Universidad Central "Marta Abreu" de Las Villas

Facultad de Matemática, Física y Computación

Trabajo de Diploma

Análisis comparativo de secuencias genómicas mediante la aplicación de la Transformada de Fourier sobre campos finitos

Autor:

Pavel Silveira Díaz

Tutores:

Dr. Robersy Sánchez Rodríguez

Dr. Ricardo Grau Ábalo

Santa Clara, Villa Clara
2008

DICTAMEN

Hago constar que el presente trabajo fue realizado en la Universidad Central “Martha Abreu” de Las Villas como parte de la culminación de los estudios de la especialidad de Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____
Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____
Firma del tutor

_____
Firma del Jefe del Seminario

RESUMEN

En esta investigación se desarrolla una nueva herramienta para análisis comparativo de secuencias genómicas. La aplicación de la transformada de Fourier sobre campos finitos (MFT) permite una recodificación en $\mathbb{Z}_p$ de las secuencias de ADN genómico representadas en $\mathbb{Z}_5$. El método se basa en la comparación de los espectros de potencia de las secuencias recodificadas mediante el uso del ANOVA bifactorial no-paramétrico para la detección de diferencias estadísticamente significativas entre los picos espectrales inducidos por la MFT. Se muestra la utilidad de la aplicación del método en el análisis filogenético de las secuencias de ADN alineadas, en la comparación directa e indirecta de los espectros de potencia. Las comparaciones indirectas de las secuencias de ADN posibilitan la inclusión, en el análisis, de secuencias aleatorias generadas con diferentes distribuciones de bases nitrogenadas. La aplicación de esta herramienta permite, incluso, la detección de la existencia de diferencias estadísticamente significativas en las arquitecturas organizativas de secuencias genómicas de especies filogenéticamente cercanas.

ABSTRACT

In this research a new tool for the comparative genomic sequence analysis is developed. The application of the Fourier's transform over finite fields (MFT) allows a recodification in $\mathbb{Z}_p$ of the DNA genomic sequences represented in $\mathbb{Z}_5$. The method is based on the comparison of the power spectra of the recodified sequences by means of the use of the nonparametric bifactorial ANOVA for the detection of the statistically significant differences between the spectral peaks induced by the MFT. There appears the usefulness of the method application in the phylogenetic analysis of the aligned DNA sequences, in the direct and indirect comparison of the power spectra. The indirect comparisons of the DNA sequences make possible the incorporation, in the analysis, of random sequences generated with different distributions of nitrogenous bases. The application of this tool allows, even, the detection of the existence of statistically significant differences in the organizational architectures of genomic sequences of phylogenetically close species.

TABLA DE CONTENIDOS

INTRODUCCIÓN

En la actualidad, la comunidad científica está fijándose nuevos objetivos, que comprenden la utilización de la gran cantidad de información que ha sido generada hasta el momento para su empleo en los análisis estructurales y funcionales de los genomas. Ahora con la secuenciación completa del genoma humano y de otras especies puestas al alcance de los investigadores, el análisis detallado de las secuencias genómicas indudablemente mejorará el entendimiento de los sistemas biológicos, lo cual requiere de sofisticadas herramientas bioinformáticas y computacionales. De esta forma estamos asistiendo al paso de la "Era Genómica" a una "Era post-Genómica" en el cual, se va a analizar y comparar los genomas y cuales son las relaciones existentes entre su estructura y su función, valiéndose de las herramientas bioinformáticas desarrolladas para este propósito (Yu et al., 2004)

Un aspecto clave dentro de las nuevas tendencias post-genómicas es la Genómica Comparativa, que analiza similitudes y diferencias de los genomas de distintos organismos, tanto estructurales como funcionales (Yu et al., 2004, Cliften, 2004). La disponibilidad de muchos genomas completos ofrece la oportunidad de realizar la comparación de genomas de especies filogenéticamente relacionadas y distantes. Esta comparación no resulta trivial pues, por ejemplo, el genoma humano y el del chimpancé son comparables en el 96% de su longitud siendo estas regiones idénticas en un 99% (Consortium and genome, 2005). El número de diferencias genéticas entre los humanos y los chimpancés es aproximadamente 60 veces menor que entre los humanos y los ratones y unas 10 veces menos que entre los ratones y las ratas. Al mismo tiempo, la cantidad de disparidades genéticas entre un hombre y un chimpancé es unas 10 veces más que entre dos personas cualesquiera. De hecho, ambos genomas son idénticos casi un 99 % si no se tienen en cuenta en el análisis ciertos aspectos del ADN que se han reorganizado de forma distinta en las dos especies. Pero si se consideran las sustituciones de nucleótidos o bases, que difieren en el 1,23%, y otras variaciones como las repeticiones que ocurren en casi el 3%, las similitudes entre las secuencias de ADN de ambas

especies sólo llegan al 96%. Luego, se origina la siguiente pregunta: ¿Si las similitudes entre el genoma del chimpancé y el humano son tan grandes entonces por qué desde el punto de vista biológico somos tan diferentes? Desde el punto de vista biológico se buscan las diferencias en el análisis comparativo de pequeñas regiones de genes codificantes para proteínas vinculadas a funciones especiales del sistema nervioso y a enfermedades (Pennisi, 2007). Evidentemente deben existir grandes diferencias en la arquitectura organizativa de estos genomas, las cuales no son perceptibles utilizando las herramientas actuales de la Bioinformática.

Antecedentes y actualidad del tema

La aplicación de la transformada discreta de Fourier ha sido ampliamente empleada en el análisis de periodicidades de secuencias de ADN genómicas en el campo de la Bioinformática. Pueden citarse algunos de los trabajos pioneros y más significativos (Fickett, 1982, Tiwari S et al., 1997, Tsonis et al., 1991) dedicados a la identificación de genes en secuencias genómicas; en esta dirección, incluso, han sido desarrollados varios software.

Desde hace unos años en el Centro de Estudios de Electrónica y Tecnologías de la Información (CEETI) se han realizado investigaciones sobre análisis de secuencias consideradas como señales digitales, entre los artículos que exponen dichos resultados pueden mencionarse (Fuentes et al., 2006a, Fuentes et al., 2006b, Fuentes et al., 2007).

En un trabajo reciente realizado por miembros del Laboratorio de Bioinformática se propone por primera vez la aplicación de la transformada discreta de Fourier en estrecha conexión con las estructuras algebraicas sobre campos finitos (Sanchez and Grau, 2008). En particular, la definición del campo de Galois de 5 elementos sobre el conjunto de bases nitrogenadas extendidas (Sanchez and Grau, 2008, Sanchez et al., 2006) ha contribuido a la formalización matemática de las secuencias genómicas y ha abierto nuevos puntos de vistas en la aplicación de la transformada discreta de Fourier.

Planteamiento del problema

No hemos encontrado en la literatura disponible ninguna aplicación de la transformada discreta de Fourier dentro del campo de la genómica comparativa. En particular, no encontramos referencias del uso de la transformada modular de Fourier en el análisis de señales de tipo alguno. Luego, podemos plantearnos:

¿En qué medida la detección de nuevas periodicidades en las secuencias genómicas mediante aplicación consecutiva de las transformadas modular (sobre campos de Galois) y discreta de Fourier permiten estimar diferencias estadísticamente significativas en la arquitectura organizativa de dos secuencias genómicas?

Hipótesis de trabajo

Teniendo en cuenta los elementos teóricos antes expuestos, así como las interrogantes existentes, se plantea la siguiente hipótesis general de investigación:

La aplicación consecutiva de las transformadas modular y discreta de Fourier induce nuevas periodicidades en las secuencias genómicas las cuales permiten la detección de diferencias estadísticamente significativas en la arquitectura organizativa de las secuencias genómicas que se comparan.

Objetivos del trabajo

Este trabajo se propone como objetivos:

- Implementar y aplicar el algoritmo FFT sobre campos finitos para su uso en la recodificación de secuencias genómicas.
- Desarrollar un método que permita el análisis comparativo de las secuencias genómicas recodificadas.
- Investigar las posibilidades de aplicación en el análisis filogenético.

Contribución del trabajo

Este trabajo crea una nueva perspectiva en la aplicación del análisis de señales genómicas en Bioinformática. En particular, la recodificación de las señales genómicas empleando la transformada modular de Fourier con el objetivo de inducir nuevas periodicidades que dependen de la naturaleza de las secuencias resulta novedosa, ya que no se han encontrado antecedentes en la literatura disponible. Por ello, los resultados obtenidos amplían los resultados previamente obtenidos en el grupo de Bioinformática y sienta las bases para nuevos desarrollos de las investigaciones en dicho grupo.

Estructura del trabajo

El trabajo se estructura esencialmente en tres capítulos.

- El capítulo 1 está dedicado, en su primera parte, a la descripción de los fundamentos biológicos y las herramientas matemático computacionales utilizadas así como al

análisis de los principales aspectos del estado del arte desde el punto de vista de la aplicación del análisis de señales en este campo. Se presenta, además, un breve sumario de la transformada modular de Fourier.

- En el capítulo 2 se describe la implementación de la transformada finita de Fourier y de los procedimientos seguidos para establecer el método de comparación de los espectros.
- En el capítulo 3 se presentan y discuten los resultados obtenidos en la aplicación del método desarrollado al análisis comparativo de secuencias genómicas. En particular, se discuten diferentes situaciones particulares.

1. FUNDAMENTOS BIOLÓGICOS, MATEMÁTICOS Y COMPUTACIONALES

En este capítulo se realiza una descripción de las bases teóricas que conducen las aplicaciones de las transformadas modular y discreta de Fourier. Dado el hecho de que las aplicaciones son realizadas en secuencias de ADN genómico y que, además, a lo largo de este trabajo se aplican una variedad de herramientas matemáticas y computacionales, este capítulo se separa en seis secciones que permiten introducir al lector en la temática abordada.

1.1. Introducción a los términos biológicos

Se presentan en esta sección algunas ideas básicas de biología molecular y algunos resultados anteriores relacionados con este trabajo. Ellos son imprescindibles para la comprensión de las hipótesis de investigación y los capítulos siguientes.

1.1.1. Las moléculas básicas de la vida

Los principales actores de la Biología Molecular son las moléculas del ADN, ARN y las proteínas. Estas moléculas son simultáneamente los principales actores de un poderoso sistema de comunicación, el sistema de información genética, entrelazados por el código de comunicación más maravilloso conocido por el hombre, el Código Genético.

La expresión de la información almacenada en el Ácido desoxirribonucleico (ADN) se produce a través de la trascripción lineal de la secuencia de nucleótidos en la secuencia de nucleótidos del Ácido ribonucleico mensajero (ARNm). La secuencia de nucleótidos del ARNm es seguidamente traducida en una secuencia lineal de aminoácidos, de manera que el flujo de información es ADN$\rightarrow$ARNm $\rightarrow$Proteína.

1.1.2. El ADN y ARNm

El ADN está compuesto por cuatro moléculas básicas llamadas nucleótidos las cuales solo se diferencian en que cada una posee una base nitrogenada diferente. Cada nucleótido contiene un fosfato, un azúcar (desoxirribosa) y una de las cuatro bases nitrogenadas: *Adenina* (*A*), *Guanina* (*G*), *Timina* (*T*), *Citosina* (*C*) (Figura 1.2.1.1)

La estructura del ADN es una doble hélice. Las bases nitrogenadas se aparean en la doble hélice formando enlaces por puentes de hidrógeno de acuerdo a las siguientes reglas: $G \equiv C$, $A = T$, donde cada "—" simboliza un enlace por puente de hidrógeno y cada par contiene una purina (*A* o *G*) y una piridamina (*C* o *T*). Las bases que forman el par se denominan bases complementarias. Una hélice simple es una cadena de nucleótidos unidos por enlaces fosfodiester. Cada hélice simple se une con la hélice complementaria a través de los enlaces de puentes de hidrógenos que forman las bases, dando lugar a la doble hélice. (Figura 1.2.1.1)

La molécula del ADN es una molécula direccional debido a la estructura asimétrica del azúcar desoxirribosa que forma el esqueleto de las hélices. Cada azúcar es conectado cadena arriba por el quinto carbono y cadena abajo por el tercer carbono, de manera que sí una de las cadenas tiene orientada el azúcar en la dirección 5'$\rightarrow$3' (se lee de cinco prima a tres prima), la cadena complementaria estará orientada en la dirección 3'$\rightarrow$5'. El ARNm está formado por una cadena de nucleótidos unidos por enlaces fosfodiester. En particular constituye una hélice en la cual el azúcar presente en el nucleótido es una ribosa, la base nitrogenada T es sustituida por la base Uracilo (U) y se mantienen las otras bases nucleotídicas: A, C y G como en el ADN. La función del ARNm es transportar hacia el citoplasma la información genética que se encuentra almacenada en el núcleo de la célula en la secuencia de nucleótidos del ADN.

Cuando se traten secuencias de ADN codificantes para proteínas, éstas se pueden escribir en la práctica utilizando las bases T o U indistintamente.

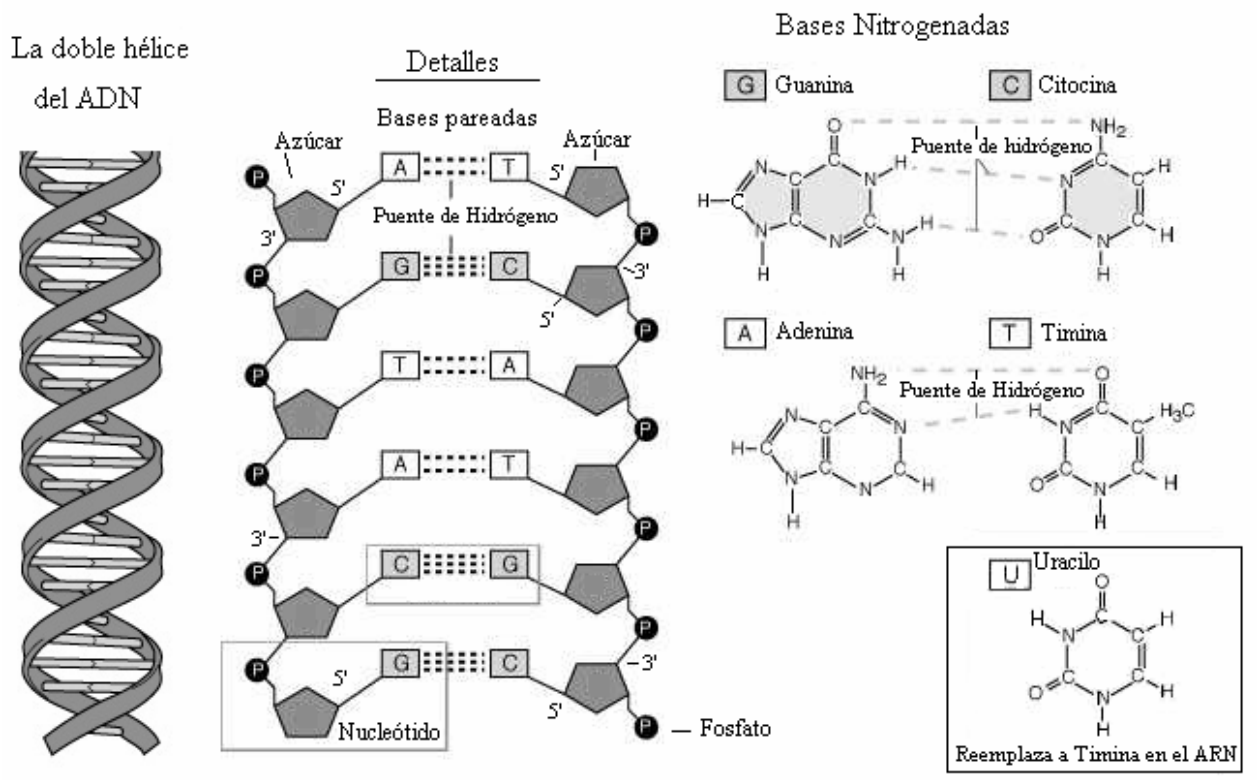

Figura 1.2.1.1. La doble hélice del ADN. Detalles de las bases nitrogenadas y su apareamiento en el ADN.

1.2. El multialineamiento de secuencias

El alineamiento múltiple de secuencias de ADN y proteínas es la piedra angular de la Bioinformática ya que es el punto de partida para los análisis de secuencias clásicos desarrollados hasta el momento. Alinear es una forma de representar y comparar dos o más secuencias o cadenas de ADN, ARN, o estructuras primarias proteicas para resaltar sus zonas de similitud, que podrían indicar relaciones funcionales o evolutivas entre los genes o proteínas analizadas. Las secuencias alineadas se escriben con las letras (representando aminoácidos o nucleótidos) en filas de una matriz en las que, si es necesario, se insertan espacios para que las zonas con idéntica o similar estructura se alineen.

Los alineamientos múltiples de secuencias son útiles para identificar similitudes y diferencias entre secuencias, producir árboles filogenéticos, y desarrollar modelos de homología sobre estructuras de proteínas (Figura 1.2.1). Sin embargo, la relevancia biológica de los alineamientos no siempre es clara. Se asume a menudo que los alineamientos reflejan un grado

de cambio evolutivo entre secuencias que descienden de un ancestro común; pero es formalmente posible que la convergencia evolutiva pueda darse para producir similitudes aparentes entre proteínas que no estén evolutivamente relacionadas, pero que lleven a cabo funciones similares y tengan parecidas estructuras.

Las representaciones visuales del alineamiento ilustran la inserción de espacios que pueden ser interpretados como mutaciones (un solo cambio de aminoácidos o nucleótidos) que aparecen como diferentes caracteres en una sola columna del alineamiento, y la inserción o eliminación (*delection*) de mutaciones, llamadas mutaciones "*indels*" que aparecen como huecos en una o varias de las secuencias en la alineación.

En la actualidad se disponen muchos software que permiten realizar el alineamiento múltiple de secuencias. Además, el servicio de alineamiento de secuencias se oferta gratuitamente en servidores en Internet., En este trabajo, en particular fueron utilizados los productos de software MEGA 4 (Tamura K et al., 2007) y BioEdit, los cuales se pueden obtener en los sitios webs: http://www.megasoftware.net/ y http://www.mbio.ncsu.edu/BioEdit/bioedit.html, respectivamente. En la Figura 1.2.1 se ilustra una región del alineamiento de secuencias codificantes para proteínas mitocondriales de 14 mamíferos obtenido con el BioEdit.

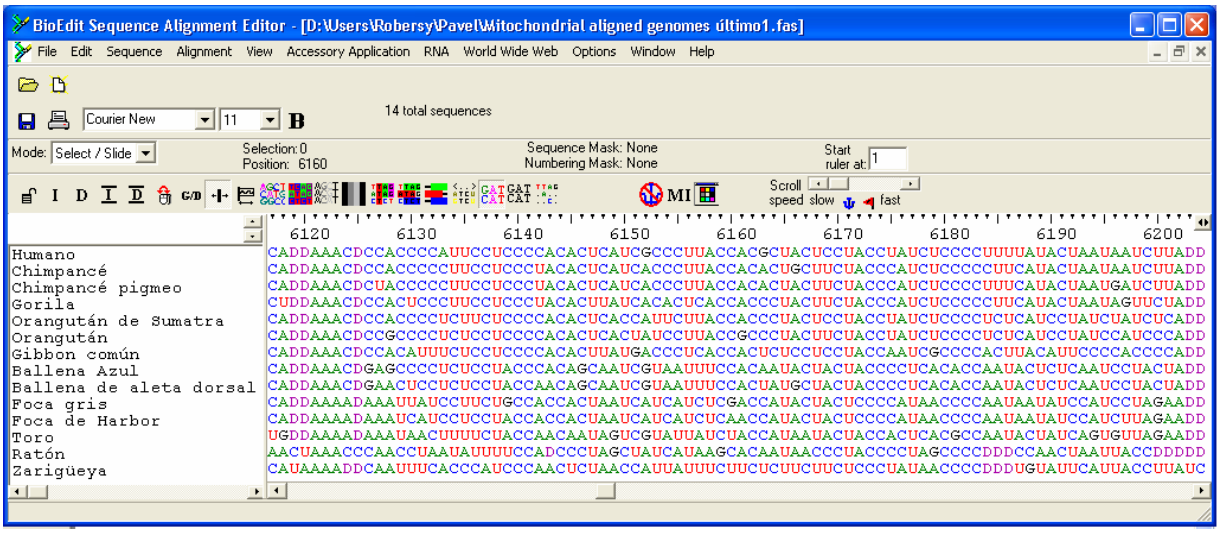

Figura 1.2.1. Región de un alineamiento de secuencias codificantes para proteínas mitocondriales de 14 mamíferos.

1.3. Herramientas algebraicas

En esta sección se abordan las principales propiedades de estas estructuras que nos permitieron obtener una descripción formal de los objetos de nuestro estudio: el conjunto de las bases del ADN. La naturaleza de conjunto finito de los nucleótidos sugiere que las álgebras abstractas constituyen herramientas ideales para desentrañar las regularidades presentes en las relaciones cuantitativas y cualitativas existentes entre los elementos de este conjunto. Las estructuras algebraicas que se abordan son bien conocidas y pueden estudiarse más profundamente en los libros de texto dados en la referencias (Birkhoff and MacLane, 1941, Dubreil and Dubreil-Jacotin, 1963, Kostrikin, 1980, Redéi, 1967, Waerden, 1970), por ello aquí solo se plasmarán las principales definiciones que serán utilizadas en secciones posteriores.

Definición 1.3.1. Sea S un conjunto no vacío ($S \neq \phi$). Toda aplicación $f: S \times S \rightarrow S$ recibe el nombre de operación binaria en S o ley de composición interna en S (para abreviar, ley interna).

En otras palabras una ley interna está dada en S cuando todo par de elementos (x, y) $(x, y \in S)$ tiene asociado otro elemento $z \in S$. Si "$\bullet$" es una ley interna en S entonces, $\bullet(x, y)$ será denotado por $x \bullet y$, es decir, la imagen z es denotada por $x \bullet y$.

Definición 1.3.2: Se denomina semigrupo al par $(G, \bullet)$ compuesto por el conjunto de elementos G y la ley interna "$\bullet$" en G, la cual para cualesquiera $x, y, z \in G$ satisface la ley asociativa:

$$(x \bullet y) \bullet z = x \bullet (y \bullet z)$$

Definición 1.3.3: Un semigrupo es un grupo si para cualesquiera $x, y, z \in G$ satisface las siguientes leyes:

i. *Existencia de neutro*: Existe en G un elemento neutral e tal que: $x \bullet e = e \bullet x$

ii. *Existencia de inversos*: Para todo $x \in G$ existe un elemento inverso x^{-1} tal que: $x \bullet x^{-1} = x^{-1} \bullet x = e$

En particular, el subconjunto $H \subset G$ se denomina subgrupo de G si $e \in H$; $h_1, h_2 \in H \Rightarrow h_1 \bullet h_2 \in H$ y $h \in H \Rightarrow h^{-1} \in H$. Además, el grupo $(G, \bullet)$ es llamado "grupo abeliano" (grupo conmutativo) si para todo $x, y \in G$ la operación binaria satisface: $x \bullet y = y \bullet x$ (ley

conmutativa). En los grupos abelianos la ley interna es usualmente denotada por el símbolo "+" y es llamada operación suma, mientras que al elemento neutro es denotado por el símbolo "0".

Definición 1.3.4: Sea $(G, \bullet)$ un grupo y $a \in G$. Se denomina orden del elemento a al menor entero positivo n, si existe, tal que $a^n = e$.

Aquí a^n denota $a\bullet\ldots\bullet a$ (n veces). En el caso de un grupo abeliano usualmente se escribe na en lugar de a^n, $na = a + \ldots + a$ (n veces).

Definición 1.3.5: Se denomina anillo al triplete $(A, +, \bullet)$ compuesto por el conjunto A y las leyes internas "+" y "•" en A, con las siguientes propiedades:

i. $(A, +)$ es un grupo conmutativo

ii. $(A , \bullet)$ es un semigrupo

iii. El producto "•" es distributivo respecto a la adición "+".

El anillo $(A, +, \bullet)$ tiene una identidad multiplicativa (elemento unidad) si existe un elemento $1_A \in A$ tal que para todo $x \in A$: $x \bullet 1_A = 1_A \bullet x = x$. Se dice entonces que A es un anillo unitario. Además, se dice que el elemento $x \in A$ es inversible o posee inverso en el anillo A, si existe un elemento $x^{-1} \in A$ tal que $x^{-1} \bullet x = 1_A$.

Definición 1.3.6: Se denomina campo a todo anillo $(F, +, \bullet)$ unitario para el cual el conjunto $F^* = F \setminus \{0\}$ con la ley interna "•" es un grupo conmutativo.

En particular, el grupo $(F^*, \bullet)$ es convencionalmente llamado grupo multiplicativo del campo.

Definición 1.3.7: Se denomina característica de un anillo A con elemento unidad 1_A al menor número entero positivo n, si existe, tal que: $1_A + \ldots + 1_A = 0$ (con n sumandos). Si no existe tal n, decimos que la característica de A es 0.

Definición 1.3.8: La aplicación $\varphi : G_1 \rightarrow G_2$ del grupo $(G_1, +)$ en $(G_2, \oplus)$ es un homomorfismo de grupo si para todo $x, y \in G_1$, se cumple:

$$\varphi(x + y) = \varphi(x) \oplus \varphi(y)$$

Definición 1.3.9: Sean $(R, +, \bullet)$ y $(S, \oplus, \otimes)$ anillos. Entonces se dice que la función $\varphi: (R, +, \bullet) \rightarrow (S, \oplus, \otimes)$ es un homomorfismo de anillos si para todo par de elementos $x, y \in R$ se cumplen las siguientes igualdades:

i. $\varphi(x + y) = \varphi(x) \oplus \varphi(y)$

ii. $\varphi(x \bullet y) = \varphi(x) \otimes \varphi(y)$

Si el homomorfismo φ es biyectivo entonces se dice que φ es un isomorfismo de anillos. Si $R = S$ entonces se dice que φ es un endomorfismo de anillos y, finalmente, si φ biyectivo y $R = S$ entonces se dice que φ es un automorfismo de anillos. En particular, cuando el anillo tratado es un campo la función φ es un homomorfismo, endomorfismo, isomorfismo o automorfismo de campos.

1.3.1. Sumario de las operaciones en los campos de Galois.

En el álgebra abstracta, un campo finito o campo de Galois (llamado así por Évariste Galois) es un campo que contiene un número finito de elementos. Todos los campos finitos tienen característica prima, y por lo tanto, su tamaño (o cardinalidad) es de la forma p^n, para p primo y $n > 0$ entero (Birkhoff and MacLane, 1941) y se denotan $GF(p^n)$ (por las siglas en inglés de Galois field). En nuestro trabajo hacemos uso de $GF(p^n)$ para n=1 y en lo adelante cuando nos refiramos a un campo de Galois o campo finito de la forma $GF(p)$ estaremos hablando de $\mathbb{Z}_p$ (considerado con las operaciones usuales de suma y multiplicación módulo p), esto lo avala el hecho de que todos los campos de orden p son isomorfos a $\mathbb{Z}_{p.}$

1.3.2. Representación de secuencias extendidas de ADN y ARN

Resulta adecuado mencionar que muchos de los primeros trabajos vinculados con las secuencias de ADN abordaron la codificación binaria de las bases y la aplicación de la teoría de la información. Se destaca el trabajo pionero de Mercer en el análisis computacional de las secuencias en el año 1968 (Mercer-Hursh, 1968) utilizando la codificación: U=00, C=01, A=10, G=11. La lista de trabajos hasta el presente es bien extensa: Bertman y Jungck en 1979 (Bertman and Jungck, 1979), Swason en 1984 (Swanson, 1984), Riveron y colaboradores en 1986 (Cantillo et al., 1986., Pérez, 1999, Riveron et al., 1986), etc.

Los resultados recientemente obtenidos en el desarrollo de nuevas estructuras algebraicas del código genético, condujeron al análisis del alfabeto extendido de las bases nitrogenadas del ADN (Sánchez and Grau, In Press) http://arxiv.org/abs/0805.1128. La extensión natural del alfabeto de ADN nos permite definir un nuevo campo de Galois $GF(5)$ sobre el conjunto del alfabeto extendido del ARN {D, A, C, G, U}, donde la letra D simboliza una (o más) base(s)

hipotética(s) alternativa(s) con apareamiento no específico presente en las moléculas de ADN y ARN primitivas.

Si una estructura algebraica de campo de Galois es definida en el alfabeto extendido de las bases sujeto a las restricciones A + U = U + A = D y A • U = U • A = G entonces las operaciones de suma y el producto pueden ser definidas en los conjunto {D, G, A, U, C}. Es decir, se requiere que las bases A y U sean inversos en las operaciones de suma y producto, con la base G como elemento neutro para el producto. De esta forma, estas definiciones reflejan el pareo de bases acorde al número de puentes de hidrógeno $G \equiv C$ y $A = U$ de Watson-Crick (Crick, 1968) distintivo en la actual molécula de ADN y en el apareamiento no específico de la(s) base(s) ancestral(es) hipotética(s) D. Las definiciones de las operaciones de suma y producto son presentadas en la Tabla 1.3.2.1. Por construcción, el campo definido en el alfabeto extendido de bases es isomorfo al campo de los enteros módulo 5 ($\mathbb{Z}_5$), una simple representación de *GF*(5). Explícitamente, la biyección es: D↔0, G↔1, A↔2, U↔3, C↔4. Las secuencias de ADN extendidas obtenidas en el multialineamiento pueden ser consideradas entonces como vectores de $(\mathbb{Z}_5)^n$.

Tabla 1.3.2.1. Tabla con las operaciones del campo de Galois definido en el conjunto ordenado del alfabeto extendido de bases B = {D, G, A, U, C}.

Suma

+	D	G	A	U	C
D	D	G	A	U	C
G	G	A	U	C	D
A	A	U	C	D	G
U	U	C	D	G	A
C	C	D	G	A	U

Producto

•	D	G	A	U	C
D	D	D	D	D	D
G	D	G	A	U	C
A	D	A	C	G	U
U	D	U	G	C	A
C	D	C	U	A	G

1.4. Transformada Discreta de Fourier

Las señales son usualmente convertidas del dominio del tiempo o el espacio al dominio de la frecuencia a través de transformadas de Fourier. En matemáticas, la Transformada Discreta de Fourier (DFT, por sus siglas en inglés) es una de las formas específicas de análisis de Fourier. Como tal, esta transforma una función en otra, la cual es llamada representación en el dominio de la frecuencia, o simplemente la DFT de la función original (que comúnmente es una función en el dominio del tiempo). Pero la DFT requiere una función de entrada que es discreta y cuyos valores tienen una limitada (finita) duración. La transformada de Fourier convierte la información contenida en la señal a una componente de magnitud y fase de cada frecuencia. Usualmente la transformada de Fourier es convertida en el espectro de potencia, el cual es el cuadrado de la magnitud de cada componente de frecuencia. En este trabajo la Transformada Discreta de Fourier fue calculada de acuerdo a la ecuación [1.4.1]:

$$F(s/N) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} g_n e^{\frac{2\pi i}{N} sn} \qquad [1.4.1]$$

donde $s = 0,\ldots, N-1$ y la frecuencia $f = s/N$. El espectro de potencia es dado por $S(f) = \left| F(s/N) \right|^2$. Nótese que hay N componentes de frecuencia y el espectro de potencia de una señal puede ser visualizado en una gráfico f contra $S(f)$. El primer valor calculado en [1.4.1] es el término de frecuencia cero. La frecuencia cero es descartada en el análisis porque correspondería a un período “infinito”.

1.5. Transformada Discreta de Fourier sobre campos finitos

La transformada discreta sobre un campo finito F es una extensión del concepto de DTF a campos finitos. Esta transformada se denota en este texto como MFT, por el acrónimo en inglés de *Modular Fourier Transform*, pues se aplica la transformada sobre $\mathbb{Z}_p$ aunque la teoría se expone de modo general, sobre cualquier campo finito.

Ahora, sea F un campo finito.

Definición 1.5.1: Llamaremos raíz primitiva n-ésima de la unidad en F a un elemento $w \in F$ de orden n.

Para el caso particular de $F=\mathbb{Z}_p$, notemos que cualquier elemento $w \in F$ distinto del cero es raíz primitiva en F, pues del pequeño teorema de Fermat tenemos que si $w \neq 0 \pmod{p}$ entonces $w^{p-1} = 1 \pmod{p}$. Además tenemos que existe raíz primitiva n-ésima de la unidad en $\mathbb{Z}_p$ si y solo si n divide a p-1 (Gao, 2001).

En lo adelante se considera $n > 1$.

Lema 1.5.1. (Propiedad de Cancelación). Sea w una raíz primitiva n-ésima de la unidad en un campo F. Entonces se cumple que

$$\sum_{j=0}^{n-1} w^{js} = \begin{cases} 0, & s \neq 0 \pmod{n} \\ n, & s = 0 \pmod{n} \end{cases}$$

Demostración:

El resultado es claro si $s = 0 \pmod{n}$ pues $w^{sj} = 1 \ \ \forall j = \overline{0,\, n-1}$. En otro caso, consideremos la siguiente factorización sobre F, $x^n - 1 = (x-1)(\sum_{j=0}^{n-1} x^j)$. Sustituyendo $x = w^s$ se hace cero el miembro izquierdo de la igualdad, luego $(w^s - 1)(\sum_{j=0}^{n-1} w^{js}) = 0$ por tanto, como $w^s \neq 1$ para $s \neq 0 \pmod{n}$ (por ser w una raíz primitiva n-ésima de la unidad) se obtiene el resultado pedido.

Definición 1.5.2: Para un entero positivo n diremos que F soporta una MFT para $x^n - 1$ si $x^n - 1$ tiene n raíces distintas en F.

Lema 1.5.2. Supongamos que F soporta una MFT para $x^n - 1$ y sean $w_1, w_2, \ldots w_n \in F$ las raíces de $x^n - 1$, entonces $w_1, w_2, \ldots w_n$ forman un subgrupo del grupo multiplicativo de F.

Demostración:

Sean w_i, w_j dos raíces de $x^n - 1$ sobre F, entonces $(w_i w_j^{-1})^n = w_i^{\,n}\, w_j^{\,-n} = 1$ y el resultado es inmediato.

El lema 1.5.2 nos permite obtener el siguiente resultado.

Lema 1.5.3. Sea w una raíz primitiva n-ésima de la unidad sobre el campo F que soporta una MFT para $x^n - 1$, entonces w es un generador del subgrupo multiplicativo formado por las raíces de $x^n - 1$.

Demostración:

Sean $w_1, w_2, \ldots w_n \in F$ las raíces de $x^n - 1$. Debemos demostrar entonces que $\{w_1, w_2, \ldots w_n\} = \{1, w, w^2, \ldots w^{n-1}\}$. Es claro que w^i es raíz de $x^n - 1$. Solo resta comprobar que $w^i \neq w^j$ si $i \neq j$. Supongamos que $i \leq j$ veamos que $w^i = w^j \Leftrightarrow w^i(w^{j-i} - 1) = 0 \Leftrightarrow w^{j-i} - 1 \Leftrightarrow i = j$ las dos últimas implicaciones se obtienen pues *w* es una raíz primitiva *n*-ésima de la unidad, completando así la demostración.

Sea $T_n(w)$ la matriz

$$\begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{n-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{n-1} & w^{2(n-1)} & \ldots & w^{(n-1)^2} \end{bmatrix}$$

donde *w* es una raíz primitiva *n*-ésima de la unidad sobre *F*.

Definición 1.5.3: Sea $a = (a_0, a_1, \ldots, a_{n-1})^t \in F^n$. La transformada discreta de Fourier de a sobre el campo *F* es $MFT(a, n, w) := T_n(w) \cdot a$.

Del lema 1.5.3 tenemos que $\{w_1, w_2, \ldots w_n\} = \{1, w, w^2, \ldots w^{n-1}\}$ por tanto es válida la relación $MFT(a, n, w) := T_n(w) \cdot a = (f(1), f(w), f(w^2), \ldots, f(w^{n-1}))$, donde w, n y $w_1, w_2, \ldots w_n$ se consideran como en el enunciado del lema 1.5.3 y $f(x) = a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}$.

Es importante destacar que cuando *w* es una raíz primitiva *n*-ésima de la unidad en *F*, podemos definir la transformada discreta inversa de Fourier sobre campos finitos (IMFT, acrónimo en inglés de *Inverse Modular Fourier Transform*).

Definición 1.5.4: Sea $A = (A_0, A_1, \ldots, A_{n-1})^t \in F^n$. La transformada discreta inversa de Fourier sobre el campo *F* es $IMFT(A, n, w) := \frac{1}{n} T_n(w^{-1}) \cdot A$.

Notemos que $w^{-1} = w^{n-1}$ y además, usando el lema 1.5.1 se puede demostrar que $T_n(w) \cdot T_n(w^{-1}) = n I_n$, donde I_n denota la matriz idéntica de $n \times n$ (Yap, 2000).

1.5.1. Transformada Rápida de Fourier sobre un campo finito *F*.

La aplicación directa de la MFT puede tornarse computacionalmente costosa a medida que aumenta el valor de *n*. De acuerdo a la definición 1.5.3 podemos obtener que la complejidad temporal de la MFT es n^2.

En 1965, Cooley y Tukey descubrieron un método que es conocido como la transformada rápida de Fourier (nombrémosla MFFT, para el caso sobre campos finitos).

El algoritmo MFFT usa la técnica de "divide y vencerás" la cual reduce un problema de gran tamaño a la solución de varios problemas similares más pequeños, aplicado este mecanismo de forma recursiva.

Supongamos que *F* soporta una MFT para $x^n - 1$ y que $n = 2t$ donde $t \geq 1$ es un entero arbitrario, entonces

$x^n - 1 = x^{2t} - 1 = (x^t - 1)(x^t + 1)$.

Sean las *t* distintas raíces de $x^t - 1$ (en cualquier orden). Entonces

$$x^t - 1 = \prod_{i=1}^{t} (x - w_i).$$

Sea $w \in F$ una raíz de $x^t + 1$, luego $w^t = -1$. Veamos que

$$x^t + 1 = x^t - w^t = w^t \cdot \left(\left(\frac{x}{w} \right)^t - 1 \right) = w^t \cdot \prod_{i=1}^{t} \left(\frac{x}{w} - w_i \right) = \prod_{i=1}^{t} (x - w \cdot w_i).$$

Por tanto todas las raíces de $x^n - 1$ son

$$w_1,\ w_2,\ \ldots\ w_t,\ w \cdot w_1,\ w \cdot w_2,\ \ldots\ w \cdot w_t.$$

Sea $f = \sum_{i=0}^{n-1} f_i x^i \in F[x]$. Queremos determinar

$$f(w_1),\ f(w_2),\ \ldots\ f(w_t),\ f(w \cdot w_1),\ f(w \cdot w_2),\ \ldots\ f(w \cdot w_t).$$

(Notemos que esto es lo mismo que hallar $f(w^i)$ para $0 \leq i \leq n-1$).

Escribamos

$$f = \sum_{i=0}^{t-1} f_i x^i + x^t \cdot \sum_{i=0}^{t-1} f_{t+i} x^i = g + x^t \cdot h = (g,\ h).$$

donde *g* denota la primera mitad de *f* y *h* la segunda mitad. Entonces

$$f = g + h \pmod{x^t - 1}, \qquad f = g - h \pmod{x^t + 1}.$$

El grado de $g + h$ es menor que t, y toma los mismo valores que f en w_i, $1 \le i \le t$. Análogamente $g - h$ toma los mismos valores que f en $w \cdot w_i$, $1 \le i \le t$. Veamos los valores de $g - h$, sea $g - h = \sum_{j=0}^{t-1} u_j x^j = u$. Entonces

$$u(w \cdot w_i) = \sum_{j=0}^{t-1} (u_j \cdot w^j) {w_i}^j, \quad 1 \le i \le t.$$

Definamos

$$WT(u, w) = \sum_{j=0}^{t-1} (u_j \cdot w^j) x^j = (u_0 \cdot 1, u_1 \cdot w, \ldots, u_i \cdot w^i, \ldots, u_{t-1} \cdot w^{t-1}),$$ al cual llamamos polinomio pesado de u en w. Entonces tenemos

$$MFFT(f, 2t, w) = (MFFT(g_1, t, w), MFFT(h_1, t, w)),$$

donde $g_1 = g + h, \quad h_1 = WT(g - h, w)$.

De esta forma un problema de tamaño $2t$ es reducido a dos problemas similares de tamaño t cada uno. Cuando t es par, podemos aplicar esta reducción nuevamente a cada uno de los problemas más pequeños.

Mostramos a continuación el algoritmo iterativo para calcular la transformada discreta de Fourier. Para cualquier arreglo $f = (f_0, f_1, \ldots, f_{n-1})$ y enteros $j \ge i \ge 0$, hagamos el convenio $f[i \ldots j] = (f_i, f_{i+1}, \ldots, f_j)$.

Algoritmo MFFT.

Entrada: $n = 2^k$, $(k \ge 1)$, $w \in F$ de orden n, y $f = (f_0, f_1, \ldots, f_{n-1}) \in F^n$.

Salida: $MFFT(f, n, w)$, o sea, los valores de f en w^i, $0 \le i \le n - 1$.

1. $T := f$, $m := n$, $s := 1$, $a := w$.
2. for i from $k - 1$ downto 0 do
3. $\quad t := m / 2$.
4. $\quad$ for j from 0 to $s - 1$ do
5. $\quad\quad h := T[jm \ldots jm + t - 1]$,
6. $\quad\quad T[jm \ldots jm + t - 1] := h + T[jm + t \ldots jm + m - 1]$,

7. $T[jm+t \ldots jm+m-1] := h - T[jm+t \ldots jm+m-1],$
8. $T[jm+t \ldots jm+m-1] := WT(T[jm+t \ldots jm+m-1],\ a).$
9. $m := m/2,\ \ s := 2s,\ \ a := a^2.$
10. Retornar *T*.

Es importante notar que como resultado de la aplicación de la MFFT el vector final no se obtiene en el orden "natural" $f(w^0),\ f(w^1), \ldots, f(w^i),\ \ldots, f(w^{n-1})$.

Para establecer el orden correcto definamos una permutación σ del conjunto $\{0, 1, \ldots, n-1\}$ donde $n = 2^k$. Para cada $0 \le i < n$ escribimos i en forma binaria, $i = i_0 + i_1 \cdot 2 + \ldots + i_{k-1} \cdot 2^{k-1} = (i_0, i_1, \ldots, i_{k-1})_2$. Agregamos los ceros necesarios para lograr que la representación en base dos de *i* tenga *k* bits.

Definimos $\sigma(i) = (i_{k-1}, \ldots, i_1, i_0)_2$, o sea $\sigma(i)$ se obtiene de *i* invirtiendo sus *k* bits en base dos.

Tenemos entonces el siguiente resultado:

$MFFT(f, n, w) = (f(w^{\sigma(0)}),\ f(w^{\sigma(1)}), \ldots, f(w^{\sigma(i)}),\ \ldots, f(w^{\sigma(n-1)}))$.

Con el fin de reorganizar el vector de salida del algoritmo MFFT, se utiliza la siguiente rutina.

Rutina EON. Establecer el orden "natural" del vector resultante en la salida de la MFFT.

Entrada: $T = (f(w^{\sigma(0)}),\ f(w^{\sigma(1)}), \ldots, f(w^{\sigma(i)}),\ \ldots, f(w^{\sigma(n-1)}))$, *n*-orden de la raíz primitiva *w*

Salida: $T = (f(w^0),\ f(w^1), \ldots, f(w^i),\ \ldots, f(w^{n-1}))$.

1. *correcto*[1…*n*] := False.
2. for *j* from 0 to *n*-1 do
3. if not *correcto*[*j*+1] then
4. *i* := *ReversoBinario*(*j*, *k*),
5. *swap*(*T*[*j*+1], *T*[*i*+1]), *correcto*[*j*+1] := True, *correcto*[*i*+1] := True.
6. Retornar *T*.

Donde la función *ReversoBinario*(*j*, *k*) devuelve el número decimal $\sigma(j) = (j_{k-1}, \ldots, j_1,\ j_0)_2$ siendo $j = (j_0,\ j_1, \ldots, j_{k-1})_2$ y la función *swap*(*x*, *y*) intercambia los valores de las variables *x*, *y*.

Luego, al combinar el algoritmo MFFT con la rutina EON se logra el mismo vector de salida que al aplicar la función $MFT(f, n, w)$.

1.5.2. Análisis de la complejidad temporal

En el algoritmo MFFT, las instrucciones en 1, 3 y 9 tienen costo $O(1)$. Las asignaciones 5, 6, 7 y 8 tienen costo $O(t)$. El ciclo 4-8 se ejecuta s veces y tiene, por tanto, costo $O(s \cdot t)$. Ahora, nótese que al iniciar la ejecución del ciclo 4-8 las variables s y t toman los valores $s = 2^{i-1}$ y $t = 2^{k-i}$ donde i simboliza la i-ésima vez que se ha ejecutado la sentencia for 4-8 en una misma corrida del algoritmo. Luego, el ciclo 4-8 tiene costo $O(2^{i-1} \cdot 2^{k-i}) = O(2^{k-1}) = O(n)$. El ciclo 2-7 se ejecuta k veces, por tanto, tiene costo computacional $O(k \cdot n) = O(n \cdot \log n)$, siendo este el costo del algoritmo.

En el caso de la rutina EON, el paso 1 tiene complejidad $O(n)$. La instrucción 4 tiene costo $O(2k) = O(\log n)$ y es la que determina la complejidad del ciclo 2-5, por lo que este tiene complejidad $O(n \cdot \log n)$. De esta manera se determina que la complejidad de la rutina EON es $O(n + n \cdot \log n) = O(n \cdot \log n)$.

En resumen, la combinación del MFFT y EON tiene complejidad $O(n \cdot \log n)$.

1.6. Sobre la programación en el paquete *Mathematica*

Mathematica se distingue de los tradicionales lenguajes, en el soporte de varios paradigmas de programación. Por ejemplo, en lenguajes como C y Java, así como en la mayoría de los *script*s, la programación procedural es el único paradigma disponible. *Mathematica* soporta todas las construcciones estándares de la programación procedural, pero a menudo las extiende a través de la integración en un entorno de programación simbólica más general.

En el corazón del *Mathematica* está su lenguaje simbólico altamente desarrollado que unifica una amplia variedad de paradigmas de programación y usa su propio concepto de programación simbólica para adicionar un nuevo nivel de flexibilidad a los conceptos clásicos de programación. Así el *Mathematica* soporta especialmente:

- Programación estructurada
- Programación orientada a listas
- Programación orientada a funciones
- Programación orientada a reglas de sustitución

Todas ellas integradas y enriquecidas con posibilidades simbólicas en un único lenguaje denominado *Mathematica Core Language*.

Las listas son construcciones centrales en *Mathematica*. Se usan para representar colecciones, arreglos, conjuntos y secuencias de todo tipo. Tienen cualquier estructura y medida y fácilmente pueden abarcar millones de elementos. Muchas funciones del *Mathematica*, y por supuesto, funciones definidas por el usuario pueden operar directamente sobre listas, haciendo de ellas un potente vehículo para la interoperabilidad, la paralelización de operaciones y evitar el uso de lazos, aun cuando estos están previstos en el lenguaje.

En el corazón de *Mathematica* está la idea fundamental de su funcionamiento: los datos, programas, fórmulas, gráficos y documentos pueden ser representados como expresiones simbólicas. Este concepto unificador es lo que subyace en el paradigma de programación simbólica y hace posible gran parte de la potencia única del lenguaje y el sistema.

El paradigma de lenguaje simbólico eleva el concepto de variables y funciones a un nuevo nivel. Una variable, además ser evaluada, puede ser usada de una forma puramente simbólica. Además para conformar un lenguaje de patrones potentes, pueden ser definidas funciones, no exactamente para tomar argumentos sino para transformar patrones con cualquier estructura. En *Mathematica* todo es una función (principio fundamental del *Core Language*), incluso el propio paquete *Mathematica* es una función.

En el corazón del paradigma de programación simbólica está el concepto de reglas de transformación para patrones simbólicos arbitrarios. El lenguaje de patrones del *Mathematica* describe convenientemente un conjunto muy general de clases de expresiones, haciendo posible programas fácilmente legibles, elegantes y eficientes.

La programación funcional ha sido considerada una importante idea desde hace tiempo, pero la teoría se materializa prácticamente en el lenguaje simbólico del *Mathematica.* El tratamiento de expresiones como $f(x)$, tanto como dato simbólico, como de la aplicación de una función f, suministra una potente manera de integrar estructura y función y es una representación elegante de muchos cálculos computacionales.

Construido sobre potentes y elegantes principios, el *Mathematica Core Language* ha sido mejorado gradualmente bajo un estricto control sobre sus veinte años de historia. La versión 6,0 agrega nuevas funciones de manipulación de listas, mejora patrones y opciones de manejo, así como un nuevo e importante sistema integrado de depuración y análisis de código.

1.7. Consideraciones finales del capítulo

De acuerdo a lo abordado, es posible representar las secuencias genómicas en $\mathbb{Z}_5$ a partir del isomorfismo de este campo con el conjunto extendido de bases. Esta representación numérica de las secuencias será el punto de partida para el análisis comparativo propuesto en el capítulo siguiente.

Se expusieron los fundamentos teóricos de la MFT y un algoritmo para su cálculo, el cual será utilizado posteriormente como método de recodificación de secuencias representadas en $\mathbb{Z}_5$. Las secuencias recodificadas pueden interpretarse como señales discretas en tiempo discreto. Esto permitirá la aplicación de la DTF a estas señales para la búsqueda de los espectros de potencia, mediante los cuales es posible detectar periodicidades en las secuencias de ADN genómico recodificadas.

Finalmente se abordó sobre la filosofía de la programación en el paquete *Mathematica,* evidenciando las ventajas que tendrá su uso en la implementación del método que se propone en las secciones siguientes.

2. IMPLEMENTACIÓN DE LOS ALGORITMOS PARA EL ANÁLISIS DE SECUENCIAS

El algoritmo MFT fundamentalmente ha sido utilizado en el área del álgebra computacional para efectuar multiplicaciones rápidas de polinomios sobre campos finitos. En la literatura disponible no hemos encontrado ninguna aplicación de este algoritmo en el análisis de secuencias de ADN. En este trabajo el algoritmo de la MFT se implementa con el propósito de recodificar las secuencias de ADN representadas sobre el campo $\mathbb{Z}_5$**.**

Por otra parte, la DFT ha sido utilizada en Bioinformática para la detección de genes en regiones de ADN genómico utilizando para ello codificaciones de las bases del ADN seleccionadas *ad hoc*. En particular, en esta investigación la aplicación de la DFT se realiza sobre las secuencias recodificadas mediante la MFT. Finalmente, para el análisis comparativo de los espectros de potencia se programó el análisis de varianza bifactorial no-paramétrico (ANOVA bifactorial no paramétrico).

En este capítulo realizamos la descripción del procedimiento seguido en la programación de estas herramientas para su aplicación en el análisis comparativo de secuencias genómicas.

2.1. MFT como método de recodificación de secuencias representadas en $\mathbb{Z}_5$

En Bioinformática, la detección de genes mediante la aplicación de la DFT en las secuencias de ADN genómico, codificante para proteínas, se ha basado en la presencia o no de picos en los espectros de potencia a las frecuencias de 1/3 y 2/3 (Fuentes et al., 2006a, Fuentes et al., 2006b). Trabajos realizados en el grupo de Bioinformática de la UCLV han demostrado que tales picos también aparecen cuando las secuencias de ADN extendidas son representadas en $\mathbb{Z}_5$ (Sánchez and Grau, 2008). Sin embargo, si nos proponemos explotar aún más esta herramienta de análisis necesitamos encontrar nuevos picos en nuevas frecuencias que nos permitan profundizar en búsqueda de similitudes y diferencias entre los espectros de potencia

de las secuencias de ADN comparadas. Para este propósito se desarrolla una forma de recodificación de las secuencias extendidas a mediante el uso de la Transformada de Fourier sobre campos finitos (MFT), la cual nos permite inducir (como se ilustra más adelante) la aparición forzada de nuevos picos en los espectros de potencia de las señales recodificadas.

La MFT, como ya hemos visto, transforma un vector de $(\mathbb{Z}_p)^n$ en otro vector de $(\mathbb{Z}_p)^n$. Este hecho nos conduce a suponer las secuencias extendidas como vectores en $\mathbb{Z}_p$. Es decir, podemos considerar $(\mathbb{Z}_5)^n$ como un subconjunto de $(\mathbb{Z}_p)^n$, para $p>5$, y sustituir las operaciones del campo $\mathbb{Z}_5$ por las de $\mathbb{Z}_p$. Notemos que $(\mathbb{Z}_5)^n$ con las operaciones de $\mathbb{Z}_p$, no constituye ninguna estructura algebraica pues las leyes no son internas. Esta transformación puede expresarse simbólicamente de la siguiente forma:

$$(\mathbb{Z}_5)^n \subset (\mathbb{Z}_p)^n \longrightarrow (\mathbb{Z}_p)^n$$

$$a \longrightarrow T_n(w)\cdot a$$

donde las operaciones efectuadas se realizan sobre $\mathbb{Z}_p$.

Es importante notar que las secuencias no necesariamente tienen que ser de longitud *n*, siendo *n* el orden de la raíz primitiva *w*. Por tal motivo, para transformar una secuencia completa, se opta por una de las siguientes alternativas:

i. Seleccionar el valor de *n* mucho menor que el tamaño de la secuencia y realizar una partición de la misma en subsecuencias de longitud *n*. Seguidamente a cada una de estas particiones aplicarle la MFT para una misma raíz *n*-ésima primitiva *w*;

ii. Seleccionar el valor de *n* como el mayor número entero menor que la longitud de la secuencia para el cual exista una raíz *n*-ésima primitiva *w* en algún campo $\mathbb{Z}_p$, y aplicar la MFT a la subsecuencia formada por las primeras *n* bases (codificadas en $\mathbb{Z}_5$).

En ambas variantes se desprecia una parte relativamente pequeña de la secuencia original en relación con su longitud inicial. Las bases que se desprecian en la transformación no se tienen cuenta en el análisis posterior, lo cual implica una limitación del método.

Observemos que esta recodificación obtenida, a partir de la aplicación de la MFT, es única para cada raíz *n*-ésima primitiva *w* en un campo $\mathbb{Z}_p$. Además, cada secuencia recodificada puede ser regresada a su representación en $\mathbb{Z}_5$ mediante la Transformada Inversa[1], la cual está bien definida por ser *w* una raíz primitiva. También es necesario destacar que la aplicación del

[1] Ver sección 1.5

algoritmo a una secuencia dependerá de los valores seleccionados para *p* (primo), *n* y *w*, de modo que *p*-1 sea divisible por *n*. Esta última restricción, es condición necesaria y suficiente para la existencia de una raíz *n*-ésima primitiva[1] *w* en $\mathbb{Z}_p$, la cual debe ser hallada luego de la selección de *p* y *n*. Notemos que al variar adecuadamente los parámetros que intervienen en la transformación se pueden obtener diferentes recodificaciones para una misma secuencia.
La secuencia de pasos seguida en la implementación del método descrito se detalla en la siguiente sección.

2.2. Análisis espectral de las señales recodificadas en $\mathbb{Z}_p$

Una vez recodificada la señal se obtiene el espectro de potencia mediante la aplicación de la DFT. Es de destacar que el número de picos espectrales inducidos por la recodificación con la MFT no es superior al orden *n* de la raíz primitiva *w* del campo $\mathbb{Z}_p$ considerado. Este hecho se debe a la partición de las secuencias en el proceso de recodificación según la variante *i* descrita en la sección 2.1. De acuerdo con esta variante cada *n* bases se aplica la MFT, lo cual puede inducir periodicidades en las secuencias observadas a frecuencias múltiplos de 1/*n*. La aparición notable o no de picos a tales frecuencias depende de la secuencia genómica analizada. Este hecho es utilizado en el análisis comparativo de los espectros de potencia al evaluar la significación estadística de las diferencias entre los picos correspondientes en las secuencias comparadas.

2.2.1. Método para hallar los espectros de potencia

Como se ha escrito anteriormente, la obtención de los espectros de potencia requiere de una serie de pasos a seguir. Ahora, sean,
B = {D, G, A, U, C} (Conjunto extendido de bases nitrogenadas)
vectorSecuencias[*i*]: la *i*-ésima secuencia
p: número primo que corresponde a la cantidad de elementos del campo $\mathbb{Z}_p$
w: raíz primitiva en $\mathbb{Z}_p$
n: orden de la raíz primitiva *w*
npart: número de subsecuencias que forman la partición
q: es la cantidad total de secuencias.

[1] Ver sección 1.5

El Método de obtención de espectros de potencia (MOEP) puede resumirse en los siguientes pasos:

1. *LeerSecuencias*(*vectorSecuencias*);
2. *Codificar*(*vectorSecuencias*, $B \leftrightarrow \mathbb{Z}_5$);
3. *Particionar*(*vectorSecuencias*, *n*);
4. Para cada *i* desde 1 hasta *q* hacer
5. Para cada *j* desde 1 hasta *npart* hacer
6. *Recodificar*(*vectorSecuencias*[*i*,*j*], *p*, *n*, *w*);
7. *Componer*(*vectorSecuencias*);
8. Para cada *i* desde 1 hasta *nsec* hacer
9. *armónicos*[*i*] := [*DFT* (*vectorSecuencias*[*i*])]^2;

El método utilizado para la recodificación de las secuencias es el MFFT combinado con la rutina EON o solamente la expresión que define la MFT[1]. El la línea 2 se codifican las secuencias de acuerdo al isomorfismo[2] entre *B* y $\mathbb{Z}_5$. En el paso 3 se realiza la partición como se describe en la variante *i* de la sección 2.1. El procedimiento *Componer* elimina las particiones llevando las secuencias a su forma original. En la instrucción 9 se determinan los espectros de potencia de cada secuencia (*armónicos*[*i*]) mediante la aplicación de la DTF[3].
La secuencia de instrucciones de este algoritmo fue programada en el *Mathematica* 6.0

2.2.2. Análisis de la complejidad temporal

Los parámetros de los cuales depende el costo computacional del método son:

- *q*: el número de secuencias a analizar
- *d*: la longitud de las secuencias alineadas
- *n*: el orden de la raíz primitiva *w*

Se considera costo $O(1)$ para las operaciones básicas en $\mathbb{Z}_p$ para un primo *p*.
Los pasos 1, 2, 3 y 7 requieren un costo $O(q \cdot d)$.
Si el paso 6 se realiza usando la combinación de métodos MFFT y EON entonces tiene costo[4] $O(n \cdot \log n)$, y tenemos que la complejidad del paso 5 es $O((d/n) \cdot n \cdot \log n) = O(d \cdot \log n)$, pues

[1] Ver definición 1.5.3
[2] Ver sección 1.3.2
[3] Ver sección 1.4
[4] Ver sección 1.5.2

npart = *d*/*n*; por tanto, el costo del paso 4 es $O(q \cdot d \cdot \log n)$. Si en el paso 6 se usa la expresión que define la MFT entoces el costo[1] del paso 4 es $O(q \cdot (d/n) \cdot n^2) = O(q \cdot d \cdot n)$.
La complejidad del paso 8 es $O(d^2)$, por lo que, la del paso 9 es $O(q \cdot d^2)$.
De aquí se obtiene que la complejidad del MOEP es $O(q \cdot d^2 + q \cdot d \cdot \log n)$ ó $O(q \cdot d^2 + q \cdot d \cdot n)$. Teniendo en cuenta que $n \leq d$ se puede concluir que, en cualquier caso, la complejidad es $O(q \cdot d^2)$.

2.3. Procedimiento seguido en el análisis comparativo de los espectros de potencia.

Para cada secuencia recodificada en $\mathbb{Z}_p$ se obtiene un único espectro, lo cual limita el potencial del análisis estadístico que deseamos aplicar. Sin embargo, si lo que nos interesa son las características periódicas globales que se preservan en las subsecuencias del genoma estudiado, entonces podemos obtener los espectros de potencia de estas subsecuencias. Es decir, para cada secuencia genómica podemos obtener varias repeticiones correspondientes a los espectros de potencia de sus subsecuencias no superpuestas. En este punto encontramos el efecto no deseado de la aparición de picos espectrales a determinadas frecuencias que no son característicos de la secuencia global.
La aparición de picos espectrales no característicos de la secuencia global puede ser filtrada utilizando el "espectro medio de una subsecuencia". Para introducir este concepto supongamos que dividimos las subsequencias en tres ventanas solapadas de igual longitud. Entendemos por espectro medio de una subsecuencia al valor medio de los espectros obtenidos de cada una de estas ventanas. En otras palabras, supongamos que tenemos una subsecuencia de una de las secuencias estudiadas recodificadas en $\mathbb{Z}_p$ (Figura 2.3.1). El espectro medio de esta subsecuencia se obtiene al dividir la misma en tres ventanas de igual longitud solapadas y obtener para cada ventana los armónicos como resultado de la aplicación de la DTF. Finalmente la magnitud de un armónico, correspondiente a una frecuencia dada, se determina como la media de las magnitudes de los armónicos de los tres espectros para la frecuencia

[1] Ver primer párrafo de la sección 1.5.1

fijada (un armónico por cada ventana). Naturalmente, este proceso se realiza para cada una de las frecuencias.

Debe destacarse que para cada secuencia genómica se pueden obtener varios espectros medios correspondientes a subsecuencias no superpuestas. Es decir, una secuencia genómica puede ser particionada en *k* subsecuencias y para cada una de estas podemos obtener un espectro medio.

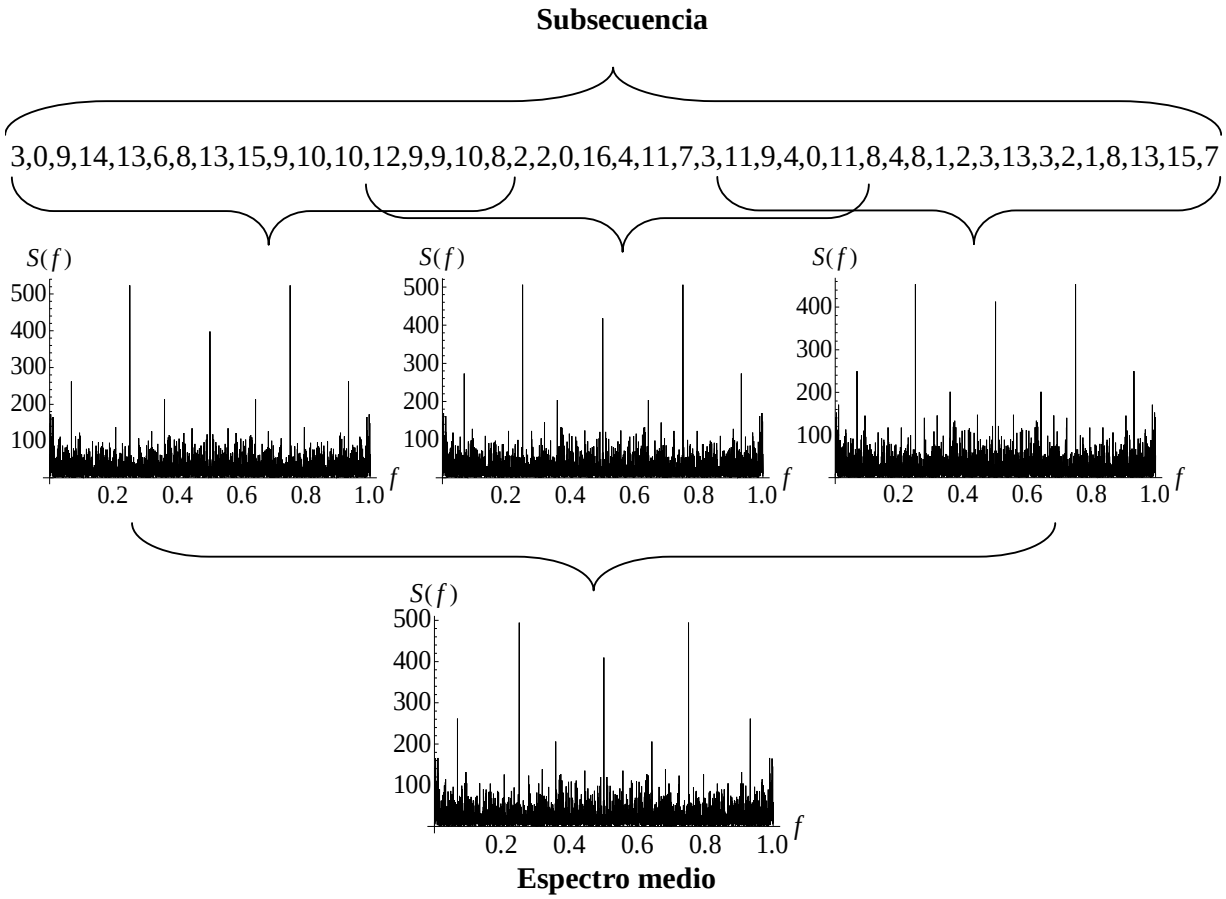

Figura 2.3.1. Ilustración del procedimiento seguido para la obtención de los espectros medios. En el ejemplo la secuencia fue recodificada en $\mathbb{Z}_{17}$, orden 8 y w=2.

2.3.1. ANOVA bifactorial no paramétrico en la comparación de los espectros de potencia

Para buscar diferencias entre los espectros de las señales recodificadas en $\mathbb{Z}_p$ realizamos un ANOVA bifactorial no paramétrico en el cual la variable analizada (variable dependiente) es la magnitud de los armónicos presentes en los espectros de potencia a determinadas frecuencias. Los dos factores involucrados son las secuencias y las frecuencias a las cuales aparecen los armónicos. De manera que cada factor puede tener dos o más niveles. Cada nivel del factor secuencia denota una secuencia analizada, mientras que cada nivel del factor frecuencia denota una frecuencia para la cual se analiza la variable dependiente. Se recomienda analizar no más de cinco niveles para cada factor. Los datos que son analizados mediante el ANOVA son conformados seleccionando los niveles de cada uno de los dos factores y las repeticiones de la variable a analizar para cada combinación secuencia-frecuencia (factor1-factor2).

Los niveles del factor secuencia de ADN (factor1) se corresponden con los genomas de los organismos que se desean comparar. Los niveles del factor frecuencia (factor2) se escogen entre los n posibles picos observables a las frecuencias múltiplos de $1/n$, pues el número de picos espectrales inducidos por la recodificación con la MFT no es superior al orden n de la raíz primitiva w del campo $\mathbb{Z}_p$ considerado[1].

Para obtener k repeticiones de cada combinación de factores de la variable dependiente se realiza una partición de cada secuencia en k subsecuencias y se obtiene el espectro medio de estas, tal como fue descrito en la sección 2.2.1. Ahora, sean s una de las secuencias seleccionadas como niveles del factor1 y f una de las frecuencias seleccionadas como niveles del factor2. Las magnitudes de los armónicos correspondientes a cada par (s, f) para la i-ésima subsecuencia de s, constituye la repetición r_i ($i = 1,2,\ldots, k$) de la variable dependiente. En otras palabras, se obtienen (s, f, r_i) repeticiones para cada combinación de factores $s \times f$. En el Anexo 1 se presentan las bases teóricas de este ANOVA, su programación utilizando el paquete Mathematica 6.0 y se ilustra con un ejemplo el procedimiento seguido en el análisis.

[1] Ver inicio de la sección 2.2

2.3.2. Generación aleatoria de secuencias de ADN

En la aplicación de las herramientas desarrolladas en este trabajo resulta útil la generación aleatoria de secuencias genómicas extendidas. Las comparaciones que se realizan requieren la generación de tres tipos de secuencias aleatorias:

i. Secuencias generadas con distribución discreta $\{p_D, p_G, p_A, p_U, p_C\}$.

ii. Secuencias generadas con distribución discreta $\{p_D, p\}$.

iii. Secuencias generadas con distribución uniforme.

Se describen a continuación como se generan cada una de las secuencias mencionadas.

En la variante *i* las secuencias se generan de acuerdo a una distribución de probabilidad discreta $\{p_D, p_G, p_A, p_U, p_C\}$ donde p_D, p_G, p_A, p_U, p_C representan respectivamente las probabilidades de aparición de las bases extendidas D, G, A, U, C en la secuencia generada, por tanto $p_D + p_G + p_A + p_U + p_C = 1$. En el análisis realizado, cada uno de los valores de p_D, p_G, p_A, p_U, p_C se obtuvo sumando las frecuencias de aparición de cada base en cada una de las secuencias naturales estudiadas y dividiendo por la suma total de las longitudes de todas las secuencias naturales en estudio. Formalmente: $p_b = \frac{n_b}{\sum_{b=1}^{5} n_b}$ donde n_b es la frecuencia de aparición de la base $b \in \{D, G, A, T (U), C\}$ en el conjunto de secuencias alineadas analizadas.

En la variante *ii* las secuencias se generan de acuerdo a una distribución de probabilidad discreta $\{p_D, p\}$, donde p_D tiene el significado ya mencionado y se determina análogamente a como se describió en el párrafo anterior. En este caso p simboliza la probabilidad de aparición del resto de las bases {G, A, U, C} en las secuencias generadas, la cual será la misma para cada una de estas bases y se determina a través de la expresión $p = (1- p_D) /4$, por lo que se satisface que $p_D + p + p + p + p = 1$.

En la variante *iii* las secuencias se generan con distribución de uniforme, o sea, cada base tiene igual probabilidad de aparición en las secuencias generadas, al ser cinco bases esta probabilidad es 1/5.

2.4. Conclusiones parciales del capítulo

Se desarrolla una nueva herramienta para el análisis comparativo de secuencias genómicas. El método se puede desglosar en tres etapas principales:

1. Recodificación en $\mathbb{Z}_p$ de las secuencias genómicas representadas en $\mathbb{Z}_5$.
2. Obtención de los espectros de potencia de las secuencias recodificadas en $\mathbb{Z}_p$
3. Comparación de los principales picos espectrales inducidos por la recodificación mediante el ANOVA bifactorial no paramétrico.

Además, se proponen tres distribuciones para la generación aleatoria de secuencias de nucleótidos que serán fundamentales en la discusión y análisis de los resultados del método propuesto

3. RESULTADOS Y DISCUSIÓN

Los procedimientos expuestos en el capítulo anterior conducen a un método general para realizar el análisis comparativo de los espectros de potencia obtenidos a partir de las secuencias de ADN genómico. Las etapas requeridas para la aplicación de dicho método se pueden representar como se muestran en la Figura 3.1. En este capítulo se aplica la nueva herramienta desarrollada a secuencias genómicas mitocondriales de mamíferos.

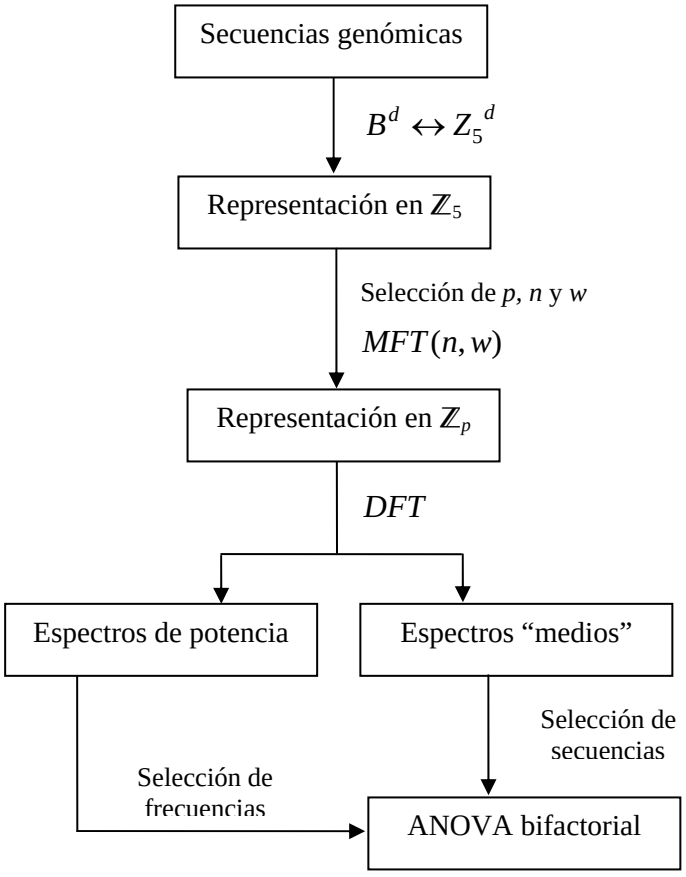

Figura 3.1. Esquema general del método propuesto para el análisis comparativo de espectros de potencia de las secuencias de ADN.

3.1. Descripción de la base de datos

Un conjunto de secuencias codificantes para proteínas mitocondriales fue construido concatenando 13 genes de 14 mamíferos Estas secuencias, conjuntamente con los 14 genomas mitocondriales completos de los mismos mamíferos, fueron utilizados como material de para esta investigación. Los datos de las secuencias genómicas de partida fueron tomadas de la base de datos del NCBI (http://www.ncbi.nlm.nih.gov/). Los organismos utilizados con sus nombres comunes y científicos y el número de acceso en la base son: humano (*Homo sapiens*, gi|115315570), chimpancé (*Pan troglodytes*, gi|5835121), chimpancé pigmeo (*Pan paniscus*, gi|5835135), gorila (*Gorilla gorilla*, gi|5835149, orangután de Sumatra (*Pongo pygmaeus abelii*, gi|5835834), orangután (*Pongo pygmaeus*, gi|5835163), gibbon común (*Hylobates lar*, gi|5835820); toro (*Bos taurus*, gi|60101824), ballena de aleta dorsal (*Balaenoptera physalus*, gi|5819095). Ballena azul (*Balaenoptera musculus*, gi|5834995); foca de Harbor (*Phoca vitulina*, gi|5834857), foca gris (*Halichoerus grypus*, gi|5835009); ratón (*Mus musculus*, gi|34538597), zarigüeya (*Didelphis virginiana*, gi|5835037). Por razones de simplicidad estos organismos serán referidos utilizando su nombre común en español y no su nombre científico.
El alineamiento múltiple de las 14 secuencias codificantes para proteínas y de los 14 genomas mitocondriales completos fueron obtenidos con el MEGA 4. Los gaps introducidos durante el proceso de alineamiento fueron reemplazados por la letra D, y las secuencias fueron representadas como se describe en la sección 1.3.3.

3.2. Espectros de potencia de secuencias naturales

En la Figura 3.2.1 se muestra los espectros de Fourier de las secuencias codificantes para proteínas mitocondriales de los 14 mamíferos. Las secuencias fueron recodificadas en $\mathbb{Z}_{17}$ (p = 17) mediante la MFT con raíz primitiva 2 de orden 8. Nótese que la recodificación con la MFT tiende a inducir picos espectrales a frecuencias múltiplos de 1/8. Sin embargo, no todos los picos múltiplos de 1/8 son igualmente significativos y las magnitudes de estos dependen de la naturaleza de la secuencia analizada. En el presente caso solo tres picos resultan ser los más significativos, encontrados a frecuencias: 1/4, 1/2 y 3/4. La recodificación con la combinación de parámetros p=11, orden 5 y w=3 induce picos múltiplos de 1/5 (Figura 3.2.2).

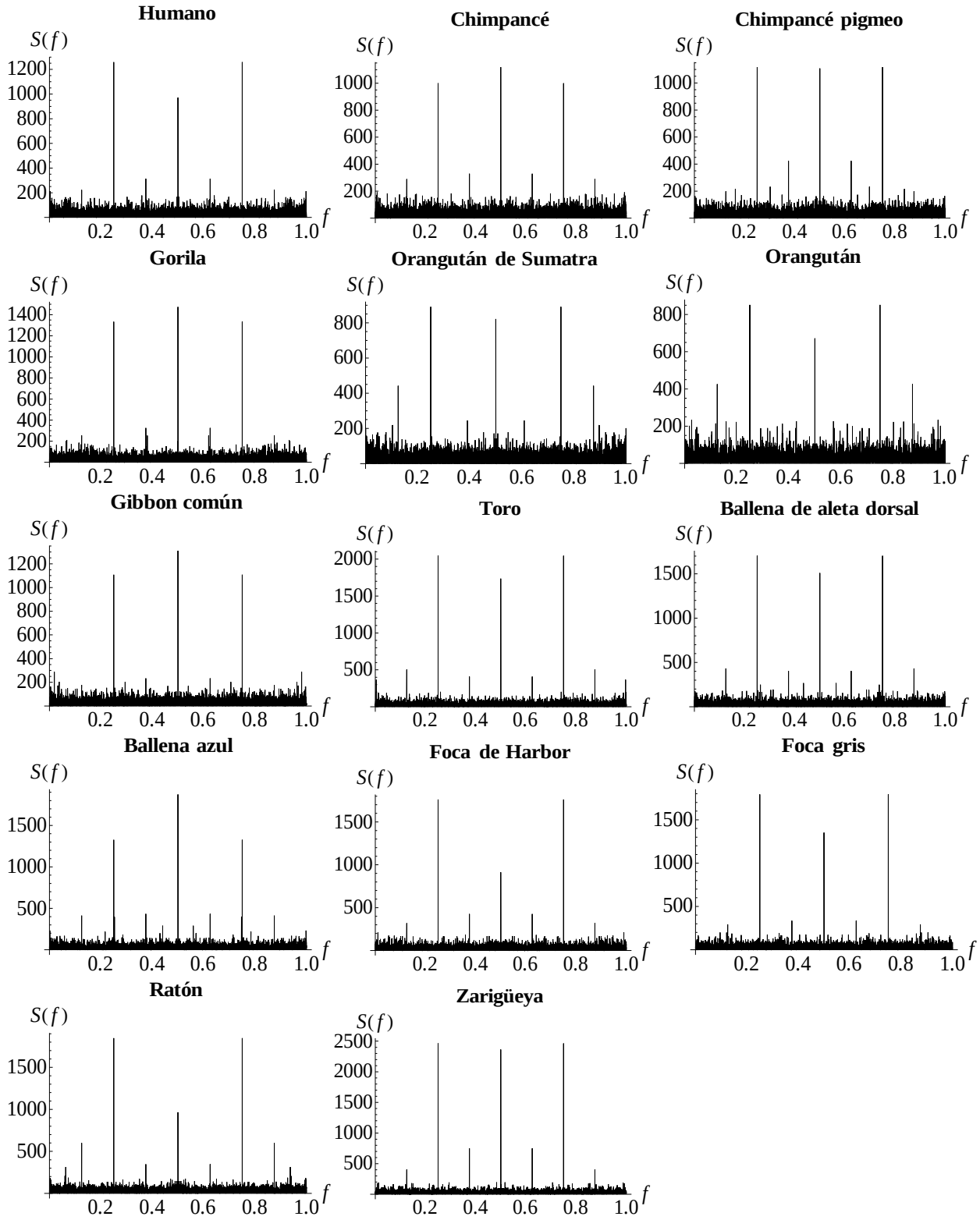

Figura 3.2.1. Espectros de potencia de las regiones codificantes para proteínas de los 14 mamíferos estudiados obtenidos para la recodificación p=17, orden 8 y w=2.

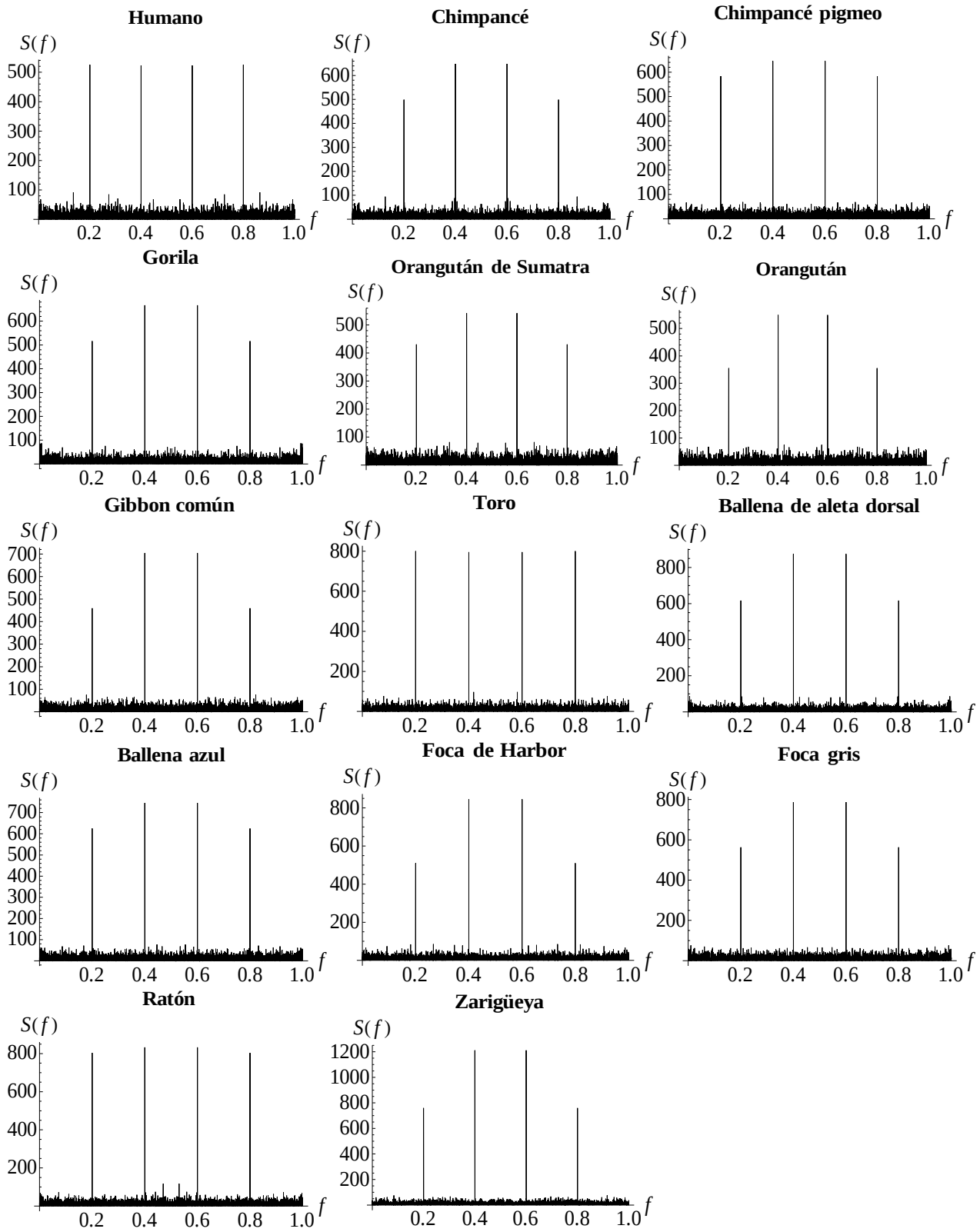

Figura 3.2.2. Espectros de potencia de las regiones codificantes para proteínas de los 14 mamíferos estudiados obtenidos para la recodificación p=11, orden 5 y w=3.

Nótese que las variabilidades observadas en las magnitudes de los espectros de potencia dependen de la combinación de parámetros utilizada. Dado que el número de combinaciones a probar es bastante elevado y no contamos con una teoría que nos permita decidir cuales combinaciones de parámetros pueden resultar apropiadas, se puede determinar tales combinaciones a partir de la aplicación de algún diseño experimental. Sin embargo, la realización de tal diseño está fuera del marco de esta tesis.

En el Anexo 2 se muestran los espectros de potencia de los genomas completos para las mismas combinaciones de parámetros utilizadas en las secuencias codificantes.

3.3. Comparaciones entre espectros de potencia de secuencias naturales

La comparación de las secuencias se realiza utilizando el ANOVA bifactorial no-paramétrico tal como se describe en la sección 2.3.1 y el Anexo 2. En la comparación entre varios niveles del factor secuencia es suficiente que existan diferencias estadísticamente significativas entre un par de secuencias para que el ANOVA detecte diferencias. Por tal motivo, las comparaciones más interesantes desde el punto de vista biológico-práctico son las par a par. En la Tabla 3.3.1 se muestra la matriz de comparaciones par a par entre las secuencias de genes concatenados codificantes para proteínas, utilizando la combinación de parámetros con p=17, w=2 y orden 8. Los números en las casillas representan el valor de la significación estadística del factor secuencia como resultado de la aplicación del ANOVA a cada par de secuencias. Solo se muestran las significaciones menores que 0.05, o sea, aquellas comparaciones en las que se detectaron diferencias estadísticamente significativas entre las secuencias comparadas. Las casillas en blanco corresponden a comparaciones en las que la significación fue superior a 0.05 y por tanto el ANOVA no asegura la existencia de diferencias estadísticamente significativas.

En el análisis comparativo de dos espectros de potencia utilizamos, además, comparaciones que incluyen un tercer espectro tomado como referencia, el cual puede ser obtenido a partir de una secuencia natural o a partir de secuencias generadas aleatoriamente. Este hecho nos conduce a definir los conceptos de comparaciones directas e indirectas. Entendemos por comparación directa de dos espectros de potencia aquella en la cual la detección o no de diferencias es derivada directamente de la significación estadística de la prueba del ANOVA, ya sea del factor secuencia o de la interacción secuencia-frecuencia. Por ejemplo, en la Tabla

3.3.1 la comparación directa arroja que existen diferencias estadísticamente significativas entre las secuencias del Humano y el Orangután (Sig.=0.015) y que no son detectadas diferencias significativas entre las secuencias del Humano y el Chimpancé (casilla en blanco).

Entendemos por comparación indirecta de dos espectros de potencia aquella en la cual las diferencias entre dichos espectros se detectan indirectamente al incluir un tercer espectro, tomado como referencia, el cual presenta diferencias estadísticamente significativas en la comparación directa con solo uno de los espectros analizados. Por ejemplo, la comparación indirecta detecta diferencias entre las secuencias del Humano y el Chimpancé, pues la secuencia del Orangután es estadísticamente diferente a la primera y no a la segunda según las comparaciones directas.

Se debe destacar que, el hecho de que no se detecten diferencias significativas no implica que tales diferencias no existan, sino que no se tienen razones estadísticas suficientes que permitan detectar las posibles diferencias mediante la aplicación de la prueba utilizada. La detección de diferencias implica la existencia de arquitecturas organizativas diferentes entre las secuencias comparadas. Sin embargo, dos secuencias de bases pueden tener altos por cientos de identidad detectados en el alineamiento par a par de ambas y simultáneamente tener diferentes arquitecturas organizativas. Por ejemplo, en la Tabla 3.3.1 se aprecia que existen diferencias estadísticamente significativas entre los espectros de potencia del Orangután y del Humano, mientras que el ANOVA no detecta diferencias significativas entre los espectros del Orangután y del Chimpancé. Sin embargo, en la Tabla 3.3.2 el número de bases diferentes por sitio entre las secuencias de ADN indica que las diferencias entre las composiciones de bases del Humano y del Orangután son menores que las diferencias entre el Orangután y el Chimpancé. Este ejemplo es indicador de que el método de análisis espectral propuesto en esta investigación no está dirigido a la detección de las diferencias en las composiciones de bases de las secuencias de ADN, tal y como se deriva de la comparación de las secuencias alineadas. En su lugar, es una evidencia más de que el método se refiere a la detección de diferencias en la arquitectura organizativa de las secuencias comparadas.

El uso de las comparaciones indirectas permite señalar que existen diferencias entre las arquitecturas organizativas de las regiones codificantes para proteínas del Gorila y el resto de los primates (los seis primeros organismos que aparecen en la tabla), tomando como referencia la secuencia del Toro. En la Tabla 3.3.1 apreciamos que existen diferencias estadísticamente

significativas entre el espectro del Toro y los espectros de los primates exceptuando al espectro del Gorila. Dada la estrecha relación filogenética entre los primates y dado el hecho de que la diferencia fue detectada respecto al Toro, sugiere que la diferencia no detectada entre los espectros del Toro y el Gorila es debida a un rasgo presente en la arquitectura organizativa del ancestro común de todos los primates. Luego, la diferencia detectada con el resto de los primates debe referirse a la pérdida del rasgo mencionado en la arquitectura organizativa, lo cual debió ocurrir después de la formación de las especies de primates analizadas. Este análisis destaca el potencial de la herramienta desarrollada en la investigación filogenética. Nótese que la detección de diferencias depende de la combinación de parámetros utilizada. Es decir, las comparaciones directas e indirectas entre dos especies para una combinación dada de parámetros (*p*, *w* y orden) pueden detectar diferencias estadísticamente significativas, mientras que para otra combinación no se detecten. Este hecho no conduce a contradicciones pues cada combinación de parámetros permite analizar características diferentes de la arquitectura organizativa de las secuencias de ADN. Además, el método no permite conocer cuales son las diferencias o dónde radican, sólo indica la existencia de las mismas.

Tabla 3.3.1. Resultados del ANOVA bifactorial no paramétrico para secuencias codificantes para proteínas.[a]

Significación	Humano	Chimpancé	Chimpancé pigmeo	Gorila	Orangután de Sumatra	Orangután	Gibbon común	Toro	Ballena de aleta dorsal	Ballena azul	Foca de Harbor	Foca gris	Ratón	Zarigüeya
Humano						0.015		0.019						0.002
Chimpancé								0.012						0.001
Chimp. pigmeo								0.012						0.003
Gorila														0.007
Orangután de S.								0.004	0.015			0.031	0.047	0.001
Orangután	0.015						0.038	0.002	0.001	0.031		0.005	0.012	0.000
Gibbon común						0.038		0.019						0.002
Toro	0.019	0.012	0.012		0.004	0.002	0.019							
Ballena de a. d.					0.015	0.001								0.024
Ballena azul						0.031								0.038
Foca de Harbor														0.031
Foca gris					0.031	0.005								0.019
Ratón					0.047	0.012								
Zarigüeya	0.002	0.001	0.003	0.007	0.001	0.000	0.002		0.024	0.038	0.031	0.019		

[a] La recodificación de las secuencias fue realizada utilizando la combinación de parámetros: *p*=17, *w*=2 y orden 8.

Tabla 3.3.2. Número de bases diferentes por sitio entre secuencias codificantes para proteínas de seis primates.[a]

	Humano	Chimpancé	Chimp. pigmeo	Gorila	Orangután de S.	Orangután
Humano		0.098	0.096	0.118	0.155	0.158
Chimpancé	0.098		0.043	0.116	0.161	0.161
Chimp. Pigmeo	0.096	0.043		0.113	0.159	0.159
Gorila	0.118	0.116	0.113		0.162	0.161
Orangután de S.	0.155	0.161	0.159	0.162		0.071
Orangután	0.158	0.161	0.159	0.161	0.071	

[a] Todos los resultados son calculados a partir del alineamiento par a par de las 6 secuencias. Los sitios conteniendo gaps fueron eliminados. Se consideraron un total 11352 de bases.

En la Tabla 3.3.3 se muestra la matriz de comparaciones par a par entre las 14 secuencias de genomas que forman parte del estudio realizado, utilizando la combinación de parámetros con p=11, w=3 y orden 5. Debe notarse que entre los espectros de potencia de los primates no se detectan diferencias significativas a partir de comparaciones directas. Sin embargo, al observar el alineamiento múltiple entre las 14 secuencias (Figura 3.3.1), se notan entre ellos grandes regiones no codificantes para proteínas donde aparecen mutaciones indel (representadas por la letra D) que provocan diferencias marcadas en la composición de bases. Esto es un indicador de que dichas mutaciones en los genomas, no ocurren en posiciones cualesquiera, sino en aquellas que sean capaces de preservar la estructura organizativa de las bases de los genomas. Se observa además, a partir de estas comparaciones, que entre los espectros de los primates y los espectros de los seis últimos organismos (excepto la Foca gris) se detectan diferencias estadísticamente significativas.

Las comparaciones indirectas muestran un resultado que no es posible obtener en los espectros de las secuencias codificantes, es la detección de diferencias entre los espectros de las secuencias del Chimpancé y el Chimpancé pigmeo, si se toma como referencia el espectro de la secuencia de la Foca gris. Notemos, además que no se detectaron diferencias estadísticamente significativas entre los espectros de la Ballena azul y los primeros cuatro primates y el Gibbon común, lo cual nos sugiere la conservación de rasgos comunes en la arquitectura genómica heredada del ancestro común de todos estos mamíferos, exceptuando los orangutanes. La diferencia detectada con los orangutanes permite realizar nuevas comparaciones indirectas.

Al comparar las Tablas 3.3.1 y 3.3.3 es notable el mayor número de diferencias en comparaciones directas que son detectadas entre los espectros de las secuencias de genomas con relación a las detectadas entre los espectros de las secuencias codificantes. Esto es de esperar, ya que en genomas completos las mutaciones indel pueden ser más frecuentes debido a que ellos contienen, además de las regiones codificantes para proteínas, regiones codificantes para ARN y regiones no codificantes. Debido a que en las secuencias codificantes para proteínas la pérdida de algún rasgo en el proceso de formación de las especies tiende a ser menor por el efecto de la presión evolutiva y las diferencias entre las arquitecturas genómicas de las especies cercanas tienden a ser mucho menores que entre especies lejanas. Es decir, la no detección de diferencias entre los espectros es consecuencia de que existe la tendencia de conservar la estructura organizativa de las regiones codificantes para proteínas, permitiendo la preservación de sus funciones biológicas.

Tabla 3.3.3. Resultados del ANOVA bifactorial no paramétrico para secuencias de genomas.[a]

Significación	Humano	Chimpancé	Chimpancé pigmeo	Gorila	Orangután de Sumatra	Orangután	Gibbon común	Ballena azul	Ballena de aleta dorsal	Foca gris	Foca de Harbor	Toro	Ratón	Zarigüeya
Humano									0.006	0.038	0.021	0.006	0.011	0.000
Chimpancé									0.038	0.038	0.021	0.006	0.006	0.000
Chimp. pigmeo									0.038		0.038	0.021	0.006	0.000
Gorila									0.006	0.021	0.038	0.006	0.006	0.000
Orangután de S.								0.001	0.001	0.001	0.003	0.001	0.001	0.000
Orangután								0.011	0.001	0.001	0.003	0.001	0.000	0.000
Gibbon común									0.011	0.011	0.011	0.003	0.006	0.000
Ballena azul					0.001	0.011			0.038			0.038	0.021	0.000
Ballena de a. d.	0.006	0.038	0.038	0.006	0.001	0.001	0.011	0.038						0.000
Foca gris	0.038	0.038		0.021	0.001	0.001	0.011							0.000
Foca de Harbor	0.021	0.021	0.038	0.038	0.003	0.003	0.011							0.011
Toro	0.006	0.006	0.021	0.006	0.001	0.001	0.003	0.038						0.021
Ratón	0.011	0.006	0.006	0.006	0.001	0.000	0.006	0.021						0.003
Zarigüeya	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.021	0.003	

[a] La recodificación de las secuencias fue realizada utilizando la combinación de parámetros: p=11, w=3 y orden 5.

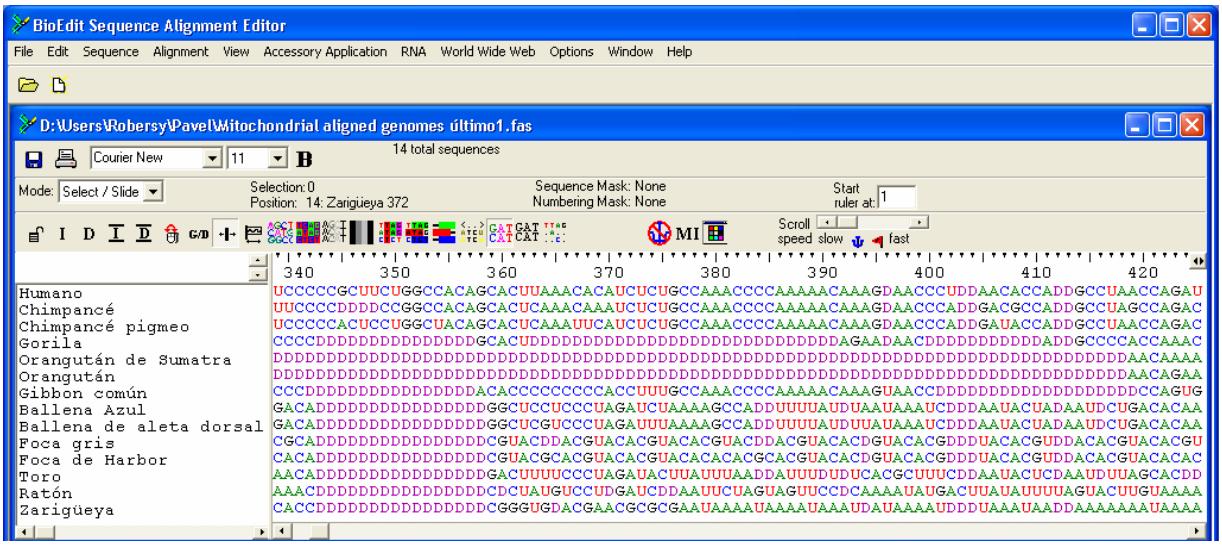

Figura 3.3.1. Región no codificante desde la 339 a la 425 de los genomas mitocondriales de los mamíferos estudiados. Se aprecia la alta frecuencia de las mutaciones indel entre los primates a partir del Gorila.

3.4. Comparaciones dos a dos entre espectros de potencia de secuencias naturales y espectros de potencia de secuencias generadas aleatoriamente

La era genómica reduce cada vez más las limitaciones en cuanto al tamaño y el número de secuencias genómicas disponibles en las bases de datos biológicas. Este hecho permite trabajar con tamaños de muestras a partir de los cuales se puede estimar las distribuciones de las bases de nucleótidos y la consecuente generación aleatoria de secuencias empleando dicha distribución. En este trabajo, el uso de los 14 genomas mitocondriales permitió realizar la estimación de tal distribución.

La generación de secuencias aleatorias incrementa sustancialmente el potencial de las comparaciones indirectas de los espectros de potencia. En particular, aquí se proponen tres tipos de comparaciones:

i. Entre secuencias naturales y secuencias generadas con distribución discreta $\{p_D, p_G, p_A, p_U, p_C\}$.

ii. Entre secuencias naturales y secuencias generadas con distribución discreta $\{p_D, p\}$.

iii. Entre secuencias naturales y secuencias generadas con distribución uniforme.

Cada una de estas comparaciones es discutida a continuación.

3.4.1. Espectros de potencia de secuencias naturales versus espectros de potencia de secuencias generadas con distribución $\{p_D, p_G, p_A, p_U, p_C\}$

La distribución de bases $\{p_D, p_G, p_A, p_U, p_C\}$ fue estimada con la muestra de secuencias codificantes para proteínas de los cuatro primeros de primates, con un total de 4×11502 bases.
En las Tablas 3.4.1.1-4 aparecen el número de secuencias de referencia que posibilitan la detección de diferencias en comparaciones indirectas entre los cuatro primeros primates. Las secuencias de referencia son obtenidas entre 1000 secuencias generadas aleatoriamente con la distribución de bases descrita, después de realizar la prueba del ANOVA bifactorial no paramétrico. Se muestran los resultados para las secuencias codificantes y para los genomas de los primates analizados, cada caso se ilustra con dos recodificaciones diferentes. Para cada tabla las muestras generadas de 1000 secuencias son diferentes.
Es conocido el carácter pseudoaleatorio de las secuencias genómicas; luego la detección de un mayor o menor número de comparaciones indirectas, entre cualquier par de secuencias, indica una mayor o menor preservación de rasgos en el carácter pseudoaleatorio de las secuencias naturales referido a las secuencias generadas. En la Tabla 3.4.1.1 se muestran las comparaciones indirectas de los espectros de potencia de secuencias naturales para p=11, orden 5 y w=3 con 1000 secuencias. Para esta combinación de parámetros se encuentra una situación muy interesante desde el punto de vista filogenético. Resulta que las diferencias encontradas en las 1000 comparaciones entre los espectros de potencias del Humano y del Gorila son menores que las encontradas entre el Humano y los chimpancés, lo cual resulta sorprendente pues el Humano y los chimpancés son filogenéticamente más cercanos que el Humano y el Gorila. Este hecho nos sugiere que las diferencias detectadas no implican cercanías o distanciamiento filogenético entre las especies comparadas sino, precisamente, la existencia de rasgos pseudoaleatorios diferentes en la arquitectura organizativa de dichas especies. Este hecho no resulta extraño pues, por ejemplo, dos edificaciones de nuestra universidad pueden tener el mismo estilo arquitectónico y, sin embargo, ser construcciones muy diferentes. De manera semejante los rasgos en la arquitectura organizativa de las especies pueden conservarse o desaparecer durante el proceso evolutivo, tal como fue comentado en la sección anterior. En la Tabla 3.4.1.2 se muestran las comparaciones para la combinación de parámetros p=17, w=2, *y* orden 8. En este caso, las diferencias detectadas entre el Humano y

el Gorila son mayores que las detectadas entre el Humano y los chimpancés. Sin embargo, las diferencias detectadas entre el Gorila y los chimpancés son mayores que las detectadas entre el Gorila y el Humano, lo cual nos indica que durante el proceso de formación de las especies de primates a partir de un ancestro común, el proceso de especiación imprimió huellas individuales en las características pseudoaleatorias de cada especie. Notemos, en las Tablas 3.4.1.1 y 3.4.1.2, que las comparaciones indirectas permiten detectar, incluso, diferencias entre especies muy cercanas como el Chimpancé y el Chimpancé pigmeo.

En las Tablas 3.4.1.3 y 3.4.1.4 se muestran las diferencias detectadas en las comparaciones indirectas entre los espectros de potencia de obtenidos para los genomas completos de los primates y 1000 secuencias aleatorias utilizando la combinación de parámetros: {p=17, orden 8 y w=2} y {p=11, orden 5 y w=3}. Ante todo podemos notar que el número de diferencias detectadas en las comparaciones indirectas de los espectros de potencia obtenidos en los genomas completos son mucho menores que las detectadas en los espectros de las regiones codificantes para proteínas. Este resultado es indicativo de que el carácter pseudoaleatorio de las secuencias genómicas está expresado más intensamente que el de las secuencias codificantes para proteínas. Pero este hecho, en cierta forma ya discutido en la sección anterior, es de esperar pues la presión evolutiva en las regiones codificantes para proteínas es mucho mayor que en las regiones no codificantes para genes, las cuales están presentes en los genomas completos. Es decir las regiones codificantes para genes poseen más libertad para que las mutaciones, inducidas aleatoriamente por el medio ambiente, sean aceptadas en la población de organismos.

Tabla 3.4.1.1. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias codificantes para proteínas generadas aleatoriamente con distribución {p_D, p_G, p_A, p_U, p_C}[a]. Recodificación para p=11, w =3 y orden 5.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		446	592	183
Chimpancé	14		159	23
Chimp. pigmeo	3	2		4
Gorila	23	295	433	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas (p < 0.05) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución {p_D, p_G, p_A, p_U, p_C} y no arrojaron diferencias estadísticamente significativas con la especie *j*.

Tabla 3.4.1.2. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias codificantes para proteínas generadas aleatoriamente con distribución $\{p_D, p_G, p_A, p_U, p_C\}$[a]. Recodificación para p=17,w=2, y orden 8.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		37	204	482
Chimpancé	144		280	589
Chimp. pigmeo	32	1		310
Gorila	0	0	0	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas ($p < 0.05$) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución $\{p_D, p_G, p_A, p_U, p_C\}$ y no arrojaron diferencias estadísticamente significativas con la especie *j*.

Tabla 3.4.1.3. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias de genomas generadas aleatoriamente con distribución $\{p_D, p_G, p_A, p_U, p_C\}$[a]. Recodificación para p=17,w=2, y orden 8.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		84	82	211
Chimpancé	163		73	286
Chimp. pigmeo	113	25		238
Gorila	5	1	1	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas ($p < 0.05$) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución $\{p_D, p_G, p_A, p_U, p_C\}$ y no arrojaron diferencias estadísticamente significativas con la especie *j*.

Tabla 3.4.1.4. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias de genomas generadas aleatoriamente con distribución $\{p_D, p_G, p_A, p_U, p_C\}$[a]. Recodificación para p=11, w =3 y orden 5.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		40	77	32
Chimpancé	138		151	115
Chimp. pigmeo	75	51		66
Gorila	47	32	83	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas ($p < 0.05$) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución $\{p_D, p_G, p_A, p_U, p_C\}$ y no arrojaron diferencias estadísticamente significativas con la especie *j*.

En este trabajo se ilustra un modesto ejemplo al generar solo 1000 secuencias aleatorias, por razones de tiempo y espacio en memoria. Sin embargo, la generación aleatoria nos permite trabajar con un número mayor y proponer, incluso, pruebas de hipótesis que nos permitan la verificación directa de hipótesis biológicas referidas a la arquitectura de las secuencias naturales comparadas.

3.4.2. Espectros de potencia de secuencias naturales versus espectros de potencia de secuencias generadas con distribución {p_D, p}

La comparación indirecta entre los espectros de potencia de los cuatro primeros primates y de las secuencias generadas aleatoriamente con distribución {p_D, p} sugiere que las mutaciones de inserciones y deleciones observadas en las secuencias codificantes para proteínas no tienen lugar arbitrariamente sino que tienden a preservar la arquitectura organizativa de los genomas. En las Tablas 3.4.2.1 y 3.4.2.2 se presentan estas comparaciones para las combinaciones de parámetros {p=17, orden 8 y w=2} y {p=11, orden 5 y w=3} respectivamente.

En las Tablas 3.4.2.3 y 3.4.2.4 se muestran las diferencias detectadas en las comparaciones indirectas de los espectros de potencia de las secuencias genómicas completas con las secuencias aleatorias generadas con las combinaciones antes mencionadas. El carácter fuertemente pseudoaleatorio de las secuencias genómicas completas en comparación con las regiones codificantes para proteínas se aprecia al contrastar las Tablas 3.4.2.1 y 3.4.2.2 con las Tablas 3.4.2.3 y 3.4.2.4. El número de diferencias detectadas en las 1000 comparaciones es notablemente menor para los genomas completos (Tablas 3.4.2.3 y 3.4.2.4) que las detectadas en las secuencias codificantes para proteínas (Tablas 3.4.2.1 y 3.4.2.2).

En la Figura 3.4.2.1 se muestran los espectros de potencias obtenidos para secuencias generadas con la distribución de bases de las secuencias codificantes para proteínas de los cuatro primeros primates (con la combinación de parámetros p= 17, orden 8 y w = 2). Resulta notable que los espectros de potencia generados con distribuciones {p_D, p_G, p_A, p_U, p_C} y {p_D, p} poseen patrones espectrales semejantes a los observados en la Figura 3.2.1. Basta fijar la frecuencia de la base D para imponer un patrón espectral semejante al encontrado en las secuencias naturales.

Tabla 3.4.2.1. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias codificantes para proteínas generadas aleatoriamente con distribución $\{p_D, p\}$[a]. Recodificación para p=17,w=2, y orden 8.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		43	179	464
Chimpancé	120		223	541
Chimp. pigmeo	35	2		3 20
Gorila	0	0	0	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas ($p < 0.05$) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución $\{p_D, p\}$ y no arrojaron diferencias estadísticamente significativas con la especie *j*.

Tabla 3.4.2.2. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias codificantes para proteínas generadas aleatoriamente con distribución $\{p_D, p\}$[a]. Recodificación para p=11, w =3 y orden 5.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		32	12	12
Chimpancé	61		19	30
Chimp. pigmeo	258	236		180
Gorila	115	104	37	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas ($p < 0.05$) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución $\{p_D, p\}$ y no arrojaron diferencias estadísticamente significativas con la especie *j*.

Tabla 3.4.2.3. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias de genomas generadas aleatoriamente con distribución $\{p_D, p\}$[a]. Recodificación para p=17,w=2, y orden 8.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		88	45	23
Chimpancé	33		24	30
Chimp. pigmeo	16	50		25
Gorila	134	196	165	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas ($p < 0.05$) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución $\{p_D, p\}$ y no arrojaron diferencias estadísticamente significativas con la especie *j*.

Tabla 3.4.2.4. Número de comparaciones indirectas encontradas entre 1000 espectros de potencias derivados de secuencias de genomas generadas aleatoriamente con distribución {p_D, p}[a]. Recodificación para p=11, w =3 y orden 5.

	Humano	Chimpancé	Chimp. pigmeo	Gorila
Humano		73	16	6
Chimpancé	1		0	2
Chimp. pigmeo	4	60		6
Gorila	3	71	15	

[a] En la celda *ij* se encuentra el número de comparaciones indirectas de la especie *i* que arrojaron diferencias estadísticamente significativas ($p < 0.05$) en 1000 comparaciones con espectros de potencia de secuencias generadas aleatoriamente con distribución {p_D, p} y no arrojaron diferencias estadísticamente significativas con la especie *j*.

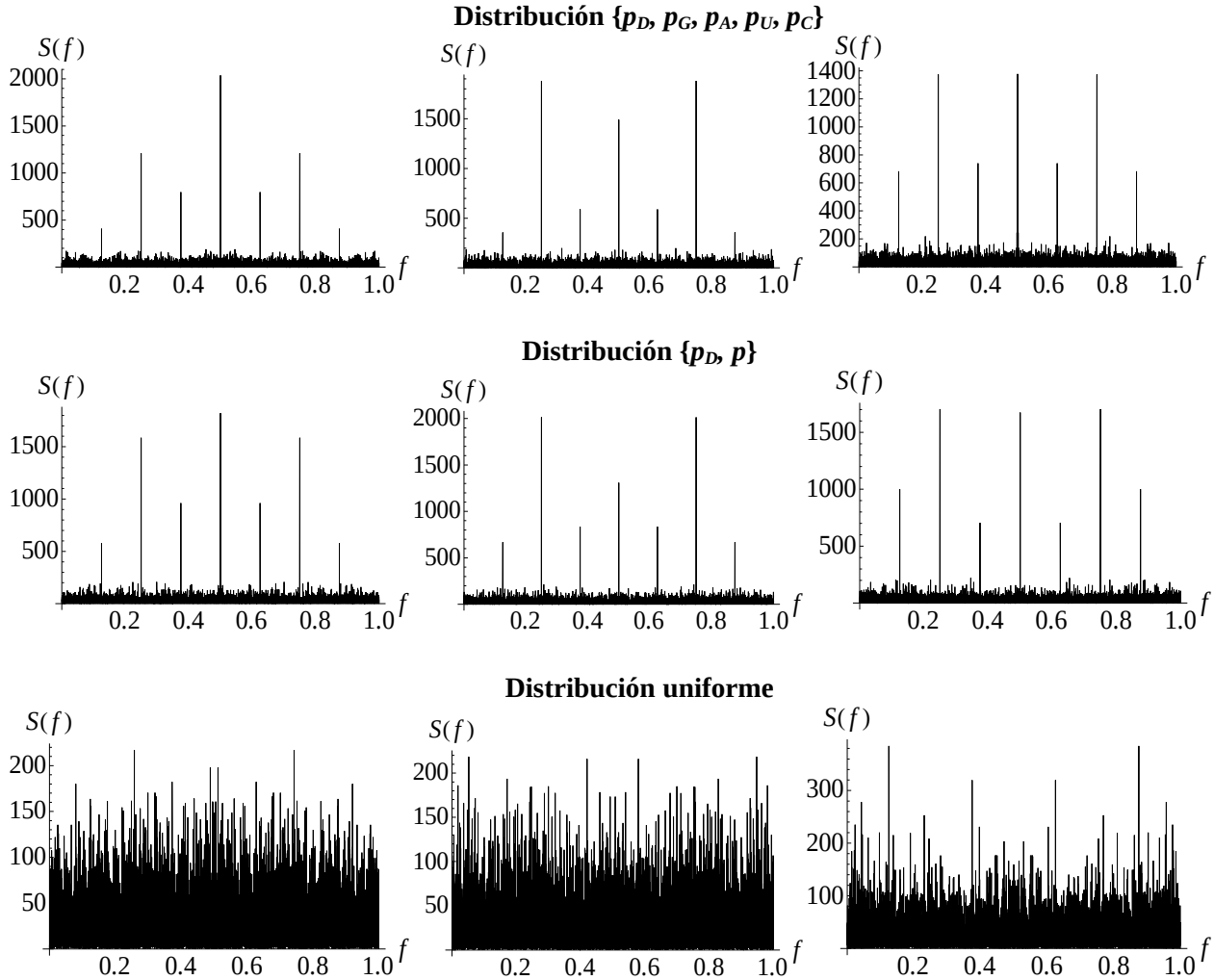

Figura 3.4.2.1. Espectros de potencia obtenidos para secuencias generadas con las tres distribuciones de bases descritas en las secciones 2.2.3. Las distribuciones de base utilizadas corresponden a las distribuciones de las secuencias codificantes.

Sin embargo, las variaciones en las magnitudes de los picos espectrales dependen de las distribuciones utilizadas pues las comparaciones indirectas arrojan mayor número de diferencias cuando se utiliza la distribución $\{p_D, p_G, p_A, p_U, p_C\}$ (Tablas 3.4.1.1 y 3.4.1.2) que cuando se utiliza la distribución $\{p_D, p\}$ (Tablas 3.4.2.1 y 3.4.2.2).

3.4.3. Espectros de potencia de secuencias naturales versus espectros de potencia de secuencias generadas con distribución uniforme

Todos los espectros de potencia de secuencias generadas con distribución uniforme de las bases mostraron diferencias estadísticamente significativas con las 14 secuencias de naturales de mamíferos, las codificantes para proteínas y las de genoma completo. Este hecho es de esperar debido a que las secuencias naturales no son totalmente aleatorias: Es decir, aunque las secuencias naturales muestran un caracter seudoaleatorio no presentan un orden arbitrario en su arquitectura. Este hecho nos evidencia que la presencia y magnitud de los picos depende de la naturaleza de la secuencias y marca una distinción sustancial entre las secuencias naturales y las no naturales. En particular, este hecho nos sugiere que las distribución de las bases del ADN y de las mutaciones indel en las secuencias naturales induce rasgos distintivos en sus arquitecturas organizativas que marcan la diferencia con las característica encontradas en las secuencias generadas con distribución uniforme de las bases. En la Figura 3.4.2.1 se puede ver que las magnitudes de los picos observados en las secuencias generadas con distribución uniforme están al nivel del ruido observado en las secuencias naturales (ver Figuras 3.2.1 y 3.2.2) y en las secuencias generadas con distribuciones $\{p_D, p_G, p_A, p_U, p_C\}$ y $\{p_D, p\}$. Estos resultados evidencian una vez más que el método de comparación desarrollado refleja rasgos pseudoaleatorios característicos de las arquitecturas organizativas de las secuencias naturales que están determinados, en gran medida por la distribución de bases encontradas en las secuencias.

3.5. Conclusiones parciales del capítulo

Se expusieron ejemplos concretos del método desarrollado en el capítulo anterior para la comparación espectral de secuencias genómicas. Se mostró la utilidad de esta herramamienta en el análisis filogenético y en la detección de la existencia de diferencias estadísticamente

significativas en la arquitectura organizativa de las secuencias analizadas, a través de comparaciones directas e indirectas. Se encontró evidencia de que el carácter pseudoaleatorio de las secuencias se expresa más intensamente en los genomas completos que en las regiones codificantes para proteínas. Los espectros de potencia de secuencias generadas con distribución uniforme mostraron marcadas diferencias con los espectros de secuencias naturales, lo que demuestra que la información contenida en las secuencias define una estructura organizativa que no es al azar. Es decir, la distribución de bases en las secuencias naturales está estrechamente relacionada con las funciones biológicas que determinan.

CONCLUSIONES

- Se implementaron exitosamente la transformada “clásica” y el algoritmo FFT sobre campos finitos para su uso en la recodificación de secuencias genómicas, lo cual induce la aparición de nuevas armónicas en los espectros de potencia.
- Se desarrolló un método que permite el análisis comparativo de las secuencias genómicas recodificadas. El método comprende la aplicación sucesiva de las transformadas modular y discreta de Fourier a las secuencias de ADN. Posteriormente, se realizó la comparación de los respectivos espectros de potencias mediante el uso del ANOVA bifactorial no parámetrico.
- La aplicación del método en el análisis filogenético de las secuencias de ADN alineadas mostró su utilidad, tanto en la comparación directa e indirecta de los espectros de potencias. De manera que es posible detectar, incluso, diferencias estadísticamente significativas entre los espectros de especies bastante cercanas. Además, se deja una puerta abierta para el desarrollo de futuras pruebas de hipótesis que pueden ser muy útiles en el análisis filogenético.

Relacionado con las dos conclusiones anteiores, se proponen además tres distribuciones para la generación aleatoria de secuencias de nucleótidos que fueron fundamentales en la discusión y análisis de los resultados del método propuesto.

RECOMENDACIONES

- Desarrollar la aplicación de métodos de filtrado de las señales recodificadas que permitan la eliminación de ruido en los espectros de potencias. Esto mejorará el análisis de regiones genómicas de pequeño tamaño.
- Puede ser posible el establecimiento de una heurística o algún diseño experimental que permita una selección más eficiente de los parámetros utilizados al aplicar la transformada modular de Fourier.
- Aplicar la metodología a regiones de genoma nuclear de organismos superiores.

REFERENCIAS

BERTMAN, M. O. & JUNGCK, J. R. (1979) Group graph of the genetic code. *The Journal of Heredety,* 70**,** 379-384.

BIRKHOFF, G. & MACLANE, S. (1941) *A survey of Modern Algebra*, The Macmillan Company, New York.

CANTILLO, J., RIVERON, A. M., KATRIB, M., BRINGAS, R., PEREZ, G., LOPEZ, L. & VALDES, F. (1986.) Utilización de un código binario en el diseño y programación de un SBDR para manipular cadenas de ADN. *Memorias del Segundo Seminario Cubano sobre Interferón y Primer Seminario Cubano sobre Biotecnología***,** 837-845.

CLIFTEN, P. (2004) The post-genomic era for a select few. *Genome Biology* 5**,** 308.1-308.2.

CONSORTIUM, T. C. S. A. A. & GENOME (2005) Initial sequence of the chimpanzee genome and comparison with the human. 437**,** 69-87.

CRICK, F. H. C. (1968) The origin of the genetic code. *J. Mol. Biol.,* 38**,** 367-379.

DUBREIL, P. & DUBREIL-JACOTIN, M. L. (1963) *Lecciones de álgebra moderna*, Editorial Reverté.

FICKETT, J. W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res,* 10**,** 5303-5318.

FUENTES, A. R., GINORI, J. V. L. & ÁBALO, R. G. (2006a) Comparison and use of different methods to convert DNA strings into numerical sequences to apply Digital Signal Processing Techniques.

FUENTES, A. R., GINORI, J. V. L. & ÁBALO, R. G. (2006b) Detection of Coding Regions in Large DNA Sequences Using the Short Time Fourier Transform with Reduced Computational Load.

FUENTES, A. R., GINORI, J. V. L. & ÁBALO, R. G. (2007) Coding Region Prediction in Genomic Sequences Using a Combination of Digital Signal Processing Approaches.

GAO, S. (2001) Fast Fourier Transforms and sparse linear systems over finite fields.

KOSTRIKIN, A. I. (1980) *Introducción al algebra*, Editorial MIR, Moscú.

MERCER-HURSH, T. (1968) Computer Analysis of Genetic Behavior at the Chromosomal Level. *American Anthropologist,* 70.

PENNISI, E. (2007) Genomicists Tackle
The Primate Tree. *SCIENCE,* 316 13.

PÉREZ, R. B. (1999) Predicción de la Estructura y la Función de dos Proteínas sobre la base del Análisis de sus Secuencias Aminoacídicas. *Subdirección de Información, Grupo de Bioinformática.* Centro de Ingeniería Genética y Biotecnología.

R.R.SOKAL & ROHLF, F. J. (1995) *Biometri.* Third edition ed., W. H. Freeman and Company, New York.

REDÉI, L. (1967) *Algebra,* Akadémiai Kiadó, Budapest.

RIVERON, A. M., BRINGAS, R., CANTILLO, J. & BAEZ, O. (1986) Codificación binaria del ADN restringida por las propiedades de las secuencias de bases. Su uso para la

determinar sitios de restricción. *Memorias del primer seminario cubano sobre Biotecnología*.

SANCHEZ, R. & GRAU, R. (2008) Extended genetic code vector space over the Galois field of five DNA bases alphabet. An hypothesis about the primeval genetic code. *In Press*.

SÁNCHEZ, R. & GRAU, R. (2008) Vector Space of the Extended Base-triplets over the Galois Field of five DNA Bases Alphabet. *Segundo Taller de Bioinformática IWOBI´08*.

SÁNCHEZ, R. & GRAU, R. (In Press) An algebraic hypothesis about the primeval genetic code architecture.

SANCHEZ, R., GRAU, R. & MORGADO, E. (2006) A novel Lie algebra of the genetic code over the Galois field of four DNA bases. *Mathematical Biosciences,* 202**,** 156–174.

SWANSON, R. (1984) A unifying concept for the amino acid code. *Bull. Math. Biol.,* 46**,** 187-203.

TAMURA K, D. J., M, N. & S, K. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol,* 24**,** 1596-1599.

TIWARI S, R., S, B., A, B. S. & R, R. (1997) Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS,* 13**,** 263-270.

TSONIS, A. A., ELSNER, J. B. & TSONIS, P. A. (1991) Periodicity in DNA coding sequence: implications in gene evolution. *J Theor. Biol,* 151**,** 323-331.

WAERDEN, B. L. V. D. (1970) *Algebra*, Ungar, New York.

YAP, C.-K. (2000) *Fundamentals Problems in Algorithmic Algebra*.

YU, U., LEE, S. H., KIM, Y. J. & KIM, S. (2004) Bioinformatics in the Post-genome Era. *Journal of Biochemistry and Molecular Biology,* 37**,** 75-82.

ANEXOS

Anexo 1. Descripción del ANOVA bifactorial no-paramétrico.

Es conocido, que hasta la Versión 15 del SPSS, y de otros paquetes de software de avanzada, no están implementadas las posibilidades de hacer Análisis de Varianza Multifactorial *no Paramétricos*, y mucho menos, Análisis de Varianza Multivariados y Multifactoriales *no Paramétricos*. Las alternativas clásicas para resolver problemas de este tipo se basan en el uso de Técnicas de Análisis de Datos Cualitativos (Categóricos). Si en particular la variable dependiente puede ser dicotomizada, o discretizada en categorías nominales, puede utilizarse la Regresión Logística Binaria o Multinomial que permite detectar efectos de factores principales o de interacciones. Cuando todas las variables predictivas son categóricas, se puede pensar en el uso de procedimientos Loglinear o Probit (ver ayuda del SPSS).

Sin embargo, hay problemas –como el que se plantea después como ejemplo- que son relativamente simples, porque se necesita un análisis de varianza con apenas dos factores en un diseño equilibrado, y las alternativas anteriormente mencionadas no producen resultados suficientemente buenos. Se regresa a la necesidad de hacer un análisis de varianza no paramétrico bifactorial, capaz de detectar la posible influencia de cada uno de los dos factores principales como su interacción.

Existe un fundamento teórico de cómo puede realizarse un tal análisis en el caso de diseños equilibrados que será comentado brevemente aquí. Más bien esta nota pretende ilustrar como lograr la materialización de esta teoría usando el paquete *Mathematica*. Comenzamos comentando un **ejemplo práctico** que forma parte de las pruebas que fueron realizadas en este trabajo y será de ayuda para ilustrar los **fundamentos teóricos y sus implicaciones prácticas.** Después concretamos **el resumen práctico de lo que se debe hacer**. Finalmente comentamos los resultados sobre el ejemplo planteado.

Un problema ejemplo

Se pretenden diferenciar las magnitudes de los armónicos (variable *Y*) correspondientes a determinadas frecuencias de los espectros de potencia de dos secuencias codificantes para proteínas. Las frecuencias fueron recodificadas mediante la MFT para p=17, w=2 de orden 8. Los datos siguieron una clasificación de doble entrada con: el factor "Secuencia", con dos niveles: Humano y Orangután, y el factor "Frecuencia" con tres niveles: 0.25, 0.5 y 0.75. Las repeticiones para cada combinación de niveles se hallan como se describe en la sección. Será importante en este análisis que el diseño sea equilibrado.

Como *Y* no tiene distribución normal, o al menos no hay forma de probarlo, se plantea un problema de análisis de varianza bifactorial no paramétrico a partir de los datos siguientes:

Secuencia (a =2 niveles)	Humano	Orangután
Frecuencia (b=3 niveles)	*Y*	*Y*
0.25	429.342	373.961
	494.502	298.975
	348.326	262.253
0.5	429.342	373.961
	494.502	298.975
	348.326	262.253
0.75	483.067	325.411
	409.614	217.381
	96.5849	110.226

Ideas del fundamento teórico y sus implicaciones prácticas

La idea esencial fundamentada en (R.R.Sokal and Rohlf, 1995) fue elaborar un Análisis de Varianza Bifactorial No Paramétrico, ranqueando la variable dependiente, como lo hace el test de Kruskal-Wallis. Se utilizan las sumas de cuadrados de la variable dependiente ranqueada y se recalculan grados de libertad de cada factor y su interacción para ofrecer finalmente una significación de cada efecto. Si algún factor tiene más de dos niveles, se pueden utilizar Test Post-Hoc clásicos que se basen fundamentalmente en rangos para obtener subconjuntos homogéneos, por ejemplo, el test de Dunnet C, válido incluso ante falta de homogeneidad de varianzas.

El ranqueo se realiza utilizando la siguiente función implementada en el paquete *Mathematica* 6.0:

```
RankValues[values_]:=
   Module[{s,m,r,a,mea11ns,ranks,rules},
```

```
    s=Split[Sort[values]];
    m=Map[Length,s];
    a=Accumulate[m];
    r=Range[1,Length[values]];
    means=Map[Mean,Drop[MapThread[Function[{i,k},
       Take[Drop[r,k],i]],{Append[m,0],Prepend[a,0]}],-1]];
    ranks=MapThread[Function[{i,j},Table[i,{j}]],{means,m}]//N;
    rules=MapThread[Function[{i,j},i[[1]]->j[[1]]],{s,ranks}];
    ReplaceAll[values,rules]
];
```

- En el caso que nos ocupa, el ranqueo conjunto de los datos conduce a los siguientes:

	Secuencia	Frecuencia	Y	Rango
1	Humano	0.25	429.342	14.5
2	Humano	0.25	494.502	17.5
3	Humano	0.25	348.326	9.5
4	Humano	0.5	429.342	14.5
5	Humano	0.5	494.502	17.5
6	Humano	0.5	348.326	9.5
7	Humano	0.75	483.067	16
8	Humano	0.75	409.614	13
9	Humano	0.75	96.5849	1
10	Orangután	0.25	373.961	11.5
11	Orangután	0.25	298.975	6.5
12	Orangután	0.25	262.253	4.5
13	Orangután	0.5	373.961	11.5
14	Orangután	0.5	298.975	6.5
15	Orangután	0.5	262.253	4.5
16	Orangután	0.75	325.411	8
17	Orangután	0.75	217.381	3
18	Orangután	0.75	110.226	2

Si identificamos a los niveles Humano y Orangután, del factor Secuencia, por 1 y 2 respectivamente y a los niveles 0.25, 0.5 y 0.75, del factor Frecuencia, por 1, 2 y 3, el conjunto de datos anteriores de puede expresar en términos de lista como:

datanew={{1,1,14.5},{1,1,17.5},{1,1,9.5},{1,2,14.5},{1,2,17.5},{1,2,9.5},{1,3,16},{1,3,13},{1,3,1},{2,1,11.5},{2,1,6.5},{2,1,4.5},{2,2,11.5},{2,2,6.5},{2,2,4.5},{2,3,8},{2,3,3},{2,3,2}}

- A continuación se hace un Análisis de Varianza Bifactorial (paramétrico) pero utilizando la variable Y ranqueada, no la variable original, mediante la función:

```
ANOVA[datanew,{namef1,namef2, All},{namef1,namef2}];
```

donde "namef1" es nombre dado al primer factor y "namef2" es el nombre dado al segundo factor.

Los resultados deben conducir esencialmente a lo siguiente:

Factor de variación	Grados de libertad (gl)	Suma de Cuadrados (SC)	Cuadrado Medio (CM)	F y Significación (no tener en cuenta)
Secuencia (A)	1	168.056	168.056	
Frecuencia (B)	2	49	24.5	
AxB	2	0.444444	0.222222	
Error	12	264	22	
Total	17	481.5		

En realidad, de la tabla anterior solo vamos a utilizar la suma de cuadrados por cada factor y su interacción así como sus grados de libertad. El Cuadrado Medio Total, y la estimación posterior de la verdadera significación de cada factor se realizarán como sigue:

- Calcule el CMT (Cuadrado Medio Total) por la fórmula

$$CMT = \frac{abn(abn+1)}{12}$$

a es el número de niveles del 1er factor (Secuencia, en este caso *a*=2)

b es el número de niveles del 2do factor (Frecuencia, en este caso *b*=3)

n es el número de réplicas de cada combinación (en este caso *n*=3)

Así, en el ejemplo:

CMT=18(19)/12=28.5

- Para cada factor y la interacción se calcula el estadígrafo *H* por la fórmula

$$H = \frac{SC(correspondiente)}{CMT}$$

En el ejemplo, tendremos,

- Para el Factor A: H = 168.056/28.5 = 5.8967
- Para el Factor B: H = 49/28.5 = 1.7193
- Para la interacción AxB H = 0.444444/28.5 = 0.0155945

- La significación para *H* se prueba como una variable con distribución Chi-cuadrado con los grados de libertad correspondiente a cada factor o la interacción.

En este caso, las significaciones respectivas resultan:

- Para el Factor A: $Sign\chi^2$ (5.8967) = 0.015 < 0.05
- Para el Factor B: $Sign\chi^2$ (1.7193) = 0.423 > 0.05
- Para la interacción AxB $Sign\chi^2$ (1.7193) = 0.992 > 0.05

Se concluye entonces que no hay influencia de interacción y tampoco de la Frecuencia. Lo determinante es la Secuencia. De los datos descriptivos de la variable original y de sus rangos, se identifica fácilmente que los resultados son más altos en el caso de la secuencia humana.

Dicha fundamentación teórica, implica desde el punto de vista práctico, que:

- Podemos utilizar el paquete *Mathematica* (podría ser incluso el Excel) para implementar como tal, el test bifactorial no paramétrico, y en particular:
  - Usar las **sumas de cuadrados de los rangos** y sus **grados de libertad** obtenidas a partir de la aplicación del test bifactorial paramétrico (función ANOVA, perteneciente al Kernel del *Mathematica* 6.0).
  - Recalcular, el valor de *CMT*, los valores de *H* y **las diferencias honestamente significativas,** desde el punto de vista no paramétrico, debidas a efectos principales y/o su interacción, pero acorde a la nueva teoría.

Resumen práctico de lo que hay que hacer

1. Prepare una lista de datos con los 2 factores y la variable dependiente. Tendrá siguiente la siguiente forma:

{{1,1, 429.342},{1,1, 494.502},{1, 348.326},{1,2, 429.342},{1,2, 494.502},{1,2, 348.326},{1,3, 483.067},{1,3, 409.614},{1,3, 96.5849},{2,1, 373.961},{2,1, 298.975},{2,1, 262.253},{2,2, 373.961},{2,2, 298.975},{2,2, 262.253},{2,3, 325.411},{2,3, 217.381},{2,3, 110.226}}

En una hoja del Exel o en el SPSS tendría el aspecto siguiente:

	Secuencia	Frecuencia	Y

1	1	1	429.342
2	1	1	494.502
3	1	1	348.326
4	1	2	429.342
5	1	2	494.502
6	1	2	348.326
7	1	3	483.067
8	1	3	409.614
9	1	3	96.5849
10	2	1	373.961
11	2	1	298.975
12	2	1	262.253
13	2	2	373.961
14	2	2	298.975
15	2	2	262.253
16	2	3	325.411
17	2	3	217.381
18	2	3	110.226

La variable que identifica los casos no es obligatoria. La variable Secuencia tiene las etiquetas de valores 1: Humano 2: Orangután y la variable Frecuencia tiene las etiquetas de valores 1: 0.25, 2: 0.5 y 3: 0.75.

2. Complete ahora los cálculos de *CMT*, los valores de *H* para cada factor y su significación a partir de la distribución Chi-cuadrado. Puede utilizarse para ello el *Excel* o el paquete *Mathematica*

Para facilitar esto se han preparado tres funciones simples en el paquete *Mathematica* 6.0 una de ellas utiliza el contexto de ANOVA dentro del *package* de Análisis de Varianza, para realizar el análisis paramétrico a la variable ranqueada.

Usted puede copiar el texto abajo a una Notebook del *Mathematica* 6.0. No olvide ejecutar todas las instrucciones (carga del paquete y de las funciones) al iniciar la sesión de trabajo. Una vez ejecutadas quedarán activas y las funciones podrán ser utilizadas en más de un análisis.

```
RankValues[values_]:=
   Module[{s,m,r,a,means,ranks,rules},
      s=Split[Sort[values]];
      m=Map[Length,s];
      a=Accumulate[m];
      r=Range[1,Length[values]];
```

```
      means=Map[Mean,Drop[MapThread[Function[{i,k},
         Take[Drop[r,k],i]],{Append[m,0],Prepend[a,0]}],-1]];
      ranks=MapThread[Function[{i,j},Table[i,{j}]],{means,m}]//N;
      rules=MapThread[Function[{i,j},i[[1]]->j[[1]]],{s,ranks}];
      ReplaceAll[values,rules]
   ];
test[nrep_,lf1_,lf2_,namef1_,namef2_,sqsumf1_,sqsumf2_,sqsumf1f2_]:=
   Module[{cmtot,grlf1,grlf2,Hf1,Hf2,Hf1f2,sigf1,sigf2,sigf1f2,finalt},
      cmtot=nrep*lf1*lf2*(nrep*lf1*lf2+1)/12;
      {Hf1,Hf2,Hf1f2}=N[{sqsumf1,sqsumf2,sqsumf1f2}/cmtot,4];
      {grlf1,grlf2}={lf1,lf2}-1;grlf1f2=grlf1*grlf2;
      sigf1=N[1-CDF[ChiSquareDistribution[grlf1],Hf1],3];
      sigf2=N[1-CDF[ChiSquareDistribution[grlf2],Hf2],3];
      sigf1f2=N[1-CDF[ChiSquareDistribution[grlf1f2],Hf1f2],3];
      finalt=PaddedForm[TableForm[
      Transpose[{{Hf1,Hf2,Hf1f2},{sigf1,sigf2,sigf1f2}}],
         TableHeadings->{{namef1,namef2,namef1<>"*"<>namef2},
         {"       H","   Sign"}}],{10,3}];
      Return[finalt]
   ];
Needs["ANOVA`"];
BifactorialNonParamANOVA[data_,nrep_,lf1_,lf2_,namef1_,namef2_]:=
   Module[{datanew,res},
      datanew=data;
      datanew=Transpose[datanew];
      datanew[[3]]=RankValues[datanew[[3]]];
      datanew=Transpose[datanew];
      res=ANOVA[datanew,{namef1,namef2,All},{namef1,namef2}];
      test[nrep,lf1,lf2,namef1,namef2,res[[1]][[2]][[1]][[1]][[2]],
         res[[1]][[2]][[1]][[2]][[2]],res[[1]][[2]][[1]][[3]][[2]]]
   ];
```

La función **`RankValues`** tiene el parámetro:

`values:` lista de valores de la variable dependiente que serán ranqueados.

La función **`test`** tiene los siguientes parámetros:

`nrep`: Representa el número de réplicas (constante en cada combinación de valores de los factores)

`lf1`: Niveles del factor 1

`lf2`: Niveles del factor 2

`namef1`: Nombre del factor 1

`namef2`: Nombre del factor 2

`sqsumf1`: Suma de cuadrados del factor 1

`sqsumf2`: Suma de cuadrados del factor 2

`sqsumf1f2`: Suma de cuadrados de la interacción

La función **`BifactorialNonParamANOVA`** tiene los siguientes parámetros:

`nrep,lf1,lf2,namef1,namef2`: Como en la función **`test`**

`data`: Conjunto de datos de la forma mostrada en 1 del Resumen Práctico

Una vez cargadas las funciones en general, será invocada la función **`BifactorialNonParamANOVA`** con los parámetros correspondientes a cada análisis. En nuestro ejemplo, sería así

```
BifactorialNonParamANOVA[{{1,1,429.342},{1,1,494.502},{1,348.326},{1,2,42
9.342},{1,2,494.502},{1,2,348.326},{1,3,483.067},{1,3,409.614},{1,3,96.5849
},{2,1,373.961},{2,1,298.975},{2,1,262.253},{2,2,373.961},{2,2,298.975},{2,
2,262.253},{2,3,325.411},{2,3,217.381},{2,3,110.226}},3,2,3,"Sequence","Fre
quency"]
```

La respuesta del *Mathematica* será una tabla como la siguiente

	H	Sign
Sequence	5.114	0.024
Frequency	3.841	0.279
Sequence*Frequency	0.364	0.948

Y se puede apreciar que los resultados coinciden con los esperados y son diferentes a las que arrojó originalmente el ANOVA paramétrico y que teníamos que despreciar. Si las variables son puramente ordinales, con un espectro de valores menor esta diferencia de resultados puede ser mucho más marcada y aparecer significaciones que por este proceso se obtienen pueden diferir mucho más de las significaciones obtenidas por el ANOVA paramétrico del cual solo tomamos las sumas de cuadrados

Alternativamente usted puede utilizar el paquete *Excel*. Los valores de *CMT* y de *H* se calculan de forma obvia. Para el cálculo de la significación de cada valor de *H* con sus grados de libertad se utiliza la fórmula CHIDIST(H,deg_freedom).

Anexo 2. Espectros de potencia de genomas mitocondriales

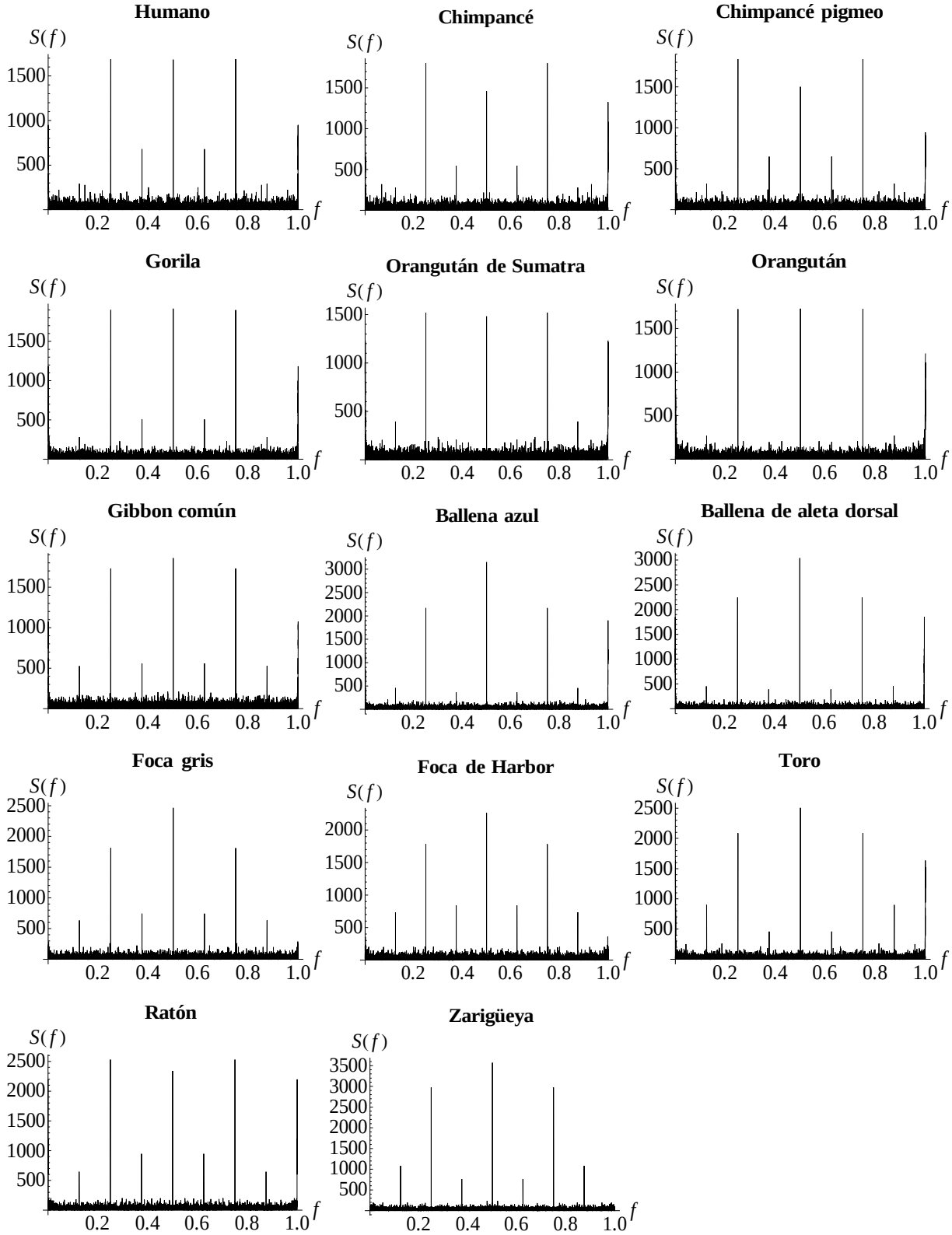

Figura A.2.1. Espectros de potencia de secuencias de genomas de los 14 mamíferos estudiados obtenidos para la recodificación *p*=17, orden 8 y *w*=2.

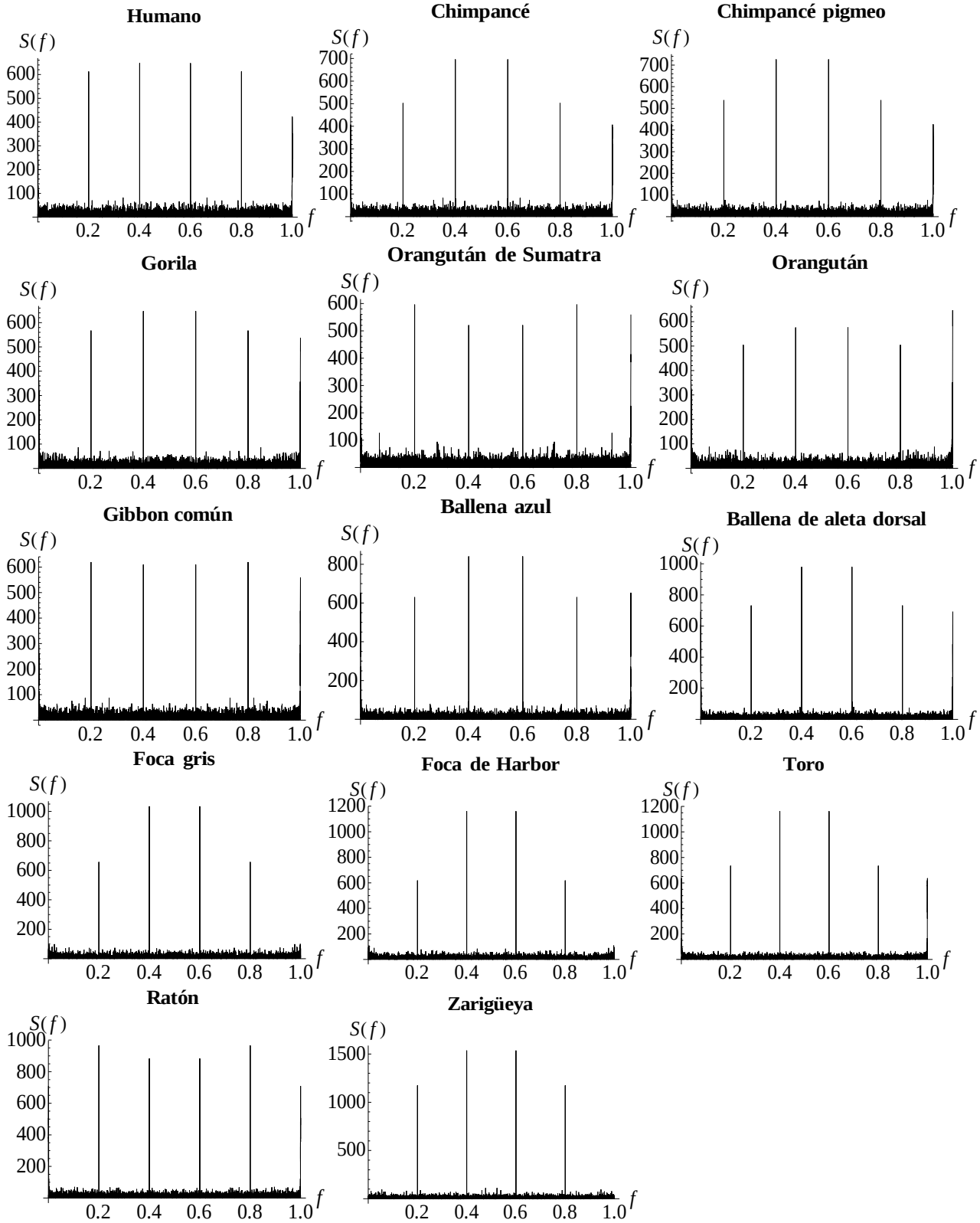

Figura A.2.1. Espectros de potencia de secuencias de genomas de los 14 mamíferos estudiados obtenidos para la recodificación p=11, orden 5 y w=3.