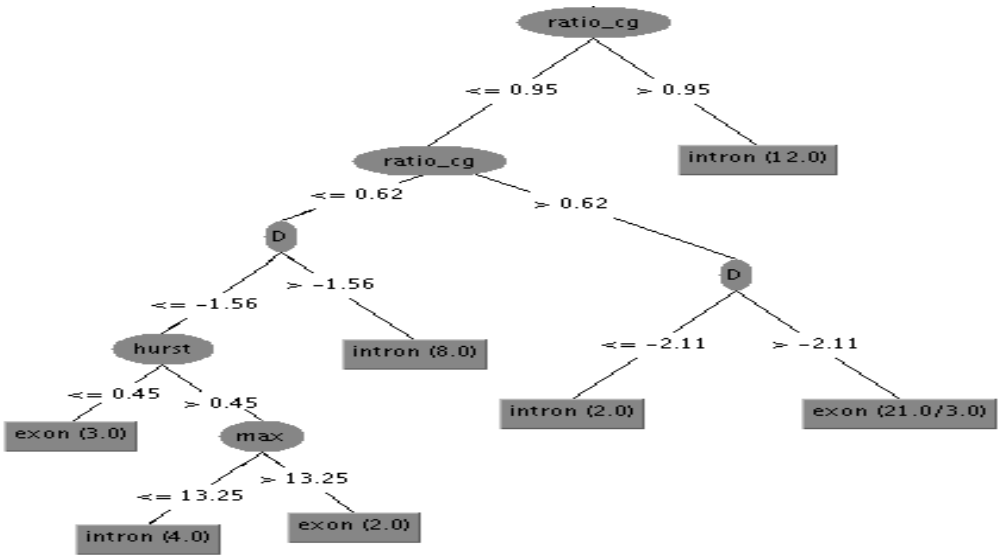
Gráfica 39: **Rama "e" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



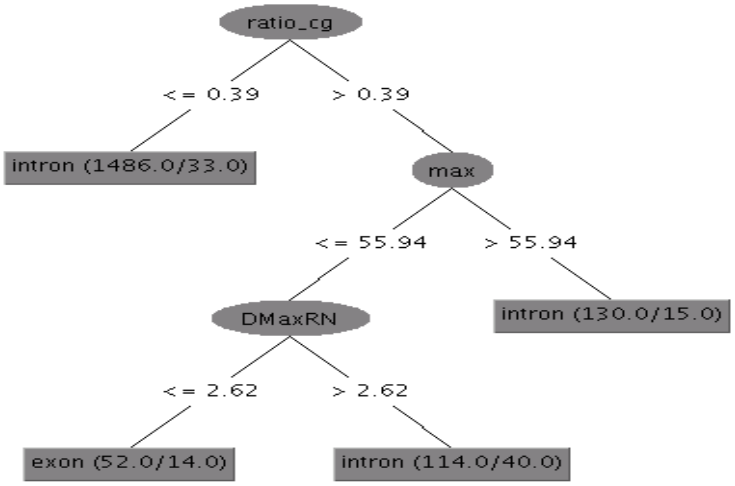
Gráfica 40: **Rama "e2" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



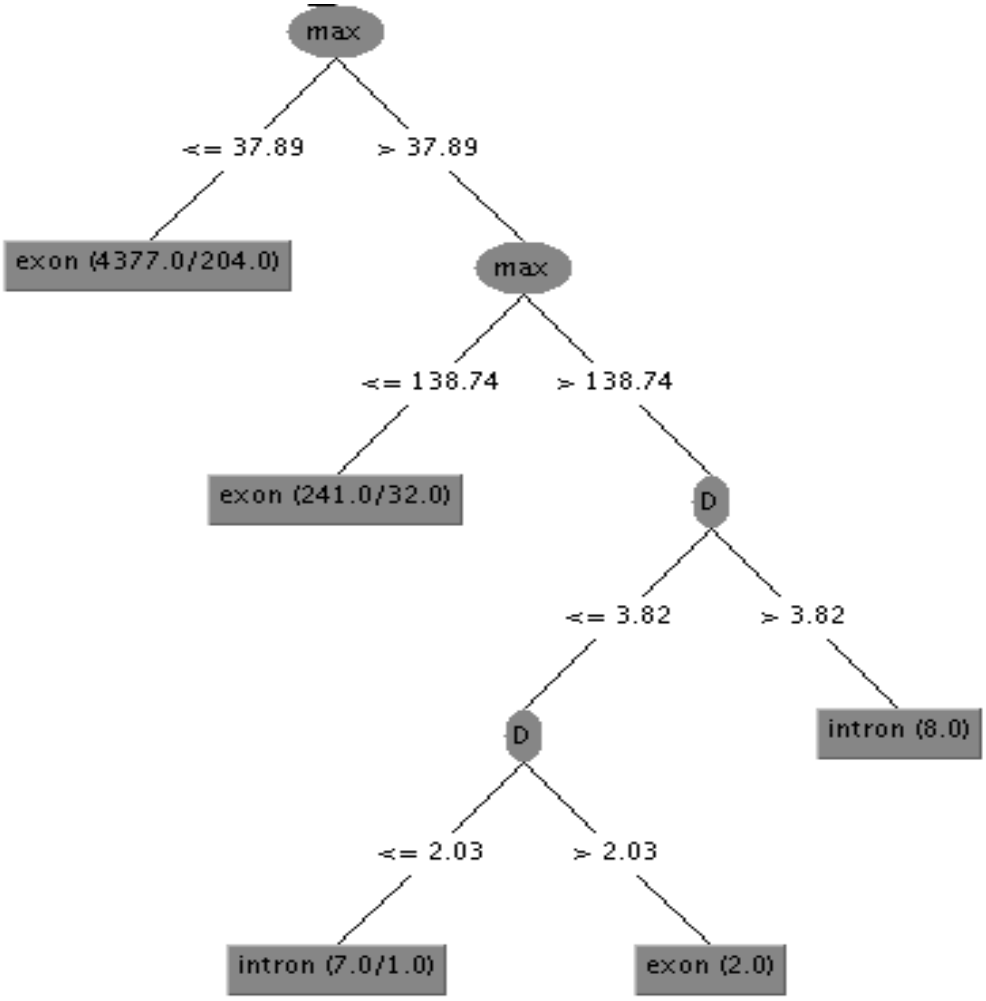
Gráfica 41: **Rama "d3" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



Gráfica 42: **Rama "d2" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de Hurst (hurst), Hurst máximo (max)



Gráfica 43: **Rama "f" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidades fractales de Hurst: exponente de hurst (hurst), Hurst máximo (max)



Gráfica 44: **Rama "g" del árbol** general de decisión del modelo 13 para secuencias de ADN en genomas eucariotes. Medidas estadísticas: porcentaje de CG (%gc), índice de CG (ratio_cg); medidas fractales bidimensionales: bidimensional fractal (D), bidimensional mínimo (DMinRN), bidimensional máximo (DmaxRN), ; medidas fractales de Hurst: exponente de Hurst (hurst), Hurst máximo (max)

Referencias auxiliares: van todas unidas pero sin paréntesis

p 1/15/68 (Wang et. al., 2004) (Mathé et. al., 2002) (Zhang, 2002)

p 1/15/68 (Majoros et. al., 2004) (Burge y Karlin, 1996)(Salzberg et. al., 1999)(Xu et. al., 1994)

p 15/29 /72 (Gao et. al., 2004) (Gao et. al., 2005)

p 21/ / (Hamori y Ruskin, 1983)

p 18/32 64 (Zipf, 1949) (Zipf, 1932)

p 26/ (Majoros et. al., 2004) (Majoros et. al., 2005)