UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA,ALIMENTARIA Y DE BIOSISTEMAS

GRADO EN BIOTECNOLOGÍA

DEPARTAMENTO DE MATEMÁTICA APLICADA

CLASIFICACIÓN DE CADENAS DE ADN USANDO TÉCNICAS MULTIFRACTALES PARA EL ANÁLISIS DE LA FLUCTUACIÓN DE SERIES TEMPORALES

TRABAJO FIN DE GRADO

Autor/a: José Javier Tejeda Sánchez

Tutor/a: Carlos García-Gutiérrez Báez y Fernado San José Martínez

Julio de 2022

UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Técnica Superior De

Ingeniería Agronómica, Alimentaria y de Biosistemas

GRADO DE BIOTECNOLOGÍA

CLASIFICACIÓN DE CADENAS DE ADN USANDO TÉCNICAS MULTIFRACTALES PARA EL ANÁLISIS DE LA FLUCTUACIÓN DE SERIES TEMPORALES

TRABAJO FIN DE GRADO

José Javier Tejeda Sánchez

MADRID, 2022

Carlos García-Gutiérrez Báez y

Fernando San José Martínez

Dpto. de Matemática Aplicada, UPM

TITULO DEL TFG- CLASIFICACIÓN DE CADENAS DE ADN USANDO TÉCNICAS MULTIFRACTALES PARA EL ANÁLISIS DE LA FLUCTUACIÓN DE SERIES TEMPORALES

Memoria presentada por José Javier Tejeda Sánchez para la obtención del título de Graduado en Biotecnología por la Universidad Politécnica de Madrid

Fdo: José Javier Tejeda Sánchez

VºBº Tutor y Director del TFG

Carlos García-Gutiérrez Báez
Dpto. de Matemática Aplicada
ETSIAAB - Universidad
Politécnica de Madrid

VºBº Cotutor

Fernando San José Martínez
Dpto. de Matemática Aplicada
ETSIAAB - Universidad Politécnica de Madrid

Madrid, 8 de julio 2022

AGRADECIMIENTOS

Me gustaría comenzar agradeciendo a las personas que me han acompañado a lo largo de este trabajo y me han guiado para hacerlo lo mejor posible, corrigiendo hasta el último momento para que saliera a tiempo y bien, efectivamente, me refiero a mis tutores, Carlos y Fernando, gracias por haberme enseñado a la importancia que pueden llegar a tener las matemáticas en la biotecnología.

A la gente maravillosa que he conocido tanto en el Aquinas como en la universidad y con los que he pasado días maravillosos. Desayunar con Jorge en el único momento del día que hablaba menos y con el que había pasado la noche anterior junto a Nacho en teatro, a veces con buenos momentos y otras menos buenos, todo dependiendo del lugar. Tras el desayuno ir a clase con Julia la coordinadora, uno con más energía que otra, pero no diré quién es quién. Llegar a clase y encontrarme con Marcos, Ana, Julia, Inés y ya al rato a Lucía. Llegar a comer al Aquinas y encontrarme con Sergio que se iba volando a clase y ya dependiendo del día estar con unos u otros.

Gracias a toda mi familia, siempre preocupándose por mí y alegrándose cuando salía todo bien. Tengo que hacer mención honorífica a mis abuelas Fita y Chon, a mis tíos Rosi, Tomás y Mila y a Mario, que seguirá siendo el niño aunque tenga 50 años y nos doble en altura a todos.

A los padres de Sara, que me ayudaron para matricularme este último curso y que siempre que han estado aquí en Madrid me han tratado como a un hijo.

Por último me gustaría dar un especial agradecimiento a 4 personas más, las cuáles han sido un gran apoyo y la fuente de mis fuerzas esta última semana.Gracias a mis padres por haber confiado en mí y darme la oportunidad de venir aquí a Madrid para estudiar. En todo momento me habéis apoyado, a lo largo de la carrera y en todas las decisiones que he tomado en ella, siempre levantándome el ánimo en los peores momentos, dándome consejos cuando ha sido necesario y estando en los buenos momentos también.

A Sara, siempre ahí con tu tremendo amor y cariño que me llenaban de energías para seguir con todo adelante aunque la situación fuese terrible, gracias por todos tus detalles día a día y por los continuos ánimos.

Y por último, pero no menos importante, la persona que más ha tenido que aguantar mis tonterías y quejas de todo Madrid, gracias Nico. Estando para todo siempre con algo positivo que decir, ayudando a cualquiera en todo momento y habiendo pasado conmigo innumerables noches ya fuera de paseo, viendo YTPH u otros vídeos a la misma altura.

Me gustaría seguir agradeciéndoos, pero las normas son las normas y solo hay una hoja para agradecimientos, si no escribiría un libro entero.

Índice

Índice de figuras

Índice de tablas

LISTA DE ABREVIATURAS

1. **ADN:** Ácido desoxirribonucleico.
2. **ARNm:** Ácido Ribonucleico mensajero.
3. **DFA:** Detrended Fluctuation Analysis. Análisis de Fluctuación sin Tendencia.
4. **MF-DFA:** MultiFractal Detrended Fluctuation Analysis. Análisis Multifractal de Fluctuación sin Tendencia.
5. **ANOVA:** ANalysis Of VAriance. Análisis de la varianza.
6. **NCBI:** National Center for Biotechnolgy Information. Centro Nacional para Información Biotecnológica.
7. **EMBL:** European Molecular Biology Laboratory. Laboratorio Europeo de Biología Molecular.
8. **EBI:** European Bioinformatics Institut. Instituo Europeo de Bioinformática.
9. **NIH:** National Human genome research Institute. Instituto Nacional de investigación de genoma Humano.

ABSTRACT

The aim of this research is study the feasibility of classifying DNA sequences using parameters obtained using mathematical tools for sequence analysis.

For this purpose, a study has been carried out on 200 DNA sequences that have been collected from different databases, such as NCBI [7] or EMBL [1]. The first step was to convert the DNA sequences into time series using a method described by Peng *et al.* [9].

Once the time series were obtained, the methods described in Peng *et al.* [9] were used to make a fluctuation analysis that provides a parameter called α.

On the other hand, the time series were also used to perform a MF-DFA [6], with which $h(q)$ values were obtained for $q \in \{-10, -9, ..., 9, 10\} \cup \{\pm 0{,}2\}$ and interpolating polynomials of different degrees 1, 2 and 3.

After calculating the parameters α and $h(q)$, we used them to perform an hypothesis testing (T-Student, ANOVA, Tukey test), depending on the characteristics we wanted to clasify. Using the p-values obtained and the α and $h(q)$ means, we can see which values could serve as classifiers.

Finally, a classification has been carried out with two machine learning methods (k-means and neural networks). In both methods the study is done using only the $h(q)$ parameters, and other classification is done using the α and $h(q)$ parameters.

The results of this research sugest that we can't clasify the DNA sequences using neural networks because the error rates for all classifications are very high (the smallest is 0.18). This situation may be due to two possible reasons. The first is that the database is not large enough to train the classifier correctly. The second possible case is that there are not enough parameters for this task. However the hypothesis testing reveals significant differences between the parameters for the selected characteristics.

CAPÍTULO 1. INTRODUCCIÓN Y OBJETIVOS

En la ciencia, desde que en 1869 Freidrich Miescher descubrió el ADN y en 1944 Avery, McLeod y McCarty demuestran la presencia de la información genética en esta molécula, ha existido un interés en descifrar esta información. Gracias a los procesos de secuenciación creados en 1953 por Sanger y a los avances posteriores, se ha llegado a un punto donde a día de hoy estos métodos son sencillos y rápidos de utilizar. Esto provoca que en la actualidad se tenga acceso a más información que nunca en las bases de datos de ADN (como por ejemplo en el NCBI [7] o en EMBL [1]), algo que es bueno pero que puede llegar a generar un problema de exceso de información [11]. Para lidiar con este problema hay que realizar una criba de los datos con los que se quiere trabajar realmente. Es aquí donde entra en juego la bioinformática, la cuál puede ser una herramienta que facilita el procesamiento de sumas tan elevadas de datos.

Para llegar al punto de partida de este trabajo han pasado una serie de acontecimientos que habría que destacar.

En 1992 Peng *et al.* [9] desarrolla un método que permite convertir las secuencias de ADN en series temporales [9]. Otro procedimiento novedoso que introduce en este trabajo es un método de estudio de las series temporales, que más tarde se mejoró hasta convertirse en un *Detrended Fluctuation Analysis* (DFA). Gracias a este método descubre que las series temporales creadas a partir de secuencias que contienen intrones, presentan correlaciones de largo alcance, frente a las series que vienen de secuencias sin intrones que no presentan estas correlaciones. Además Peng *et al.* revelan a partir de este estudio que la escala cuantitativa de las correlaciones es similar a la de numerosos fenómenos que tienen un origen fractal, por los que deja en el aire que podrían existir propiedades fractales dentro de las series temporales.

En 2002 Kantelhardt *et al.* [6] emplea un nuevo algoritmo para caracterización multifractal de las series temporales no estacionarias, basándose en DFA. Este método es similar al análisis multifractal y permite identificar las fluctuaciones en las series temporales utilizando exponentes para describir todo el comportamiento de la fluctuación a lo largo de la serie. A este algoritmo lo llama *MultiFractal Detrendred Fluctuation Analysis* (MF-DFA) [6].

Con todos estos precedentes, el objetivo principal de este trabajo es el estudio de la viabilidad de clasificar secuencias de ADN cuyo origen es desconocido, usando parámetros obtenidos mediante herramientas matemáticas de análisis de secuencias. Para lograrlo hay que cumplir una serie de pasos previos. Lo primero es crear una base de datos con secuencias de ADN, que han sido buscadas en base a las características que se quieren clasificar. Estas características son la presencia o ausencia de intrones, el reino al que pertenece la secuencia, en el caso de las plantas concretar si son monocotiledóneas o dicotiledóneas, en el caso de los animales comprobar si son o no son humanos y en el caso de las bacterias

diferenciar entre bacilli y gammaproteobacterias. A continuación hay que convertir las cadenas de ADN en series temporales para poder utilizar los métodos matemáticos de Peng *et al.* [9] y de MF-DFA [6]. Estos métodos se programarán en Python para obtener los datos que se utilizarán en un estudio estadístico clásico (test de hipótesis con método T-Student, ANOVA y test de Tukey) que dará la información sobre la posibilidad o no de de realizar distinciones entre los grupos de estudio. Para concluir, se emplearon métodos de aprendizaje automático (k-means y redes neuronales) para ver si se pueden agrupar y clasificar las secuencias en base a los parámetros que obtenidos,

En el caso de k-means se pretende comprobar si los parámetros se pueden agrupar en base a las características que se están estudiando (presencia de intrones, reino, tipo de planta, tipo de bacteria o tipo de animal). En el caso de las redes neuronales se tratará de construir un clasificador usando los parámetros para los grupos que se han descrito. Además para comprobar la eficacia de estos métodos se usarán en el caso de k-means el índice de Rand y en el caso de las redes neuronales la tasa de error.

Pese a que los test de hipótesis muestran diferencias significativas entre los parámetros para las características seleccionadas, los clasificadores obtenidos tienen unas tasas de error que en principio pueden parecer elevadas (la más baja de todas es 0,18). Esto puede deberse a dos posibles motivos. El primero es que la base de datos no es lo suficientemente grande para entrenar correctamente al clasificador. El segundo caso posible es que no haya suficientes parámetros para esta tarea.

CAPÍTULO 2. MATERIALES Y MÉTODOS

2.1. Secuencias obtenidas del NCBI

Para este estudio se utilizaron 200 secuencias de ADN, las cuales se obtuvieron de las siguientes bases de datos: Gene del National Center for Biotechnology Information (NCBI) [7]; y Ensembl [1], EnsemblBacteria [2], EnsemblPlants [4] y EnsemblFungi [3] de EMBL's European Bioinformatics Institute (EMBL-EBI). Las secuencias recopiladas pertenecen a distintos organismos y están seleccionadas en base a una serie de características.

- **Reino.**

  El reino es la segunda categoría taxonómica, solo por debajo del dominio. Clasifica a los seres vivos en 5 subdivisiones taxonómicas según su parentesco evolutivo. Estas subdivisiones son Protista, Animales, Plantas, Fungi y Mónera. Para este trabajo, como indica la tabla 1, se han recopilado 86 secuencias de animales, 17 de fungi, 46 de móneras, 43 de planta y 8 de protistas. Hay que remarcar que en este trabajo, todas las secuencias que se han utilizado del reino mónera son de bacterias, por lo tanto a partir de ahora se hablará de bacterias en lugar de móneras.

- **Clase.**

  Es la cuarta categoría taxonómica. En este trabajo se han obtenido secuencias de distintas clases (en total hay recogidas secuencias de 38 clases), aunque las que son utilizadas para el trabajo serán monocotiledóneas ($N = 13$) y dicotiledóneas ($N = 21$) dentro de las plantas y bacilli ($N = 10$) y gammaproteobacterias ($N = 21$) de las bacterias. Estas clases usadas son las que aparecen en la tabla 1

- **Presencia de intrones.**

  Según el National Human Genome Research Institute (NIH) "Un intrón es una región que reside en el interior de un gen, pero no permanece en la molécula madura final de ARNm después de la transcripción de ese gen y no codifica para los aminoácidos que conforman la proteína codificada por ese gen" [8]. En la base de datos utilizada para este trabajo se han recogido 78 secuencias sin intrones y 122 con intrones.

- **Función de la proteína.**

  Para este trabajo se han buscado secuencias de ADN en base a la función de la proteína que codifica. Estas funciones son: defensiva ($N = 27$), enzimática ($N = 43$), estructural ($N = 23$), hormonal ($N = 35$), reguladora ($N = 11$), de reserva ($N = 20$), de transporte ($N = 19$) y de movilidad ($N = 22$).

Como muestra la tabla 1, se realizarán estudios de diferenciación con distinto número de secuencias. Por lo que habrá estudios que incluyen las 200 secuencias como es el caso de los estudios de presencia de intrones, del reino y de la función. Por otro lado, se realizan estudios sobre un reino o una clase en concreto del conjunto de secuencias; este es el caso de diferenciar entre plantas monocotiledóneas y dicotiledóneas, distinguir entre humanos y el resto de animales o separar entre gammaproteobacterias y bacilli. En los casos de plantas y bacterias no se ha hecho el estudio de todas sus clases y se han elegido solo las clases que más número de secuencias tienen para realizar un test estadístico fiable. En el caso de secuencias para distinguir a humanos del resto de animales se hizo porque se quería ver hasta que punto existen diferencias de ADN entre los seres humanos y el resto de especies animales.

Estudio	Grupo	N
Presencia de intrones	No	78
	Sí	122
	N TOTAL	200
Reino	Animal	86
	Fungi	17
	Bacteria	46
	Planta	43
	Protista	8
	N TOTAL	200
Función	Defensiva	27
	Enzimática	43
	Estructural	23
	Hormonal	35
	Movilidad	22
	Reguladora	11
	Reserva	20
	Transporte	19
	N TOTAL	200
Número de cotiledones	Monodicotiledónea	13
	Dicotiledónea	21
	N TOTAL	34
Humano	Sí	23
	No	63
	N Total	86
Clase de bacteria	Bacilli	10
	Gammaproteobacteria	21
	N TOTAL	31

Tabla 1: Tabla que resume las poblaciones que se han elegido para los estudios.

Hay que aclarar que todas las secuencias que se han seleccionado pertenecen a más de un grupo. Una misma secuencia puede ser, por ejemplo, del reino vegetal, la clase monocotiledónea, tener intrón

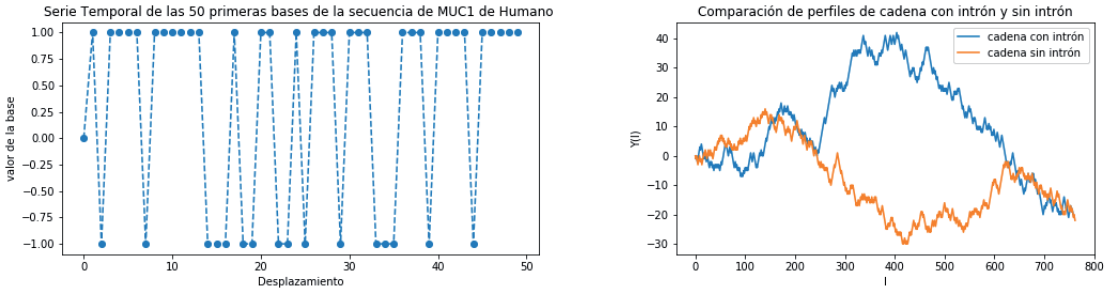
y generar una proteína reguladora.

2.2. Series temporales

Una serie temporal es una secuencia de variables aleatorias que se obtienen a lo largo de un periodo de tiempo de forma regular o irregular [10], por ejemplo, la medición de la temperatura a lo largo de un año. En este ejemplo nuestra variable aleatoria es la temperatura que se ha ido obteniendo cada día, de forma regular, durante un año. Esto nos permite, entre otras cosas, ver qué factores influyen en la serie y en base a estas observaciones se pueden elaborar métodos de previsión, como es el caso de predecir la temperatura que hará un día futuro.

Para el caso concreto que se está estudiando en este trabajo, lo primero que hay que hacer es convertir el código de la secuencia de ADN, compuesto por letras, en números. Para ello se utilizó un algoritmo desarrollado por Peng *et al.* [9], el cual, utiliza un método similar al de paseo aleatorio o *random walk* que permite obtener una serie temporal a partir de una secuencia de ADN.

Este algoritmo que presenta Peng consiste en moverse por la cadena de ADN dando "pasos" de nucleótido en nucleótido donde i será el nucleótido y $u(i)$ será el valor que se le asignará al nucleótido de la posición i : $u(i) = +1$ si la base es púrica (adenina A y guanina G) y $u(i) = -1$ si la base es pirimídica (timina T y citosina C). De esta forma se ha crado la serie temporal $u(i)$. A continuación, en la imagen 1a, se muestra una gráfica de ejemplo de la serie temporal creada a partir de los 50 primeros nucleótidos de la secuencia de MUC1 de humano. Por otra parte la figura 1b muestra una gráfica del perfil de la serie temporal Y(l), para las secuencias de PomA de Vibrio alginolyticus (secuencia sin intrón) y la de la paratohormona de Chelonoidis abingdonii (secuencia con intrón), este perfil se explicará en el siguiente.



(a) Serie temporal de MUC1

(b) Perfil de secuencias con y sin intrones

Figura 1: Gráficas de serie temporal elaborada por el método de Peng [9] (a) y de perfiles de secuencias con y sin intron (b) donde la línea azul es la secuencia con intrón y la naranja sin intron.

2.3. Análisis de la fluctuación de la serie temporal

Se comienza calculando el desplazamiento o perfil (Y), que será lo que se utilice para hacer el análisis.

$$Y(l) = \sum_{i=1}^{l} \left[u(i) - \overline{u}\right], \quad l = 1, ..., N. \tag{1}$$

donde $u(i)$ es una serie de tamaño N $(i = 1, ..., N)$ y $\overline{u}$ es la media de las serie $u(i)$. Esta ecuación también se puede realizar sin la media, como hace Peng *et al.* para el DFA [9]

2.3.1. Análisis de fluctuación de Peng (DFA)

Una vez calculado el perfil, Peng caracteriza la serie temporal calculando el valor cuadrático medio de la fluctuación.

$$F^2(l) = \overline{\left[\Delta Y(l) - \overline{\Delta Y(l)}^2\right]} = \overline{\left[\Delta Y(l)\right]^2} - \overline{\Delta Y(l)}^2 \tag{2}$$

donde $\Delta Y(l)$ viene determinado por la fórmula 3 y las líneas encima de los elementos de la ecuación son la media :

$$\Delta Y(l) = Y(l_0 + l) - Y(l_0) \tag{3}$$

donde l es el tamaño de la ventana con inicio en la posición l_0 y final en la posición $l_0 + l$. Según Peng *et al.*, los valores de l van desde 2 hasta una décima parte de la longitud de la cadena, ya que para valores mayores a esta décima parte el error estadístico aumenta en análisis de este tipo. Debido a esto l_0 va desde 1 hasta $N - (N/10)$.

Este cálculo de $F(l)$ nos puede llevar a tres casos distintos [9]. El primero que el comportamiento de la serie fuese totalmente aleatorio y $C(l)$ sería 0 de media a excepción del caso $C(0) = 1$, por lo que $F(l) \sim l^{1/2}$. El segundo caso es que exista una correlación local en un rango de la serie R, en ese caso $C(l) \sim exp(-1/R)$, sin embargo seguirá ocurriendo que $F(l) \sim l^{1/2}$. La tercera opción es que el rango R sea muy extenso y no pueda caracterizase, en ese caso habrá una correlación de larga distancia, $C(l)$ no será exponencial y $F(l) \sim l^{\alpha}$, donde $\alpha \neq \frac{1}{2}$. En todos estos casos el parámetro α, se obtiene al realizar una regresión lineal de la gráfica log-log de $F(l)$.

Por lo tanto, si $\alpha \approx 0,5$, la serie tendrá un comportamiento aleatorio o correlaciones locales (de corto alcance), mientras que si $\alpha > 0,5$ existirán correlaciones de largo alcance.

En base a este parámetro α, en Peng *et al.* [9], se llega a la conclusión de que en los casos donde existe correlación de largo alcance ($\alpha > 0,5$), se corresponden con las secuencias con intrones, mientras que en los casos donde hay correlaciones locales ($\alpha \approx 0,5$) coinciden con las secuencias sin intrones.

2.3.2. Análisis Multifractal de Fluctuación sin Tendencia (MF-DFA)

A partir de las series temporales se puede realizar otro estudio más general de la fluctuación, el cuál utiliza una serie de exponentes para poder estudiar los distintos valores de la serie temporal centrándonos en los valores más bajos o más alto en función del exponente. El MF-DFA permite el estudio de la variación de las fluctuaciones de la serie temporal teniendo en cuenta los exponentes utilizados para el estudio de valores de la serie. El método de MF-DFA según Kantelhardt *et al.* [6] consta de 5 pasos,

- **Paso 1.** Determinar el perfil como indica la ecuación 1.
- **Paso 2.** Dividir el perfil $Y(i)$ en segmentos no solapantes de la misma longitud s ($N_s \equiv int(N/s)$). Habrá casos donde N no sea divisible por s, y sobra una parte del perfil al final, para evitar el problema se realiza la misma división empezando por el lado contrario al anterior y se tendrá $2N_s$ segmentos. El valor de s irá desde 10 hasta $N/4$, porque tal y como menciona Kantelhardt *et al.* [6] para $s > N/4$ el número de segmentos N_s que se usarán en el paso 4 se vuelve muy pequeño y por lo tanto habrá pocos datos incorrelados.
- **Paso 3.** Calcular la tendencia local para cada uno de los $2N_s$ segmentos y posteriormente calcular la varianza.

$$F^2(\nu, s) = \frac{1}{s}\sum_{i=1}^{s}\left\{Y\left[(\nu-1)s+i\right] - y_\nu(i)\right\}^2, \quad \nu = 1, ..., N_s \tag{4}$$

$$F^2(\nu, s) = \frac{1}{s}\sum_{i=1}^{s}\{Y[N-(\nu-N_s)s+i] - y_\nu(i)\}^2, \quad \nu = N_s, ..., 2N_s \tag{5}$$

  donde ν es el índice del segmento. Tanto en la ecuación 4 como en la 5 $y_\nu(i)$ es un polinomio de ajuste de los valores de la serie en el segmento ν. El polinomio de ajuste es el polinomio que pasa más cerca del conjunto de datos que hay, en este caso los datos son los del segmento. El polinomio se utiliza para eliminar tendencias del perfil del mismo orden que el polinomio. Cada vez que se realice el MF-DFA hay que elegir el grado de este polinomio, el cuál no varía a lo largo del análisis. En este trabajo se han han elegido polinomios de grado 1, 2 y 3..

  La ecuación 5 solo se utiliza en el caso de que en el paso 2, la longitud N no sea divisible por la longitud s de los segmentos.
- **Paso 4.** Promedio de todos los segmentos para obtener la función de fluctuación de orden q.

$$F_q(s) = \left\{ \frac{1}{2N_s} \sum_{\nu=1}^{2N_s} \left[F^2(\nu, s)\right]^{q/2} \right\}^{1/q} \tag{6}$$

donde q es un valor variable que nos permite el estudio de unos valores y otros de la serie, como se mencionaba al inicio de este punto. Hay que repetir los pasos del 2 al 4 para distintos s y q; en el trabajo se realiza el MF-DFA para 23 valores de q ($q \in \{-10, -9, ..., 9, 10\} \cup \{\pm 0{,}2\}$).

En el caso de que en el paso 2, la cadena de longitud N fuese divisible por los segmentos de longitud s, en la ecuación 6, dentro del sumatorio, se sustituye $2N_s$ por N_s.

- **Paso 5.** Determinar el comportamiento asintótico de F respecto a s para cada valor q.

$$F_q(s) \sim s^{h(q)}. \tag{7}$$

La ecuación 6 no puede utilizarse en el caso de $q = 0$, debido ha que habría un $\frac{1}{0}$. Para este caso la ecuación que se utiliza es la siguiente.

$$F_0(s) \equiv exp\left\{ \frac{1}{4N_s} \sum_{\nu=1}^{2N_s} ln\left[F^2(\nu, s)\right] \right\} \sim s^{h(0)}. \tag{8}$$

Según Kantelhardt *et al.* [6], si la serie temporal es monofractal $h(q)$ no depende de q, y el comportamiento será el mismo en toda la serie temporal.

La q permite filtrar los valores de fluctuaciones, cuánto más alta es, más diferencia hay entre los resultados de los valores altos y bajo. Por ejemplo si se tiene una serie con los valores 2 y 10, para $q = 10$, 2^{10} y 10^{10} =10.000.000.000, como se puede ver en este caso, donde antes estaban los valores 2 y 10, ahora se tiene 1024 y 10.000.000.000 y el 1024 es despreciable. Lo contrario ocurre en el caso de las q negativas, ya que elevar por un número negativo es lo mismo que dividir 1 entre el número elevado a q con signo positivo, por lo que en este caso, serán los valores que en principio eran altos los que se convierten en despreciables. Por ejemplo de nuevo con 2 y 10 y $q = -10$, $\frac{1}{1024}$ es mucho mayor que $\frac{1}{10000000000}$ y en este caso el segundo valor es muchísimo más bajo y por tanto despreciables.

Sin embargo si $h(q)$ es dependiente de q, significa que existe un comportamiento asintótico distinto de las fluctuaciones(F_q) pequeñas frente a las grandes.

Por último Kantelhardt [6] dice que el parámetro $h(2)$ es similar a α ya que como se puede observar en la ecuación 6 si q vale 2, la ecuación se corresponde a la de un análisis DFA y por lo tanto para este caso concreto no se tiene en cuenta la multifractalidad.

2.4. Análisis estadístico.(T-Student, ANOVA y test de Tukey)

Una vez obtenidos los valores α del método de Peng *et al.* [9] y los $h(q)$ del MF-DFA [6], se pasa a realizar un análisis estadístico para poder ver si existen diferencias entre los grupos de estudio (tabla 1).

Para ello se realizará un test de hipótesis en el que la hipótesis nula H_0 será que las medias de los grupos son iguales, mientras que la hipótesis alternativa, H_1 es que las medias son distintas. Para rechazar o no rechazar H_0 se han utilizado distintos métodos estadísticos clásicos. En el caso de que existan solo 2 grupos se realiza un test T-Student, mientras si existen más de 2 grupos se hace un test ANOVA. Tras realizar estos test obtenemos el p-valor, el cuál es "el nivel de significación mínimo no arbitrario con el que podemos rechazar la hipótesis nula (H_0) dada una función de distribución y un estadístico de contraste"[5]. Cuanto más bajo sea este valor más fiable es el rechazo de la hipótesis nula. Lo más común es poner como límite de 0,05, y cuando son menores a este número se rechaza H_0. En los casos de que el p-valor del ANOVA sea menor que 0,05, se lleva a cabo un test de Tukey para estudiar los p-valores 2 a 2 entre los distintos grupos. Con este p-valor se va a determinar si existen diferencias notables entre los grupos que se han propuesto al inicio del trabajo.

2.5. Métodos de aprendizaje automático

Una vez obtenidos todos los datos tras realizar el MF-DFA [6] y el estudio que realiza Peng [9], se realizará un estudio con aprendizaje automático para ver si los grupos que hemos visto que tenían diferencias entre sí se se pueden clasificar. Para ello se empleará un método de clustering, el k-means y un metodo de redes neuronales.

2.5.1. Clustering con el método de k-means

El método de K-means, agrupar datos en base a sus patrones. Para realizar este algoritmo solo son necesarios los datos $h(q)$ y α y el número de grupos que se quieran establecer (K). Por lo tanto en el trabajo se va a utilizar este método para agrupar las secuencias en base a los valores α y $h(q)$ y se estudiará si logran formar los grupos que estamos estudiando.

Para comprobar la calidad los agrupamientos se utilizará el índice de Rand, el cuál se usará para comparar los resultados obtenidos con k-means con los datos de la base de datos. Este índice se calcula como:

$$R = \frac{a+b}{a+b+c+d} \tag{9}$$

Donde a es el número de verdaderos positivos, b número de verdaderos negativos, c el numero de falsos positivos y d el de falsos negativos.

Este valor va de 0 a 1, si es cercano a 1 significa que el método de k-means sirve para realizar esta clasificación si es cercano a 0 significa que no.

2.5.2. Redes neuronales.

Para el trabajo se utilizarán redes neuronales para ver si se pueden clasificar las secuencias en base a α y a los $h(q)$. El funcionamiento de estas redes consiste en que los datos entran a la red, son procesados en las distintas capas y al final de la red se da un resultado.

El objetivo de utilizar las redes neuronales es el de conseguir clasificar las secuencias en base a los parámetros α y $h(q)$ que se han obtenido. Para ello, lo primero que hay que hacer es entrenar a la red con una parte de los datos que obtenidos y a continuación con el resto de datos se realiza un test para probar la efectividad de la clasificación. Con los datos del test de efectividad se calcula la tasa de error, que es

$$T.E. = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{10}$$

Donde $\hat{Y}_i$ es el vector de valores predichos en el test, Y_i el vector de valores reales y n el número total de secuencias.

Para el trabajo se han utilizado redes neuronales de 2 capas con 10 y 4 neuronas para cada capa.

2.6. Herramientas Informáticas

Para realizar este trabajo se utilizaron las páginas webs del NCBI [NCBI] y EMBL [Ensembl] para la obtención de las secuencias de ADN. Por otra parte se utilizaron los siguientes paquetes de Python, para programar los métodos utilizados (DFA [9] y MF-DFA [6]): pandas, numpy, matplotlib, seaborn, scipy, pingouin, Bio, klearn, math y os. Para la parte de aprendizaje automático se utilizaron las siguientes librerias de R: tidyverse, car, rstatix, cluster, 'fpc', neuralnet, nnet y NeuralNetTools.

Por último, para escribir este trabajo se ha utilizado LaTeX.

CAPÍTULO 3. RESULTADOS Y DISCUSIÓN

Una vez obtenidas las series temporales a partir de las secuencias de la base de datos creada, se pasa a aplicar los métodos de Peng [9] y MF-DFA [6] sobre ellas. Tras obtener los parámetros α y $h(q)$ se realizará el estudio estadístico clásico de los mismos y para finalizar aplicarán herramientas de aprendizaje automático.

3.1. Estudio estadístico del parámetro α

Obtenidos los valores α [9], se pasa a comprobar si las conclusiones que sacó Peng [9] se repiten para la base de datos utilizada en el trabajo, para ello se hizo un test T de Student como se muestra en las filas correspondientes a intrones de la tabla 2.

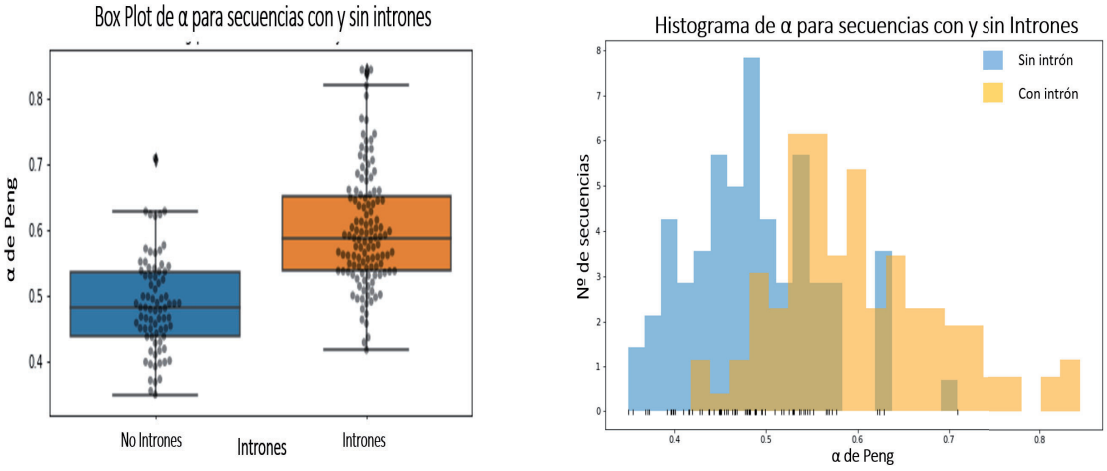
Posteriormente se hace un estudio del resto de los grupos que se están analizando para ver si existen diferencias entre ellos. Para ellos se realizará un estudio estadístico mediante un análisis de T de Student o ANOVA cuyos resultados se muestran en la tabla 2.

	Grupo	N	$\overline{\alpha}$	σ^2	p-valor
Intrones	Sin Intrones	78	0,486	0,072	$1,11 \cdot 10^{-17}$
	Intrones	122	0,6	0,09	
Reino	Animal	86	0,597	0,093	$7,55 \cdot 10^{-13}$
	Fungi	17	0,548	0,052	
	Bacteria	46	0,467	0,067	
	Planta	43	0,579	0,096	
	Protista	8	0,512	0,101	
Función de la proteína	Defensiva	27	0,557	0,109	0,0048
	Enzimática	43	0,514	0,095	
	Estructural	23	0,612	0,123	
	Hormonal	35	0,576	0,082	
	Movilidad	22	0,554	0,093	
	Reguladora	11	0,510	0,095	
	Reserva	20	0,539	0,078	
	Transporte	19	0,587	0,088	
Número de cotiledones	Monodicotiledónea	13	0,548	0,067	0,26
	Dicotiledónea	21	0,576	0,071	
Humano	Sí	23	0,605	0,084	0,59
	No	63	0,593	0,096	
Bacterias	Bacilli	10	0,482	0,043	0,48
	Gammaproteobacteria	21	0,465	0,067	

Tabla 2: Análisis estadístico para α . Tabla de los distintos grupos que se van a estudiar. Contiene la media del parámetro α ($\overline{\alpha}$) y su varianza σ^2 para cada grupo. Por último está representado el p-valor para cada estudio (si hay 2 grupos será un T-Student, si hay más de 2 será un ANOVA).

3.1.1. Estudio de los α para la de presencia de intrones realizado por Peng

Lo primero que se calcula es la media y varianza de los α de secuencias en función de la presencia de intrones en ellas, que se encuentran en la tabla 2. Por otra parte se ha realizado un box plot (figura 2a) para ver cómo se distribuyen estos valores α en función de la presencia o ausencia de intrones. Además se ha realizado un histograma de los valores de α para las secuencias con intrones y sin intrones (figura 2b).



(a) Box Plot para la presencia de intrones

(b) Histograma para la presencia de intrones

Figura 2: Representación de los valores de α mediante un box plot(a) donde la caja azul representa al grupo sin intrones y la narana al con intrones y un histograma(b) donde en azul aparece representado el grupo sin intrones y en amarillo el grupo con intrones.

Como se puede observar en la figura 2a parece que existe una diferencia entre el valor de α sin intrón y con intrón, esta misma idea se puede intuir de nuevo gracias a la gráfica 2b. Se realiza un Test de T-Student, en el cuál se considera como hipótesis inicial que la media de los α sin intrón es igual a la media de los α con intrón. Tras realizar esta prueba, se obtiene un p-valor de $1,11 \cdot 10^{-17}$ por lo tanto se puede rechazar la hipótesis inicial de que la media de los α de secuencias con intrones son distintos a la de los α sin intrones de la muestra. El valor de la media de los α para las secuencias sin intrones es 0,486 $\pm 0,005$ y el de las secuencias con intrones 0,6 $\pm 0,008$. Por lo tanto, tal y como decía Peng la media está en torno al 0,5 para los α de las secuencias sin intrones, indicando que no existe correlaciones de larga distancia en las series temporales correspondientes. Para los α de secuencias con intrones es mayor a 0,5, lo que muestra que sí existen correlaciones de larga distancia a lo largo de las series temporales.

3.1.2. Estudio de los α para el resto de grupos.

Una vez comprobada la efectividad del método de Peng [9] para separar secuencias que tienen intrones de las que no los tienen, se pasó a comprobar si existen diferencias del parámetro α para las otras características de estudio.

Lo primero que se comprobó fue si se pueden encontrar diferencias significativas en el valor de α para los distintos reinos, para ello se tomó una muestra de animales, hongos, bacterias, plantas y protistas, tal y como indica la tabla 2. Se realizó el box plot que aparece en la figura 3. En esta figura se puede ver como la media de α de las bacterias es bastante inferior al resto de reinos, probablemente debido a que solo contiene secuencias que no tiene intrones. A continuación se realizó un test ANOVA y como el p-valor era $7,55 \cdot 10^{-13}$, se puede ver como existen diferencias entre los reinos. Para ver entre qué reinos en concreto había diferencias notables de α se realizó un test de Tukey, el cuál determinó que solo existían diferencias entre animal con bacteria (p-valor=0,001), fungi con bacteria (p-valor=0,009) y planta con bacteria (p-valor=0,001). Sin embargo no hay diferencias entre los α de bacterias y protistas (p-valor=0,64).

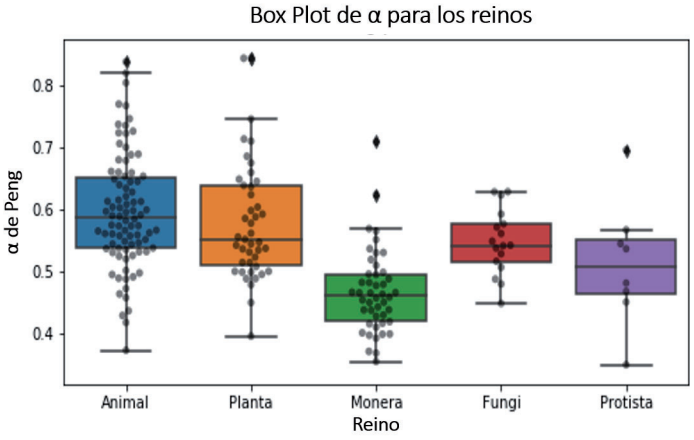


Figura 3: Box plot donde la caja azul representa al grupo de animales, la caja naranja al de plantas, la verde al de bacterias, la roja al fungi y la morada al protista.

Este resultado podría deberse a que todas las secuencias bacteria que se han utilizado no tienen intrones, mientras que las secuencias de animales y plantas la mayoría sí contienen intrones (aunque existe algunos sin intrón). Para el reino fungi, sí que hay más variedad de intrón/no intrón. Por otra parte, similitud entre los alfa de protista y bacteria podría deberse a que salvo 1 secuencia de protista, el resto son secuencias sin intrón como las de bacteria.

Además de con los reinos, también se realizo un estudio del parámetro α para la función de la proteína que codifica la secuencia de ADN. La muestra de datos se encuentra en la tabla 2. Después

de hacer un test ANOVA, sale que el p-valor es de 0,0048, por lo que de nuevo se puede rechazar que los valores de α para las distintas funciones sean los mismos. Tras realizar el test de Tukey se llega a la conclusión de que hay diferencias significativas entre los α de funciones enzimáticas y estructurales debido a que se obtiene un p-valor=0,005.

Debido a estos resultados, se puede concluir que utilizando el α se puede diferenciar entre las secuencias de ADN que codifican proteínas enzimáticas de las que producen para proteínas estructurales. La media de los α de las enzimáticas es 0,5 y por tanto no hay correlaciones de larga distancia y la de las estructurales es superior indicando que sí las hay. En este caso esta diferenciación es más fiable que en el de los reinos, ya que en ambas poblaciones existen secuencias con y sin intrones. Además se puede ver algo curioso en el box plot de la figura 4, y es que aunque parece que los α de la población función enzimática están a la misma altura o muy similar que los α de las funciones de movilidad, reserva y reguladora (cuya media de valores α es incluso menor), cuando se realiza el test de Tukey, el único p-valor que da un valor inferior a 0,05 que nos indica que existen diferencias de valores, es el de enzimática con estructural; esto puede deberse a que como se puede ver en la figura 4 y en la tabla 2, hay un mayor número de datos para las secuencias que generan proteínas enzimáticas, generando un resultado más preciso.

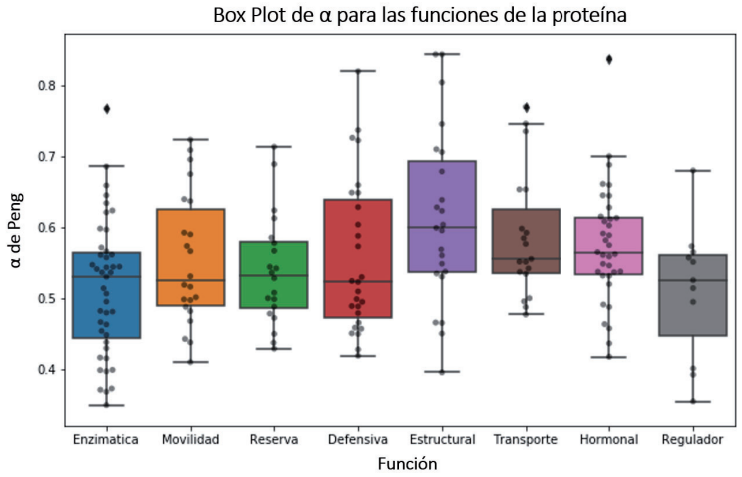


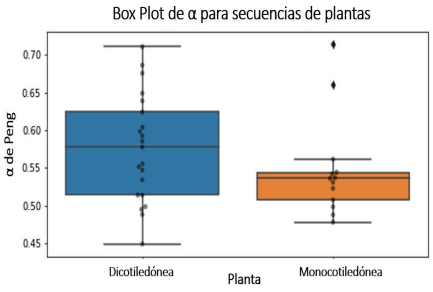
Figura 4: La figura muestra un box plot donde la caja azul representa la distribución de todos los α de las secuencias con función enzimática, la caja naranja movilidad, la verde reserva, la roja defensiva, la morada estructural, la marrón transporte , la rosa hormonal y la gris reguladora.

Por último se realizaron 3 test de T-Student entre dos de todas las clases de plantas y bacterias, por un lado y dentro del reino animal por otro, para ver si existen diferencias dentro los α para las secuencias de un mismo grupo. Para plantas se trabajó con las clases monocotiledóneas y dicotiledóneas, ya que eran las plantas de las que se disponía de una mayor cantidad de datos. Para el resto de plantas no

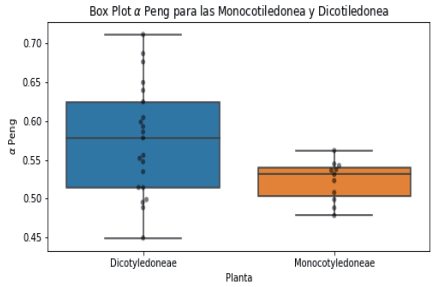
hay una cantidad suficiente de secuencias para realizar un buen análisis estadístico. Por otra parte se estudiaron las diferencias en el reino animal, más concretamente si existían diferencias significativas en los parámetros hallados para los grupos de humanos frente a no humanos. El tercer test que se hizo fue sobre las bacterias. En este caso al igual que con plantas y humanos solo se comparan dos clases (bacilli y gammaproteobacterias), ya que del resto de bacterias no hay suficientes datos para el análisis. Se podrían haber hecho más análisis dentro de los reinos, pero para ello es necesaria una base de datos de secuencias con más elementos.

En las tres situaciones descritas el p-valor es superior a 0,05: no existen diferencias significativas entre los α de los grupos. En la figura 5a se puede observar como no hay mucha diferencia entre las medias.

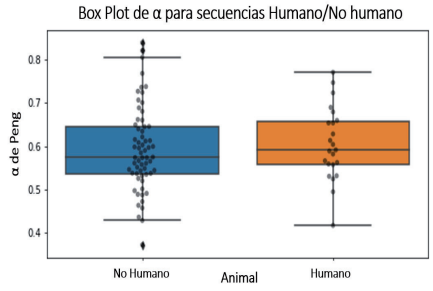
Cabe destacar que en la figura 5a hay un par de valores de α de monocotiledóneas que son mucho mayores que el resto de valores del grupo. Para comprobar si el test está siendo afectado por estas secuencias, que son posibles *outliers*, se procedió a eliminar esos dos valores y se realizó de nuevo un box-plot (imagen 5b) y un T-Student, para este caso el p-valor es 0,02, por lo que en este caso sí que existen diferencias entre los valores α de monocotiledóneas y dicotiledóneas.



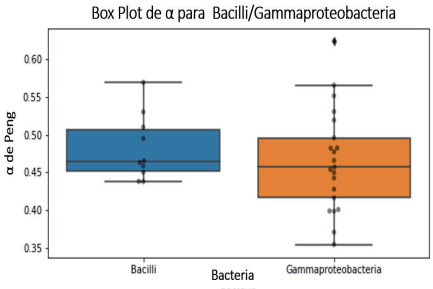
(a) Box Plot para plantas



(b) Box Plot para plantas sin los datos extremos



(c) Box Plot para animales



(d) Box Plot para bacterias

Figura 5: Representación mediante box plot de la población de de α de plantas (a), plantas sin los datos extremos de monocotiledóneas (b), animales (c) y bacterias (d) mediante un box plot donde la caja azul representa a dicotiledóneas(a y b), no humanos (c) y bacilli (d) y la naranja a monocotiledóneas (a y b), humanos (c) y gammaproteobacterias(d).

3.2. MF-DFA

Para el MF-DFA de nuevo se utilizaron las mismas secuencias y los mismos grupos que están en la tabla 2, con la diferencia de que esta vez, en lugar de calcular un parámetro α, se calcularán los parámetros $h(q)$, donde $q \in \{-10, -9, ..., 9, 10\} \cup \{\pm 0{,}2\}$. Además el MF-DFA se ha realizado en tres ocasiones cambiando el grado del polinomio, siendo estos grados 1, 2 y 3.

Para comenzar, se ha realizado un estudio de los $h(q)$ según la presencia de intrones utilizando los polinomios de los tres grados y así poder ver si se pueden utilizar para distinguir entre los $h(q)$ de cada grupo. Para ello, se ha ido grado a grado y q por q realizando test de T-student entre los grupos con intrón y sin intrón. La tabla 3 y la imagen 6 muestran los resultados de estos estudios.

	Grado1			Grado2			Grado3		
Q	p-valor	media sin intrones	media con intrones	p-valor	Media sin intrones	Media con intrones	p-valor	media sin intrones	media con intrones
10	$5{,}54 \cdot 10^{-10}$	0,471 ± 0,091	0,611 ± 0,174	$1{,}38 \cdot 10^{-10}$	0,499 ± 0,077	0,586 ± 0,0966	$1{,}16 \cdot 10^{-08}$	0,515 ± 0,067	0,587± 0,099
9	$4{,}29 \cdot 10^{-14}$	0,476 ± 0,090	0,6 ± 0,113	$6{,}87 \cdot 10^{-11}$	0,502 ± 0,076	0,591 ± 0,096	$5{,}93 \cdot 10^{-09}$	0,517 ± 0,066	0,093 ± 0,093
8	$2{,}24 \cdot 10^{-14}$	0,481 ± 0,088	0,605 ± 0,113	$2{,}92 \cdot 10^{-11}$	0,505 ± 0,074	0,595 ± 0,095	$2{,}67 \cdot 10^{-09}$	0,519 ± 0,065	0,594 ± 0,092
7	$9{,}57 \cdot 10^{-15}$	0,486 ± 0,086	0,611 ± 0,112	$1{,}03 \cdot 10^{-11}$	0,508 ± 0,073	0,599 ± 0,095	$1{,}05 \cdot 10^{-09}$	0,521 ± 0,064	0,597 ± 0,091
6	$3{,}21 \cdot 10^{-15}$	0,491 ± 0,083	0,617 ± 0,111	$2{,}99 \cdot 10^{-12}$	0,51 ± 0,071	0,603± 0,094	$3{,}05 \cdot 10^{-07}$	0,52 ± 0,063	0,6 ± 0,089
5	$8{,}35 \cdot 10^{-16}$	0,496 ± 0,080	0,623 ± 0,110	$7{,}12 \cdot 10^{-13}$	0,513 ± 0,069	0,606 ± 0,093	$1{,}08 \cdot 10^{-10}$	0,525 ± 0,062	0,602 ± 0,087
4	$1{,}74 \cdot 10^{-16}$	0,501 ± 0,076	0,628 ± 0,109	$1{,}45 \cdot 10^{-13}$	0,515 ± 0,066	0,609 ± 0,091	$3{,}06 \cdot 10^{-11}$	0,526 ± 0,061	0,604 ± 0,085
3	$3{,}37 \cdot 10^{-17}$	0,504 ± 0,071	0,631 ± 0,107	$2{,}90 \cdot 10^{-14}$	0,516 ± 0,063	0,611 ± 0,088	$9{,}34 \cdot 10^{-12}$	0,527 ± 0,060	0,605 ± 0,082
2	$8{,}42 \cdot 10^{-18}$	0,507 ± 0,067	0,633 ± 0,104	$8{,}29 \cdot 10^{-15}$	0,518 ± 0,061	0,611 ± 0,085	$4{,}46 \cdot 10^{-12}$	0,528 ± 0,060	0,605 ± 0,078
1	$7{,}60 \cdot 10^{-18}$	0,513 ± 0,063	0,634 ± 0,101	$1{,}52 \cdot 10^{-14}$	0,522 ± 0,061	0,612 ± 0,082	$1{,}38 \cdot 10^{-11}$	0,533 ± 0,061	0,606 ± 0,075
0,2	$1{,}35 \cdot 10^{-10}$	0,552 ± 0,095	0,655 ± 0,110	$7{,}39 \cdot 10^{-07}$	0,561 ± 0,100	0,633 ± 0,097	$8{,}37 \cdot 10^{-05}$	0,571 ± 0,103	0,628 ± 0,092
0	$3{,}07 \cdot 10^{-17}$	0,512 ± 0,051	0,633 ± 0,063	$8{,}69 \cdot 10^{-15}$	0,519 ± 0,066	0,612 ± 0,082	$4{,}51 \cdot 10^{-12}$	0,529 ± 0,066	0,605 ± 0,075
-0,2	0,0459	0,718 ± 0,426	0,816 ± 0,259	0,150	0,72 ± 0,420	0,79 ± 0,248	0,259	0,73 ± 0,422	0,78 ± 0,247
-1	0,166	1,28 ± 1,013	1,45 ± 0,675	0,180	1,28 ± 1,007	1,44 ± 0,665	0,220	1,29 ± 1,009	1,43 ± 0,661
-2	0,1189	1,36 ± 1,094	1,56 ± 0,741	0,160	1,36 ± 1,088	1,54± 0,724	0,190	1,37 ± 1,091	1,53 ± 0,720
-3	0,114	1,39 ± 1,122	1,6 ± 0,761	0,115	1,39 ± 1,115	1,58 ± 0,744	0,18	1,39 ± 1,118	1,57 ± 0,741
-4	0,109	1,41 ± 1,136	1,62 ± 0,771	0,14	1,40 ± 1,122	1,6 ± 0,755	0,18	1,41 ± 1,133	1,59 ± 0,751
-5	0,1048	1,41 ± 1,415	1,63 ± 0,778	0,13	1,41 ± 0,755	1,61 ± 1,129	0,17	1,42 ± 1,141	1,6 ± 0,758
-6	0,1011	1,42 ± 1,149	1,65 ± 0,781	0,13	1,42 ± 1,143	1,62 ± 0,766	0,17	1,42 ± 1,147	1,61 ± 0,763
-7	0,098	1,43 ± 1,154	1,66 ± 0,784	0,12	1,42 ± 1,147	1,63 ± 0,769	0,16	1,43 ± 1,151	1,62 ± 0,766
-8	0,0954	1,43 ± 1,156	1,66 ± 0,786	0,12	1,43 ± 1,151	1,64 ± 0,772	0,16	1,43 ± 1,154	1,62 ± 0,468
-9	0,0932	1,44 ± 1,159	1,67 ± 0,788	0,12	1,43 ± 1,153	1,64 ± 0,773	0,16	1,44 ± 1,157	1,63 ± 0,770
-10	0,0914	1,44 ± 1,160	1,68 ± 0,789	0,12	1,43 ± 1,155	1,65 ± 0,775	0,159	1,44 ± 1,159	1,64 ± 0,771

Tabla 3: Representación de los resultados tras el test de Student para todos los grados y todas las q, para secuencias con y sin intrones. Se muestran los p-valores y las medias de cada grupo. Además en amarillo está la celda del p-valor más bajo para cada grado.

Figura 6: Gráfica de los p-valores en función de q para los diferentes grados del polinomio. La línea roja representa el límite del p-valor para aceptar o rechazar la hipótesis nula.

Como se puede observar en la tabla 3 y en la imagen 6, para todos los grados, las q positivas dan lugar a valores de $h(q)$ que presenta diferencias significativas entre los grupos con y sin intrón. Existe un valor de q negativo ($q = -0,2$) del grado 1 que también se puede utilizar para esta distinción; sin embargo es con mucha diferencia la peor, ya que tiene un p-valor muy cercano a 0,05. Para el resto de valores $q < 0$ los p-valores son mayores que 0,05.

Utilizando un polinomio de grado 1, el valor $h(1)$ es el que tiene el menor p-valor con $7,6 \cdot 10^{-18}$. El siguiente es $h(2)$ con un un p-valor de $8,42 \cdot 10^{-18}$.

Para el polinomio de grado 2 el valor $h(2)$ tiene el menor p-valor siendo $8,29 \cdot 10^{-15}$, aunque también es bastante bueno $h(0)$ con un p-valor de $8,69 \cdot 10^{-15}$.

Para el polinomio de grado 3 el mejor p-valor es el de h(2) siendo este $4,46 \cdot 10^{-12}$, seguida muy de cerca por $h(0)$ con p-valor de $4,51 \cdot 10^{-12}$. Por lor tanto $h(1)$ de grado 1 es la mejor $h(q)$. Si se comparan estos resultados con la tabla de este mismo estudio para el parámetro α (tabla 2), se puede observar que el p-valor es de $1,11 \cdot 10^{-17}$, por lo que existe una mejor distinción entre los $h(1)$ de grado 1 que entre los valores α.

Debido a todos estos resultados se ve que el grado de polinomio que es mejor para distinguir los α es el grado 1 y la mejor q es 1. También se ve como el peor grado en este caso es el grado 3, cuyos p-valores son los más altos para cada q.

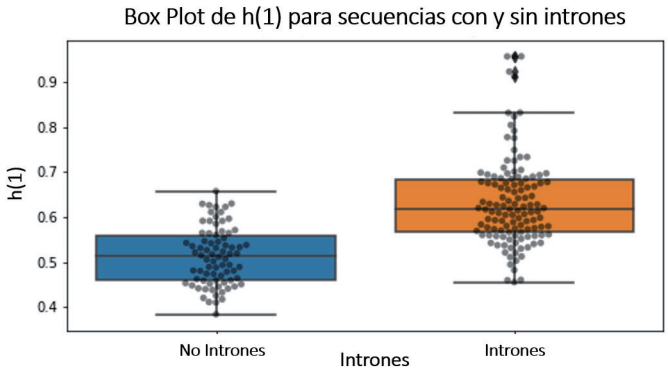


Figura 7: La figura muestra un box plot donde la caja azul representa la distribución de todos los $h(1)$ de las secuencias con intrones y la caja naranja las secuencias sin intrones,

Posteriormente se estudió las relaciones entre los reinos, para ello se hizo un test ANOVA (tabla 4 y cuando el p-valor es menor que 0,05 se realiza un test de Tukey para ver las relaciones de los reinos 2 a 2, tal y como nos muestra las tablas 9, 10, 11 del anexo A, donde solo aparecen representadas las parejas de reinos para las que el p-valor del test de Tukey es menor a 0,05.

Lo primero que se puede observar en estas tres tablas es que para los valores de q negativos no existen diferencias significativas entre los valores de $h(q)$ para los reinos.

En la tabla 4 también se puede observar como los mejores p-valores tras el ANOVA corresponden al los grados 1 y 2, mientras que los de grado 3 son mayores. Dentro de los grados encontramos que el mejor p-valor es para la $q = 0$ del polinomio de grado 2 y tiene un valor de $4{,}11 \cdot 10^{-13}$. Sin embargo no existe mucha diferencia con el p-valor de $q = 1$ y $q = 0$ del grado 1, siendo $7{,}98 \cdot 10^{-13}$ y $7{,}31 \cdot 10^{-13}$ respectivamente.

Pasando a las tablas del test de Tukey (tabla 9, tabla 10 y tabla 11) del anexo A, otra cosa que se puede llegar a deducir es que para todas las q positivas utilizando los polinomios de los 3 grados, se puede llegar a distinguir bien las $h(q)$ de animales con bacterias y de plantas con bacterias, con una excepción en el $q = 0,2$ del grado 3. Además en todos los casos el p-valor de Tukey es de 0,001, al igual que en el caso del estudio de α. Estas distinciones de nuevo pueden deberse a que en bacterias solo hay secuencias sin intrones, mientras que en animales y plantas, la inmensa mayoría de secuencias sí los tienen.

Por otro lado, se ve como entre algunas medias de $h(q)$ existen diferencias en la comparación fungi y bacterias; cuyo mejor p-valor es 0,025 correspondiente a $q = 9$ y a un polinomio de grado 1; este valor es mayor que el correspondiente al α que se obtuvo en el apartado del estudio de α, que era 0,009.

El resto de diferencias entre las $h(q)$ de reinos es entre animales con otros reinos. En primer lugar, hay medias de $h(q)$ que se distinguen para el reino animal y fungi, donde destacan las $q = 1$ y $q = 0$ en polinomio de grado 2 y las q positivas menores o iguales que 5 en el de grado 3, todas con un p-valor

de 0,001, por lo tanto cualquiera sería válida para ver esta distinción, y la media de los $h(q)$ está en 0,53 $\pm 0,057$, mientras que las de animales es 0,62 $\pm 0,02$.

Las $h(q)$ de los animales también se pueden distinguir de las de los protistas usando una q = 1 y el polinomio de grado 1 para hacer el MF-DFA, en este caso la media de las $h(1)$ de protista está en 0,54$\pm 0,124$ y la de animales es 0,637 $\pm 0,022$. No obstante, el p-valor de Tukey en este caso es de 0,048, muy cercano a 0,05, por lo que para poder afirmar esto último habría que trabajar con más secuencias de protistas.

Por último, existe la posibilidad de diferenciar entre las $h(q)$ de animales y plantas, en este caso el mejor p-valor es 0,017 y nos lo da $q = 0$ utilizando un polinomio de grado 3, donde las medias para las plantas están en 0,57$\pm 0,017$ y las de animales en 0,61$\pm 0,032$.

p-valor ANOVA			
Q	Polinomio de Grado 1	Polinomio de Grado 2	Polinomio de Grado 3
10	$4{,}96{\cdot}10^{-7}$	$5{,}31{\cdot}10^{-9}$	$1{,}92{\cdot}10^{-8}$
9	$3{,}83{\cdot}10^{-10}$	$3{,}4{\cdot}10^{-9}$	$1{,}05{\cdot}10^{-8}$
8	$2{,}65{\cdot}10^{-10}$	$1{,}77{\cdot}10^{-9}$	$5{,}09{\cdot}10^{-9}$
7	$1{,}62{\cdot}10^{-10}$	$8{,}36{\cdot}10^{-10}$	$2{,}17{\cdot}10^{-9}$
6	$5{,}44{\cdot}10^{-11}$	$3{,}35{\cdot}10^{-10}$	$8{,}02{\cdot}10^{-10}$
5	$3{,}71{\cdot}10^{-11}$	$1{,}12{\cdot}10^{-10}$	$2{,}56{\cdot}10^{-10}$
4	$1{,}38{\cdot}10^{-11}$	$3{,}08{\cdot}10^{-11}$	$7{,}17{\cdot}10^{-11}$
3	$4{,}55{\cdot}10^{-12}$	$7{,}20{\cdot}10^{-12}$	$1{,}88{\cdot}10^{-11}$
2	$1{,}53{\cdot}10^{-12}$	$1{,}69{\cdot}10^{-12}$	$5{,}94{\cdot}10^{-12}$
1	$7{,}98{\cdot}10^{-13}$	$1{,}05{\cdot}10^{-12}$	$7{,}25{\cdot}10^{-12}$
0,2	$1{,}65{\cdot}10^{-8}$	10^{-6}	0,000014
0	$7{,}31{\cdot}10^{-13}$	$4{,}11{\cdot}10^{-13}$	$4{,}36{\cdot}10^{-12}$
0,2	0,052	0,49	0,15
-1	0,091	0,077	0,09
-2	0,0615	0,069	0,08
-3	0,0596	0,066	0,08
-4	0,0578	0,064	0,08
-5	0,0561	0,062	0,08
-6	0,0578	0,06	0,08
-7	0,0561	0,059	0,08
-8	0,0547	0,058	0,08
-9	0,0534	0,057	0,08
-10	0,052	0,056	0,07

Tabla 4: Tabla de test ANOVA para los $h(q)$, para todos los grado de polinomio, para el estudio de los reinos. Se muestran los p-valores y en amarillo está el p-valor más bajo para cada grado.

Otro estudio de grupos realizado tras el MF-DFA ha sido el de las medias de $h(q)$ para las funciones

de las proteínas que dan lugar nuestras secuencias. Para ello se ha realizado un ANOVA (tabla 5) y para las q que daban un p-valor $< 0{,}05$ se ha realizado un test de Tukey (tablas 12, 13 y 14 del anexo B).

A diferencia de lo que ocurría en el estudio de α donde solo existían diferencias significativas en los valores de α para las funciones enzimática y estructural, con las $h(q)$ existen más diferencias. Las diferencias entre las $h(q)$ de las funciones existen para los grupos enzimática con estructural, enzimática con hormonal, enzimática con transporte, hormonal con movilidad y hormonal con defensiva.

Como se puede ver en la tabla 5 el mejor grado para hacer las distinciones es el grado 1 para la mayoría de q, aunque el grado 2 también es útil para algunas de las q. También como se muestra en la tabla 14 del anexo B, los valores del q que van del 10 al 7 no aparecen, esto se debe a que tras realizar el test de Tukey no había ninguna pareja de funciones para la que el p-valor fuese menor a 0,05.

Para diferenciar las $h(q)$ enzimática y estructural, la mejor opción es utilizar $q = 9$ con polinomio de grado 1, la cual nos da un p-valor de Tukey de 0,004 y que además, es ligeramente mejor que el del α que era de 0,005. En el caso de $q = 9$ la media es de $0{,}507 \pm 0{,}11$ para las proteínas enzimáticas y $0{,}621 \pm 0{,}12$ para las estructurales.

Para distinguir entre la media de $h(q)$ de función enzimática y hormonal, el mejor valor de q que se puede tomar es el 0 con un polinomio de grado 2, cuya comparación nos da un p-valor de 0,009, y la media de las proteínas enzimáticas es $0{,}549 \pm 0{,}097$, y la de las hormonales $0{,}618 \pm 0{,}084$.

Si se desean estudiar los $h(q)$ entre secuencias que codifican para proteínas enzimáticas y secuencias que dan proteínas de transporte las mejores opciones son tomar los valores de $q = 8$, $q = 9$ o $q = 10$ y utilizar un polinomio de grado 2, las tres q nos dan un p-valor de 0,009, y en ambos casos la media de la función enzimática está en $0{,}52 \pm 0{,}097$ mientras que la de la función de transporte es $0{,}615 \pm 0{,}09$.

También se puede realizar una diferenciación entre la $h(q)$ de función hormonal con la de movilidad, y en este caso la mejor opción es la de tomar una $q = 0, 2$ para un polinomio de grado 2, aunque también se puede coger el mismo valor de q pero con un polinomio de grado 3, en ambos casos nos da un p-valor de 0,005 y la media de hormonal es $0{,}55 \pm 0{,}11$, mientras que la de movilidad es de $0{,}66 \pm 0{,}082$.

Para terminar con las funciones, se hizo el análisis de medias de $h(q)$ para la función hormonal y la de defensa y el resultado fue que la mejor opción es tomar un valor de $q = 0$ con polinomio de grado 2, dándonos un p-valor de 0,037, aunque también se podría tomar la misma q pero con el grado del polinomio 1, en este caso el p-valor es 0,038, pero la diferencia de medias de $h(q)$ es mayor, aunque en ambos casos la media de la función de defensa es $0{,}55 \pm 0{,}07$, la media de la función hormonal es 0,63 $\pm$ 0,097 para el grado 1 y $0{,}618 \pm 0{,}084$ para el grado 2.

p-valor ANOVA			
Q	Polinomio de Grado 1	Polinomio de Grado 2	Polinomio de Grado 3
10	0,01	0,0008	0,012
9	0,000626	0,0008	0,011
8	0,000643	0,0009	0,01
7	0,000666	0,0009	0,01
6	0,000696	0,0009	0,009
5	0,000751	0,0009	0,009
4	0,000846	0,001	0,008
3	0,001025	0,001	0,008
2	0,001358	0,0014	0,008
1	0,00179	0,0016	0,008
0,2	0,001337	0,0017	0,003
0	0,0017	0,00116	0,007
-0,2	0,167	0,14	0,2
-1	0,09	0,078	0,09
-2	0,0683	0,072	0,08
-3	0,067	0,069	0,08
-4	0,0665	0,067	0,08
-5	0,066	0,065	0,07
-6	0,0658	0,064	0,07
-7	0,06565	0,063	0,07
-8	0,06555	0,063	0,07
-9	0,06545	0,062	0,07
-10	0,06537	0,062	0,07

Tabla 5: Tabla de ANOVA tras MF-DFA para todos los grado de polinomio para el estudio de las funciones. En amarillo están destacados los p-valores más bajos para cada grado del polinomio

Tras el estudio de las funciones, se realizó el estudio de las diferencias significativas entre dos de las clases dentro del reino planta para ver si se podía distinguir entre las $h(q)$ de monocotiledóneas y dicotiledóneas. Una vez realizada la prueba de T-Student, se llegó a la conclusión de que, ni para ningún grado de polinomio ni para ninguna q se podían distinguir las medias, ya que los p-valores para todas las q y grados era mayor a 0,05, tal y como muestra la tabla 15 en el anexo C.

A continuación, se intentó ver si existía alguna diferenciar las medias de las $h(q)$ de los humanos de las del resto de animales, para ello se hizo un test T-Student, mediante el cuál se pude discernir que utilizando una $q = 10$ y un polinomio de grado 1, se pueden distinguir estas medias, ya que es de 0,69 $\pm$ 0,3 para humanos, mientras que para el resto de animales es 0,58 $\pm$ 0,14, en este caso, solo había otra q para la que los p-valores eran menor a 0,05, y es el caso de $q = 0$ para grados 1 y 2, donde sus p-valores eran 0,033 y 0,04 respectivamente.

El último análisis estadístico que se realizó con los datos obtenidos, mediante el método de MF-DFA, fue entre los $h(q)$ de bacterias, como muestra la tabla 6, de nuevo se tomaron solo dos clases (bacilli y gammaproteobacteria) para este estudio, ya que son las clases dentro de bacteria para las que se tenía un mayor número de secuencias. Tras un estudio de T-Student se llegó a la conclusión de que cualquier valor de q negativo tiene un p-valor menor que 0,05 (al contrario de los estudios anteriores donde los p-valores menores a 0,05 se correspondían con las q positivas), aunque principalmente, si se usan polinomios de grado 2 o 3, cualquier q de -1 hasta -10 nos dan unos resultados que como muestra la tabla 6 tienen todos p-valores menores o iguales a 0,0079. Todas las q menores o iguales que -1 y de grados 2 y 3, dan el p-valor más bajo (p-valor=0,002), en estos casos la media de bacilli es 1,83 $\pm 0,071$ y la de gammaproteobacterias es 0,9 $\pm 0,027$.

Grado 1				Grado 2				Grado 3			
Q	p-valor	Media B	Media G	Q	p-valor	Media B	Media G	Q	p-valor	Media B	Media G
10	0,59	0,47 ± 0,083	0,44 ± 0,09	10	0,41	0,5 ± 0,048	0,48± 0,068	10	0,13	0,53 ± 0,048	0,5 ± 0,057
9	0,59	0,47 ± 0,089	0,45 ± 0,088	9	0,39	0,5 ± 0,047	0,48 ± 0,073	9	0,12	0,53 ± 0,049	0,5 ± 0,056
8	0,59	0,48 ± 0,086	0,46 ± 0,088	8	0,35	0,51 ± 0,053	0,49 ± 0,071	8	0,11	0,53 ± 0,048	0,5 ± 0,058
7	0,58	0,49 ± 0,083	0,47 ± 0,087	7	0,32	0,51± 0,051	0,49 ± 0,071	7	0,1	0,54 ± 0,047	0,5 ± 0,058
6	0,58	0,49 ± 0,086	0,47 ± 0,089	6	0,28	0,51 ± 0,054	0,49 ± 0,070	6	0,095	0,54 ± 0,048	0,51 ± 0,056
5	0,57	0,5 ± 0,079	0,48 ± 0,093	5	0,24	0,52 ± 0,050	0,5 ± 0,073	5	0,09	0,54 ± 0,049	0,51 ± 0,059
4	0,56	0,5 ± 0,085	0,48 ± 0,098	4	0,2	0,52 ± 0,048	0,5 ± 0,069	4	0,08	0,55 ± 0,048	0,51 ± 0,063
3	0,55	0,5 ± 0,089	0,49 ± 0,092	3	0,17	0,53 ± 0,053	0,5 ± 0,074	3	0,08	0,55 ± 0,048	0,51 ± 0,065
2	0,54	0,51 ± 0,088	0,49 ± 0,095	2	0,14	0,53 ± 0,057	0,5 ± 0,077	2	0,08	0,55 ± 0,051	0,51 ± 0,063
1	0,54	0,51 ± 0,092	0,5 ± 0,091	1	0,15	0,54 ± 0,056	0,5 ± 0,074	1	0,09	0,56 ± 0,049	0,52 ± 0,062
0,2	0,74	0,54 ± 0,095	0,53 ± 0,093	0,2	0,47	0,57 ± 0,059	0,54 ± 0,071	0,2	0,37	0,59 ± 0,053	0,55 ± 0,064
0	0,24	0,52 ± 0,088	0,49 ± 0,09	0	0,03	0,544 ± 0,065	0,498 ± 0,065	0	0,022	0,565 ± 0,059	0,509 ± 0,063
-0,2	0,0079	0,91 ± 0,86	0,58± 0,28	-0,2	0,004	0,93 ± 0,039	0,58 ± 0,065	-0,2	0,005	0,94 ± 0,056	0,59 ± 0,059
-1	0,0029	1,82 ± 0,98	0,91 ± 0,58	-1	0,002	1,83 ± 0,100	0,9 ± 0,059	-1	0,002	1,84 ± 1,02	0,91 ± 0,058
-2	0,0028	1,95 ± 1,07	0,95 ± 0,57	-2	0,002	1,95 ± 1,07	0,94 ± 0,053	-2	0,002	1,97 ± 0,97	0,95 ± 0,050
-3	0,0027	1,99± 1,10	0,97± 0,58	-3	0,002	2 ± 1,10	0,96 ± 0,057	-3	0,002	2,01 ± 1,08	0,97 ± 0,57
-4	0,0027	2,02 ± 1,10	0,974 ± 0,55	-4	0,002	2,02 ± 1,10	0,97 ± 0,057	-4	0,002	2,03 ± 1,07	0,98 ± 0,056
-5	0,0027	2,03 ± 1,15	0,98 ± 0,55	-5	0,002	2,03 ± 1,09	0,98 ± 0,057	-5	0,002	2,05 ± 1,08	0,98 ± 0,057
-6	0,0027	2,04 ± 1,09	0,98 ± 0,59	-6	0,002	2,04 ± 1,11	0,98 ± 0,058	-6	0,002	2,06 ± 1,05	0,98 ± 0,057
-7	0,0027	2,05 ± 1,06	0,99 ± 0,56	-7	0,002	2,05 ± 1,12	0,98 ± 0,056	-7	0,002	2,07 ± 1,06	1 ± 0,059
-8	0,0027	2,058 ± 1,09	0,99 ± 0,53	-8	0,002	2,06 ± 1,12	0,98 ± 0,056	-8	0,002	2,08 ± 1,09	1 ± 0,056
-9	0,0027	2,064 ± 1,14	0,996 ± 0,57	-9	0,002	2,06 ± 1,08	0,98 ± 0,058	-9	0,002	2,08 ± 1,06	1 ± 0,056
-10	0,003	2,069 ± 1,14	0,999 ± 0,62	-10	0,002	2,07 ± 1,10	0,99 ± 0,059	-10	0,002	2,09 ± 1,08	1,004 ± 0,054

Tabla 6: Tabla que contiene los p-valores tras el test T-Student de todas las q en los grados 1,2 y 3 para las secuencias de bacilli (B) y gammaproteobacteriaG. En amarillo están destacados los p-valores menores para cada grado.

3.3. Clasificación de secuencias con aprendizaje automático

Para terminar se utilizaron métodos de aprendizaje automático para ver si se podían clasificar las secuencias en base a todos los $h(q)$ por una parte y en base a todos los $h(q)$ junto al parámetro α por otro lado. Esto se realizó para cada grado de polinomio, para ver si existe un grado mejor que otro para el agrupamiento.

3.3.1. K-means

Los estudios realizados fueron con los grupos que aparecían en la tabla 1.

Para el k-means se utilizó toda la población de secuencias para ver si se podían agrupar en base a los intrones, a los reinos o a la función. Por otro lado se utilizó la población de monocotiledóneas y dicotiledóneas para estudiar si se podían distinguir estos grupos en el *clustering*, También se utilizó el grupo animales para ver si existía la posibilidad de diferenciar los humanos del resto de animales en los grupos formados por el método. Por último se utilizaron las secuencias de bacterias para ver si se podían diferenciar en base a bacilli y gammaproteobacterias. Por otra parte en el caso de los reinos se realizaron 3 estudios más, se trató de ver si se podían diferenciar los animales del resto de reinos, las plantas del resto o las bacterias del resto. Para el conjunto de bacterias se hizo algo parecido y se hicieron 2 estudios más, uno para comprobar si se podían distinguir los bacilli del resto de bacterias y el otro para hacer lo mismo con las gammaproteobacterias.

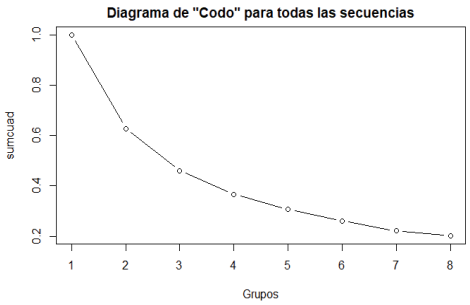
Para cada población descrita se hizo un k-means, en todos los casos con un $k = 2$, a excepción de los reinos, que se hizo con un $k = 5$ y la función con un $k = 8$. Los valores de k elegidos son el número de grupos en cada clasificación. Después se utilizó el índice de Rand con los grupos obtenidos para cuantificar la clasificación de los grupos mencionados. En todos estos casos, como muestran las tablas 7 y las tablas del anexo D 16 y 17 el método k-means nos da unos índices de Rand menores a 0,14. Esto puede indicar que no se puede utilizar los parámetros α y $h(q)$ para distinguir los grupos que nos interesan, o que es necesaria una mayor cantidad de datos, ya sean más muestras, otras q u otros grados de polinomio para el MF-DFA o incluso que haya que utilizar valores de q en un rango en vez de todas.

Grado	Intrones	Plantas: Monocotiledónea/Dicotiledónea	Animales: Humano/No Humano	Funciones
1	-0,016	0,06	0,14	0,01
2	0,02	0,027	0,1	0,0066
3	0,002	0,027	0,1	0,01
$1 + \alpha$	0,007	0,059	0,14	0,0082
$2 + \alpha$	0,02	0,002	0,013	0,0084
$3 + \alpha$	0,011	0,027,	0,107	0,019

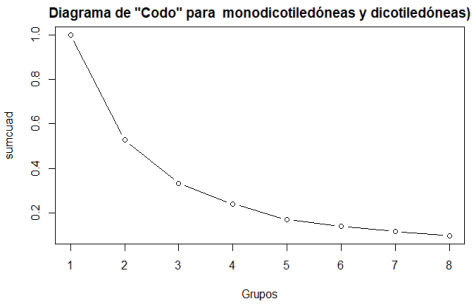
Tabla 7: Gráfica de los índices de Rand del método k-means para cada Grado del polinomio del MF-DFA y para cada Grado + el α . Este método de *clustering* se aplica sobre presencia o ausencia de intrones, diferencia entre monocotiledónea y dicotiledónea , deferencia entre humano y el resto de animales y funciones

Debido a estos resultado para k-means se hizo un diagrama de codo con cada conjunto de datos usados para ver cuál es el número óptimo de grupos para el conjunto de datos. Estos conjuntos eran todas las secuencias de la base de datos (usadas para el estudio de intrones y de reinos); las de plantas

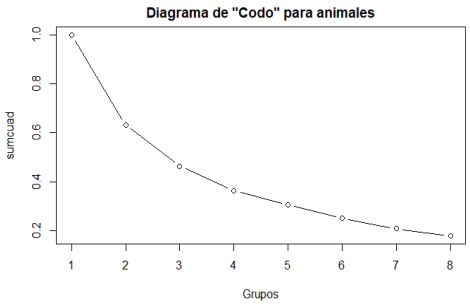
monodicotiledóneas y dicotiledóneas; las de animales y las de bacterias. Estos diagramas se encuentran representados en las imágenes 8, 9 y 10. En los diagramas se puede ver cómo para todas las secuencias en todos los grados y para todas las poblaciones de datos usadas habría que coger un numero de grupos $k = 3$ ó $k = 4$. Los números de grupos del trabajo son $k = 2$, $k = 5$, $k = 7$. Se puede decir que el algoritmo de k-means no agrupa las secuencias en los grupos propuestos usando los parámetros citados.



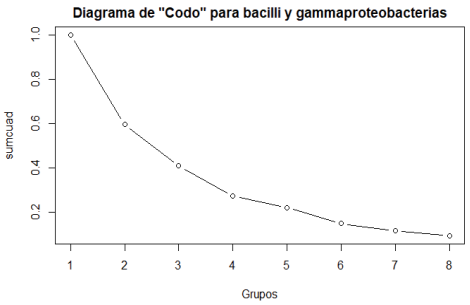
(a) D.codo de todas las secuencias. Grado 1



(b) D. codo para todas las plantas. Grado 1

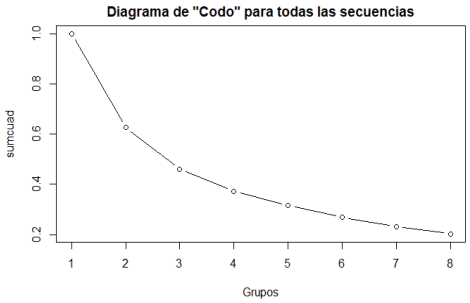


(c) D. codo para todos los animales. Grado 1

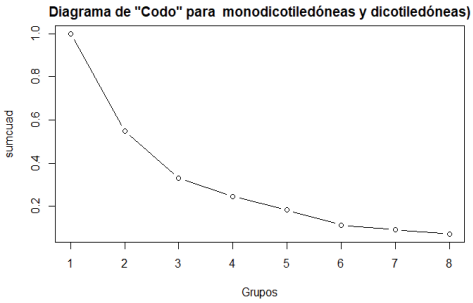


(d) D. codo para todas las bacterias. Grado 1

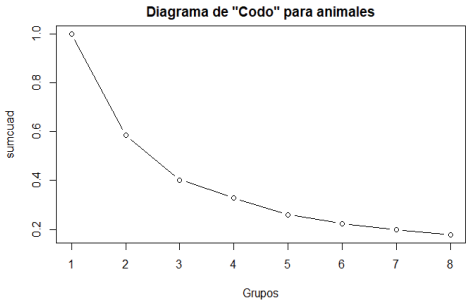
Figura 8: Diagramas de codo para cada población de secuencias estudiada con las $h(q)$ obtenidas con un MF-DFA que se ha hecho con un polinomio grado 1.



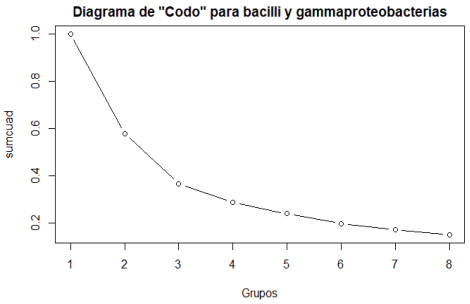
(a) D.codo de todas las secuencias. Grado 2



(b) D. codo para todas las plantas. Grado 2

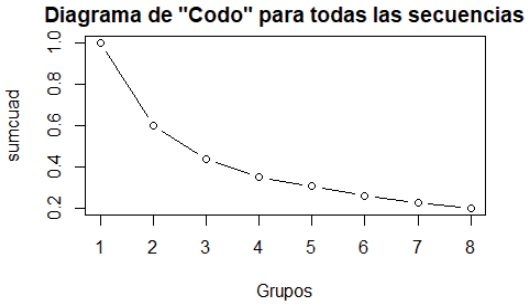


(c) D. codo para todos los animales. Grado 2

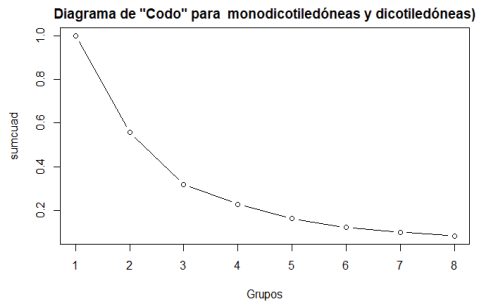


(d) D. codo para todas las bacterias. Grado 2

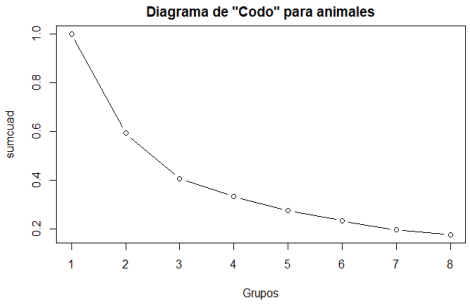
Figura 9: Diagramas de codo para cada población de secuencias estudiada con las $h(q)$ obtenidas con un MF-DFA que se ha hecho con un polinomio de grado 2.



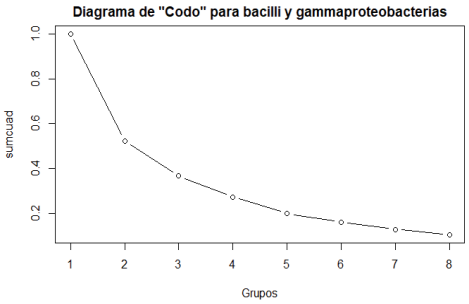
(a) D.codo de todas las secuencias. Grado 1



(b) D. codo para todas las plantas. Grado 3



(c) D. codo para todos los animales. Grado 3



(d) D. codo para todas las bacterias. Grado 3

Figura 10: Diagramas de codo para cada población de secuencias estudiada con las $h(q)$ obtenidas con un MF-DFA que se ha hecho con un polinomio de grado 3.

3.3.2. Redes neuronales

Por último se emplearon redes neuronales para poder ver si era posible la clasificación según los grupos descritos anteriormente, junto a nuevos grupos que son dentro de los reinos: animales y resto de reinos, plantas y resto de reinos y por último Bacterias y resto de reinos. En el caso de bacterias también se hizo modificaciones, ya que se estudió la clasificación de todas las clases de bacterias y luego se intenta distinguir gammaprotobacterias del resto de bacterias y bacilli del resto de bacterias.

Para el estudio utilizaron redes neuronales de dos capas con 10 y 6 neuronas cada una. El entrenamiento se hizo con un 70 % de los elementos de cada grupo, y el test con el 30 % restante. Para ver la calidad de las clasificaciones, se calculó la tasa de error (fórmula 10).

En las tablas 8 y las tablas del anexo E 18 y 19 se puede ver como al añadir el parámetro α, la tasa de error aumenta para todos los estudios con todos los grados salvo para los de bacterias. Para la clasificación de intrones se ve como para el grado 3 la tasa de error sube hasta el 0,5, frente al 0,3 y 0,28 de los grados 1 y 2.

Grado	Intrones	Plantas: Monocotiledónea/Dicotiledónea	Animales: Humano/No Humano	Funciones
1	0,3	0,3	0,38	0,86
2	0,28	0,3	0,35	0,9
3	0,5	0,3	0,35	0,85
$1 + \alpha$	0,4	0,3	0,68	0,81
$2 + \alpha$	0,4	0,5	0,44	0,95
$3 + \alpha$	0,55	0,8	0,59	0,91

Tabla 8: Gráfica de las tasas de error de la red neuronal para cada Grado del polinomio del MF-DFA y para cada Grado + el α . Estos métodos de clasificación se aplican sobre presencia o ausencia de intrones, diferencia entre monocotiledónea y dicotiledónea , deferencia entre humano y el resto de animales y funciones

En las tablas 18 y 19 del anexo E se encuentran las menores tasa de error de todos los estudios, ambas tienen un valor de 0,18 y corresponden a la clasificación de bacilli frente al resto de bacterias para los $h(q)$ sin α y gammaproteobacteria frente al resto de bacterias para los $h(q)$ con α.

A pesar de esto, cabe la posibilidad de que si la base de datos hubiese presentado una mayor cantidad de secuencias para el entrenamiento, se podría llegar a mejores tasas de error.

CAPÍTULO 4. CONCLUSIONES

1. Existen diferencias en los valores α para el estudio de presencia de intrones en las secuencias de la base de datos ($\overline{\alpha}_{sinintrones} = 0,486 \pm 0,072$, $\overline{\alpha}_{conintrones} = 0,6 \pm 0,09$, p-valor $= 1,11 \cdot 10^{-17}$, como ya dijo Peng [9].

2. Existen diferencias de los α para secuencias de animales y bacterias ($\overline{\alpha}_{animal} = 0,597 \pm 0,093$, $\overline{\alpha}_{bacteria} = 0,467 \pm 0,067$, p-valor = 0,001), fungi y bacterias ($\overline{\alpha}_{fungi} = 0,548 \pm 0,052$, $\overline{\alpha}_{bacteria} = 0,467 \pm 0,067$, p-valor = 0,009), y plantas y bacterias ($\overline{\alpha}_{planta} = 0,579 \pm 0,096, \overline{\alpha}_{bacteria} = 0,467 \pm 0,067, p - valor = 0,001$). Para las funciones se pueden distinguir los α de enzimática y estructural ($\overline{\alpha}_{enzimatica} = 0,514 \pm 0,095$, $\overline{\alpha}_{estructural} = 0,612 \pm 0,123$, p-valor = 0,005).

3. Hay diferencias significativas en los valores de $h(q)$ para los grupos que se han utilizado.En el caso de los intrones usando $h(1)$ con polinomio de grado 1. Para animales y plantas respecto a bacterias se pueden usar todas las q positivas de todos los grados, con la excepción de $q = 0, 2$ del grado 3. Las diferencias significativas de $h(q)$ entre fungi y bacteria se encuentran utilizando una $q = 9$ con el polinomio de grado 1. Los parámetros $h(1)$ y $h(0)$ del polinomio de grado 2 y las q positivas de grado 3 son significativamente distintos para animales y fungi. La media de los $h(1)$ con polinomio de grado 1 son distintos para animales y protisitas. También se pueden distinguir las $h(q)$ de animales y plantas siendo el mejor valor de q 0 con un polinomio de grado 3. Para las funciones de la proteína $h(q)$ tiene diferencias en los casos de función enzimática respecto a estructural ($q = 9$ y polinomio de grado 1), enzimática con hormonal ($q = 0$ y polinomio de grado 2), enzimáticas con transporte ($q = 8, q = 9 y q = 10$ y polinomio de grado 2), para la función hormonal hay diferencias significativas de $h(q)$ con las funciones de movilidad ($q = 0, 2$ y polinomio de grado2 o grado 3) y defensa ($q = 0$ y polinomio de grado 2). Por último también hay diferencias de los $h(q)$ para bacilli y gammaproteobacterias, para ello hay que utilizar cualquier valor de q entre -1 y -10 y polinomios de grado 2 o 3.

4. Una clasificación usando el método de k-means obtiene los índices de Rand menores a 0,14. Usando redes neuronales las tasas de error son superiores a 0,18.

CAPÍTULO 5. BIBLIOGRAFIA

Referencias

[1] EMBL-EBI, actualizado en 2022, Ensembl. Disponible en Internet: https://www.ncbi.nlm.nih.gov/gene/, consultado el 15-06-2022.

[2] EMBL-EBI, actualizado en 2022, Ensembl Bacteria. Disponible en Internet: https://bacteria.ensembl.org/index.html, consultado el 15-06-2022.

[3] EMBL-EBI, actualizado en 2022, Ensembl Fungi. Disponible en Internet: https://fungi.ensembl.org/index.html, consultado el 15-06-2022.

[4] EMBL-EBI, actualizado en 2022, Ensembl Plants. Disponible en Internet: https://plants.ensembl.org/index.html, consultado el 15-06-2022.

[5] J. Ollé, 2019, Economispedia, Disponible en Internet: https://conceptosclaros.com/que-es-el-p-valor/, consultado el 3-07-2022.

[6] J. W Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde & H. E. Stanley, Multifractal detrended fluctuation analysis of nonstationary time series, *Physica A: Statistical Mechanics and its Applications*, 316(1-4), pp.87-114, 2002.

[7] National Center for Biotechnology Information Bethesda (MD): National Library of Medicine (US), Actualizado en 2022, Gene [internet]. Disponible en Internet: https://www.ncbi.nlm.nih.gov/gene/, consultado el 15-06-2022.

[8] NIH National Human Genome Research Institute, aactualizado en 2022. Disponible en Internet: https://www.ncbi.nlm.nih.gov/gene/, consultado el 8-06-2022.

[9] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons & H.E. Stanley, Long-range correlations in nucleotide sequences, *Nature*, 356(6365), pp.168-170, 1992.

[10] Wikipedia, actualizado en 2022, Wikipedia:Serie Temporal, Disponible en Internet: https://es.wikipedia.org/wiki/Serie_temporal, consultado el 01-06-2022.

[11] Thomas D. Wilson, 2001, Information overload: implications for healthcare services, *Health Informatics Journal*, 7(2), pp.112-117.

ANEXO A. TABLAS TEST DE TUKEY PARA REINOS

Grado 1					
Q	Reino 1	Reino 2	Media 1	Media 2	p-valor Tukey
10	Animal	Bacteria	0,614± 0,032	0,449 ± 0,029	0,001
	Planta	Bacteria	0,572 ± 0,022	0,449 ± 0,029	0,001
9	Animal	Bacteria	0,596± 0,034	0,454 ± 0,030	0,001
	Planta	Bacteria	0,546 ± 0,22	0,454 ± 0,030	0,001
	Fungi	Bacteria	0,577 ± 0,049	0,454 ± 0,030	0,025
8	Animal	Bacteria	0,602 ± 0,026	0,46 ± 0,031	0,001
	Planta	Bacteria	0,582 ± 0,021	0,46± 0,031	0,001
	Fungi	Bacteria	0,55 ± 0,043	0,46 ± 0,031	0,026
7	Animal	Bacteria	0,608± 0,024	0,465± 0,022	0,001
	Planta	Bacteria	0,587 ± 0,019	0,465 ± 0,022	0,001
	Fungi	Bacteria	0,554 ± 0,039	0,465 ± 0,022	0,028
6	Animal	Bacteria	0,614 ± 0,038	0,471± 0,020	0,001
	Planta	Bacteria	0,592 ± 0,019	0,471± 0,020	0,001
	Fungi	Bacteria	0,558 ± 0,034	0,471 ± 0,020	0,031
5	Animal	Bacteria	0,621± 0,033	0,477± 0,027	0,001
	Planta	Bacteria	0,596 ± 0,023	0,477 ± 0,027	0,001
	Fungi	Bacteria	0,56 ± 0,041	0,477± 0,027	0,037
4	Animal	Bacteria	0,626± 0,019	0,482± 0,021	0,001
	Planta	Bacteria	0,599 ± 0,016	0,482± 0,021	0,001
	Fungi	Bacteria	0,562 ± 0,028	0,482± 0,021	0,046
3	Animal	Bacteria	0,631 ± 0,030	0,487± 0,028	0,001
	Planta	Bacteria	0,601 ± 0,024	0,487± 0,028	0,001
2	Animal	Fungi	0,635 ± 0,028	0,56± 0,032	0,028
	Animal	Bacteria	0,635 ± 0,028	0,492± 0,024	0,001
	Planta	Bacteria	0,6 ± 0,019	0,493 ± 0,024	0,001
1	Animal	Fungi	0,637 ± 0,022	0,558 ± 0,042	0,01
	Animal	Bacteria	0,637 ± 0,022	0,499 ± 0,024	0,001
	Animal	Protista	0,637 ± 0,022	0,544 ± 0,124	0,048
	Planta	Bacteria	0,599 ± 0,017	0,499 ± 0,024	0,001
0,2	Animal	Fungi	0,661± 0,020	0,569 ± 0,038	0,011
	Animal	Bacteria	0,661 ± 0,020	0,537 ± 0,019	0,001
	Planta	Bacteria	0,628 ± 0,017	0,537 ± 0,019	0,001
0	Animal	Fungi	0,639 ± 0,026	0,556± 0,036	0,007
	Animal	Bacteria	0,639 ± 0,026	0,498 ± 0,020	0,001
	Planta	Bacteria	0,593± 0,015	0,498± 0,020	0,001

Tabla 9: Tabla de resultados de reino tras el test de Tukey para todas las q en grado 1 que dieron un p-valor en el test ANOVA menor que 0,05 con las medias de cada grupo. Solo los valores que den p-valor menor que 0,05 para el test de Tukey se muestran en esta tabla.

Grado 2					
Q	Reino 1	Reino 2	Media 1	Media 2	p-valor Tukey
10	Animal	Bacteria	0,588± 0,028	0,477 ± 0,031	0,001
	Planta	Bacteria	0,572 ± 0,022	0,477 ± 0,031	0,001
9	Animal	Bacteria	0,592± 0,025	0,481± 0,029	0,001
	Planta	Bacteria	0,575± 0,019	0,481± 0,029	0,001
8	Animal	Bacteria	0,597± 0,027	0,484 ± 0,027	0,001
	Planta	Bacteria	0,578± 0,018	0,484 ± 0,027	0,001
7	Animal	Bacteria	0,601± 0,022	0,488± 0,024	0,001
	Planta	Bacteria	0,581± 0,013	0,488± 0,024	0,001
6	Animal	Bacteria	0,606± 0,028	0,491± 0,030	0,001
	Planta	Bacteria	0,583± 0,019	0,491± 0,030	0,001
5	Animal	Bacteria	0,61± 0,024	0,494± 0,026	0,001
	Planta	Bacteria	0,584± 0,016	0,494± 0,026	0,001
4	Animal	Fungi	0,614± 0,021	0,55 ± 0,049	0,033
	Animal	Bacteria	0,614± 0,021	0,497± 0,026	0,001
	Planta	Bacteria	0,585± 0,018	0,497± 0,026	0,001
3	Animal	Fungi	0,616± 0,026	0,548± 0,053	0,013
	Animal	Bacteria	0,616± 0,026	0,5± 0,029	0,001
	Planta	Bacteria	0,584± 0,021	0,5± 0,029	0,001
2	Animal	Fungi	0,618± 0,028	0,546± 0,058	0,005
	Animal	Bacteria	0,618± 0,028	0,503± 0,031	0,001
	Planta	Bacteria	0,582± 0,021	0,503± 0,031	0,001
1	Animal	Fungi	0,62± 0,021	0,544± 0,057	0,001
	Animal	Bacteria	0,62± 0,021	0,508± 0,023	0,001
	Animal	Planta	0,62± 0,021	0,581± 0,019	0,043
	Planta	Bacteria	0,581± 0,019	0,508± 0,023	0,001
0,2	Animal	Fungi	0,644± 0,028	0,557± 0,046	0,007
	Animal	Bacteria	0,644± 0,028	0,545± 0,023	0,001
	Planta	Bacteria	0,613± 0,020	0,545± 0,023	0,001
0	Animal	Fungi	0,621 ± 0,020	0,543 ±0, 057	0,001
	Animal	Bacteria	0,621± 0,020	0,505± 0,024	0,001
	Animal	Planta	0,621± 0,020	0,578 ± 0,016	0,023
	Planta	Bacteria	0,578± 0,016	0,505± 0,024	0,001

Tabla 10: Tabla de resultados de reino tras el test de Tukey para todas las q en grado 2 que dieron un p-valor en el test ANOVA menor que 0,05 con las medias de cada grupo. Solo los valores que den p-valor menor que 0,05 para el test de Tukey se muestran en esta tabla.

Grado 3					
Q	Reino 1	Reino 2	Media 1	Media 2	p-valor Tukey
10	Animal	Fungi	0,594 ± 0,028	0,522 ± 0,045	0,01
	Animal	Bacteria	0,594 ± 0,028	0,499 ± 0,024	0,001
	Planta	Bacteria	0,575 ± 0,018	0,499 ± 0,024	0,001
9	Animal	Fungi	0,598 ± 0,034	0,524 ± 0,046	0,008
	Animal	Bacteria	0,598 ± 0,034	0,501 ± 0,022	0,001
	Planta	Bacteria	0,577 ± 0,019	0,501 ± 0,022	0,001
8	Animal	Fungi	0,601 ± 0,031	0,526 ± 0,051	0,006
	Animal	Bacteria	0,601 ± 0,031	0,504 ± 0,021	0,001
	Planta	Bacteria	0,579 ± 0,017	0,504 ± 0,021	0,001
7	Animal	Fungi	0,605 ± 0,029	0,528 ± 0,040	0,004
	Animal	Bacteria	0,605± 0,029	0,506 ± 0,019	0,001
	Planta	Bacteria	0,581 ± 0,013	0,506 ± 0,019	0,001
6	Animal	Fungi	0,608 ± 0,032	0,53 ± 0,052	0,002
	Animal	Bacteria	0,608 ± 0,032	0,508 ± 0,021	0,001
	Planta	Bacteria	0,582 ± 0,012	0,508 ± 0,021	0,001
5	Animal	Fungi	0,611 ± 0,030	0,531 ± 0,056	0,001
	Animal	Bacteria	0,611 ± 0,030	0,511 ± 0,020	0,001
	Planta	Bacteria	0,582 ± 0,015	0,511 ± 0,020	0,001
4	Animal	Fungi	0,613 ± 0,037	0,531 ± 0,057	0,001
	Animal	Bacteria	0,613 ± 0,037	0,512 ± 0,018	0,001
	Planta	Bacteria	0,582 ± 0,019	0,512 ± 0,018	0,001
3	Animal	Fungi	0,614 ± 0,027	0,532 ± 0,055	0,001
	Animal	Bacteria	0,614 ± 0,027	0,514 ± 0,017	0,001
	Planta	Bacteria	0,58 ± 0,011	0,514 ± 0,017	0,001
2	Animal	Fungi	0,615 ± 0,034	0,532 ± 0,058	0,001
	Animal	Bacteria	0,615 ± 0,034	0,516 ± 0,015	0,001
	Animal	Planta	0,615 ± 0,034	0,578 ± 0,016	0,04
	Planta	Bacteria	0,578 ± 0,016	0,516 ± 0,015	0,001
1	Animal	Fungi	0,617 ± 0,031	0,534 ± 0,056	0,001
	Animal	Bacteria	0,617 ± 0,031	0,522 ± 0,019	0,001
	Animal	Planta	0,617 ± 0,031	0,578 ± 0,011	0,023
	Planta	Bacteria	0,578 ± 0,011	0,522 ± 0,019	0,001
0,2	Animal	Fungi	0,641 ± 0,036	0,549 ± 0,057	0,039
	Animal	Bacteria	0,641 ± 0,036	0,558 ± 0,019	0,001
0	Animal	Fungi	0,617 ± 0,032	0,535 ± 0,057	0,001
	Animal	Bacteria	0,617 ± 0,032	0,517 ± 0,023	0,001
	Animal	Planta	0,617 ± 0,032	0,576 ± 0,017	0,017
	Planta	Bacteria	0,576 ± 0,017	0,517 ± 0,023	0,001

Tabla 11: Tabla de resultados de reino tras el test de Tukey para todas las q en grado 3 que dieron un p-valor en el test ANOVA menor que 0,05 con las medias de cada grupo. Solo los valores que den p-valor menor que 0,05 para el test de Tukey se muestran en esta tabla.

ANEXO B. TABLAS TEST DE TUKEY PARA FUNCIONES

Grado 1					
Q	Función 1	Función 2	Media Función 1	Media Funcióneino 2	p-valor Tukey
10	Enzimática	Hormonal	0,502 ± 0,13	0,619 ± 0,095	0,027
9	Enzimática	Estructural	0,507 ± 0,11	0,621 ± 0,120	0,004
	Enzimática	Transporte	0,507± 0,11	0,606 ± 0,102	0,042
8	Enzimática	Estructural	0,513± 0,14	0,625 ± 0,114	0,005
	Enzimática	Transporte	0,513± 0,14	0,611 ± 0,092	0,041
7	Enzimática	Estructural	0,518 ± 0,16	0,629 ± 0,126	0,005
	Enzimática	Transporte	0,518± 0,16	0,617 ± 0,099	0,041
6	Enzimática	Estructural	0,524 ± 0,18	0,633 ± 0,121	0,006
	Enzimática	Transporte	0,524 ± 0,18	0,622 ± 0,110	0,04
5	Enzimática	Estructural	0,53 ± 0,17	0,636 ± 0,117	0,007
	Enzimática	Transporte	0,53 ± 0,17	0,627 ± 0,098	0,04
4	Enzimática	Estructural	0,536 ± 0,15	0,638 ± 0,130	0,01
	Enzimática	Transporte	0,536 ± 0,15	0,63 ± 0,089	0,042
3	Enzimática	Estructural	0,54 ± 0,15	0,63 ± 0,120	0,015
	Enzimática	Hormonal	0,54 ± 0,15	0,637 ± 0,94	0,043
	Enzimática	Transporte	0,54 ± 0,15	0,632 ± 0,098	0,048
2	Enzimática	Estructural	0,545 ± 0,13	0,633 ± 0,111	0,03
	Enzimática	Hormonal	0,545 ± 0,13	0,622 ± 0,86	0,034
1	Enzimática	Estructural	0,55± 0,19	0,626 ± 0,136	0,084
	Enzimática	Hormonal	0,55± 0,19	0,627 ± 0,090	0,023
0,2	Enzimática	Hormonal	0,578 ± 0,13	0,665 ± 0,099	0,017
	Hormonal	Movilidad	0,665 ± 0,099	0,556 ± 0,084	0,009
0	Defensiva	Hormonal	0,547 ± 0,07	0,63 ± 0,097	0,038
	Enzimática	Hormonal	0,551 ± 0,13	0,63 ± 0,097	0,021

Tabla 12: Tabla de resultados de función tras el test de Tukey para todas las q en grado 1 que dieron un p-valor en el ANOVA menor que 0,05. Solo los valores de q que den p-valor menor que 0,05 para el test de Tukey se muestran en esta tabla.

Grado 2					
Q	Función 1	Función 2	Media Función 1	Media Función 2	p-valor Tukey
10	Enzimática	Hormonal	0,517 ± 0,097	0,585 ± 0,085	0,041
	Enzimática	Transporte	0,517± 0,097	0,611 ± 0,090	0,009
9	Enzimática	Hormonal	0,521 ± 0,096	0,589 ± 0,083	0,039
	Enzimática	Transporte	0,521 ± 0,096	0,615 ± 0,097	0,009
8	Enzimática	Hormonal	0,525 ± 0,099	0,593 ± 0,083	0,037
	Enzimática	Transporte	0,525 ± 0,099	0,618 ± 0,092	0,009
7	Enzimática	Hormonal	0,529 ± 1,02	0,597 ± 0,120	0,034
	Enzimática	Transporte	0,529 ± 1,02	0,621± 0,102	0,01
6	Enzimática	Hormonal	0,533 ± 0,098	0,601 ± 0,119	0,031
	Enzimática	Transporte	0,533 ± 0,098	0,624± 0,093	0,01
5	Enzimática	Hormonal	0,537 ± 0,091	0,605 ± 0,104	0,027
	Enzimática	Transporte	0,537 ± 0,091	0,626± 0,097	0,012
4	Enzimática	Hormonal	0,54 ± 0,092	0,608 ± 0,105	0,023
	Enzimática	Transporte	0,54 ± 0,092	0,627± 0,089	0,014
3	Enzimática	Hormonal	0,543 ± 0,102	0,611 ± 0,116	0,019
	Enzimática	Transporte	0,543 ± 0,102	0,626± 0,96	0,019
2	Enzimática	Hormonal	0,546 ± 0,097	0,613 ± 0,112	0,015
	Enzimática	Transporte	0,546 ± 0,097	0,623± 0,090	0,027
1	Enzimática	Hormonal	0,55 ± 0,099	0,618 ± 0,106	0,01
	Enzimática	Transporte	0,55 ± 0,099	0,621± 0,096	0,044
	Hormonal	Movilidad	0,618 ± 0,106	0,548 ± 0,093	0,048
0,2	Enzimática	Hormonal	0,577 ± 0,098	0,655 ± 0,11	0,019
	Hormonal	Movilidad	0,655 ± 0,110	0,552 ± 0,082	0,005
0	Defensiva	Hormonal	0,55± 0,07	0,618 ± 0,084	0,037
	Enzimática	Hormonal	0,549 ± 0,097	0,618 ± 0,084	0,009
	Enzimática	Transporte	0,549 ± 0,097	0,62 ± 0,087	0,05

Tabla 13: Tabla de resultados de función tras el test de Tukey para todas las q en grado 2 que dieron un p-valor en el ANOVA menor que 0,05. Solo los valores de q que den p-valor menor que 0,05 para el test de Tukey se muestran en esta tabla.

Grado 3					
Q	Función 1	Función 2	Media Función 1	Media Función 2	p-valor Tukey
6	Enzimática	Hormonal	0,546 ± 0,105	0,606 ± 0,129	0,048
5	Enzimática	Hormonal	0,549 ± 0,098	0,609 ± 0,137	0,043
4	Enzimática	Hormonal	0,551 ± 0,101	0,611 ± 0,133	0,039
3	Enzimática	Hormonal	0,553 ± 0,099	0,612 ± 0,130	0,035
2	Enzimática	Hormonal	0,555± 0,106	0,612 ± 0,126	0,03
1	Enzimática	Hormonal	0,559 ± 0,099	0,616 ± 0,129	0,024
0,2	Hormonal	Movilidad	0,653 ± 0,090	0,554 ± 0,082	0,005
0	Enzimática	Hormonal	0,557 ± 0,096	0,615 ± 0,124	0,027

Tabla 14: Tabla de resultados de función tras el test de Tukey para todas las q en grado 1 que dieron un p-valor en el ANOVA menor que 0,05. Solo los valores de q que den p-valor menor que 0,05 para el test de Tukey se muestran en esta tabla. Para las q de 10 a 7 aunque el ANOVA diera un p-valor menor que 0,05 el test de Tukey no dio ningún p-valor menor que 0,05

ANEXO C. TABLA DE MF-DFA PARA CLASIFICACIÓN DE PLANTAS

Grado 1		Grado 2		Grado 3	
Q	p-valor	Q	p-valor	Q	p-valor
10	0,08	10	0,1	10	0,31
9	0,08	9	0,11	9	0,31
8	0,08	8	0,11	8	0,31
7	0,08	7	0,12	7	0,32
6	0,09	6	0,13	6	0,32
5	0,09	5	0,14	5	0,33
4	0,09	4	0,16	4	0,35
3	0,12	3	0,2	3	0,37
2	0,16	2	0,28	2	0,43
1	0,32	1	0,5	1	0,66
0,2	0,36	0,2	0,26	0,2	0,21
0	0,12	0	0,11	0	0,084
-0,2	0,4	-0,2	0,56	-0,2	0,06
-1	0,22	-1	0,67	-1	0,069
-2	0,27	-2	0,67	-2	0,069
-3	0,28	-3	0,67	-3	0,069
-4	0,29	-4	0,66	-4	0,068
-5	0,29	-5	0,66	-5	0,068
-6	0,29	-6	0,66	-6	0,067
-7	0,29	-7	0,065	-7	0,067
-8	0,3	-8	0,065	-8	0,067
-9	0,3	-9	0,065	-9	0,067
-10	0,3	-10	0,065	-10	0,067

Tabla 15: Representación de los p-valores tras el test T-Student de todas las q en los grados 1,2 y 3 para las secuencias de monodicotiledóneas y Dicotiledóneas.

ANEXO D. TABLAS DE CLUSTERING PARA MF-DFA CON BACTERIAS Y REINOS

Grado	Reino	Bacterias
1	0,04	0,03
	Animales / Resto	Gamm / Resto
	0,006	0,008
	Plantas / Resto	Bacilli / Resto
	0,006	0,12
	Bacteria / Resto	-
	0,006	-
2	Reino	Bacterias
	0,04	0,003
	Animales / Resto	Gamma / Resto
	0,009	0,02
	Plantas / Resto	Bacilli / Resto
	0,009	0,18
	Bacteria / Resto	-
	0,009	-
3	Reino	Bacterias
	0,04	0,0004
	Animales / Resto	Gamma / Resto
	0,01	0,001
	Plantas / Resto	Bacilli / Resto
	0,01	0,12
	Bacteria / Resto	-
	0,01	-

Tabla 16: Representación del índice de Rand del método k-means para cada grado de los $h(q)$ del MF-DFA cada Grado del polinomio del MF-DFA . Este método de *clustering* se aplica sobre todos los reinos a la vez, y luego tratando de diferenciar, animales, plantas y bacterias del resto, también se trata de agrupar todas las bacterias y después agrupar bacilli respecto del resto de bacterias y gammaproteobacteria del resto.

Grado	Reino	Bacterias
$1 + \alpha$	0,04	0,05
	Animales / Resto	Gamma / Resto
	0,005	0,018
	Plantas / Resto	Bacilli / Resto
	0,005	0,08
	Bacteria /Resto	-
	0,005	-
$2 + \alpha$	Reino	Bacterias
	0,04	0,06
	Animales / Resto	Gamma / Resto
	0,009	0,025
	Plantas / Resto	Bacilli / Resto
	0,009	0,16
	Bacteria /Resto	-
	0,009	-
$3 + \alpha$	Reino	Bacterias
	0,031	0,001
	Animasles / Resto	Gamma / Resto
	0,006	0,001
	Plantas / Resto	Bacilli / Resto
	0,006	0,127
	Bacteria /Resto	-
	0,006	-

Tabla 17: Representación del índice de Rand del método k-means para cada grado de los $h(q) + \alpha$ del MF-DFA cada Grado del polinomio del MF-DFA . Este método de *clustering* se aplica sobre todos los reinos a la vez, y luego tratando de diferenciar, animales, plantas y bacterias del resto, también se trata de agrupar todas las bacterias y después agrupar bacilli respecto del resto de bacterias y gammaproteobacteria del resto.

ANEXO E. TABLAS DE CLASIFICIACIÓN MEDIANTE REDES NEURONALES PARA MF-DFA CON BACTERIAS Y REINOS

Grado	Reino	Bacterias
1	0,85	0,72
	Animales / Resto	Gamma / Resto
	0,28	0,45
	Plantas / Resto	Bacilli / Resto
	0,35	0,36
	Bacteria / Resto	-
	0,31	-
2	Reino	Bacterias
	0,9	0,82
	Animales / Resto	Gamma / Resto
	0,3	0,45
	Plantas / Resto	Bacilli / Resto
	0,28	0,18
	Bacteria / Resto	-
	0,28	-
3	Reino	Bacterias
	0,85	0,72
	Animales / Resto	Gamma / Resto
	0,3	0,54
	Plantas / Resto	Bacilli / Resto
	0,28	0,48
	Bacteria / Resto	-
	0,22	-

Tabla 18: Representación de la tasa de error de la red neuronal para cada grado de los $h(q)$ del MF-DFA cada Grado del polinomio del MF-DFA . Este método se aplica sobre todos los reinos a la vez, y luego tratando de diferenciar, animales, plantas y bacterias del resto, también se trata de clasificar todas las bacterias y después clasificar bacilli respecto del resto de bacterias y gammaproteobacteria del resto.

Grado	Reino	Bacterias
$1 + \alpha$	0,75	0,68
	Animales / Resto	Gamma / Resto
	0,41	0,18
	Plantas / Resto	Bacilli / Resto
	0,42	0,36
	Bacteria /Resto	-
	0,33	-
$2 + \alpha$	Reino	Bacterias
	0,85	0,7
	Animales / Resto	Gamma / Resto
	0,3	0,45
	Plantas / Resto	Bacilli / Resto
	0,28	0,45
	Bacteria /Resto	-
	0,22	-
$3 + \alpha$	Reino	Bacterias
	0,82	0,54
	Animales / Resto	Gamma / Resto
	0,26	0,54
	Plantas / Resto	Bacilli / Resto
	0,37	0,36
	Bacteria /Resto	-
	0,28	-

Tabla 19: Representación de la tasa de error de la red neuronal para cada grado de los $h(q) + \alpha$ del MF-DFA cada Grado del polinomio del MF-DFA . Este método se aplica sobre todos los reinos a la vez, y luego tratando de diferenciar, animales, plantas y bacterias del resto, también se trata de clasificar todas las bacterias y después clasificar bacilli respecto del resto de bacterias y gammaproteobacteria del resto.