POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA,ALIMENTARIA Y DE BIOSISTEMAS

GRADO EN BIOTECNOLOGÍA.

DEPARTAMENTO DE MATEMÁTICA APLICADA.

Análisis y caracterización de secuencias de nucleótidos como series temporales.

TRABAJO FIN DE GRADO

Autor: **Manuel Sánchez Díaz**

Tutor: **Carlos García-Gutiérrez Báez y Fernado San José Martínez.**

8 JULIO 2021

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TECNICA SUPERIOR DE INGENIERIA AGRONOMICA, ALIMENTARIA Y DE BIOSISTEMAS

GRADO DE BIOTECNOLOGÍA

ANALISIS Y CARACTERIZACIÓN DE SECUENCIAS DE NUCLEOTIDOS COMO SERIES TEMPORALES

TRABAJO FIN DE GRADO

Manuel Sánchez Díaz

MADRID, JULIO 2021.

Directores:

Carlos García-Gutiérrez Báez y Fernando San José Martínez

Dpto. de Matemática Aplicada, UPM

E.T.S. DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

ANÁLISIS Y CARACTERIZACIÓN DE SECUENCIAS DE NUCLEOTIDOS COMO SERIES TEMPORALES.

Memoria presentada por Manuel Sánchez Díaz para la obtención del título de Graduado en Biotecnología por la Universidad Politécnica de Madrid

Fdo: Manuel Sánchez Díaz

VºBº Tutor

Carlos García-Gutiérrez Báez

Dpto. de Matemática Aplicada

ETSIAAB - Universidad Politécnica de Madrid

Firmado por GARCIA-GUTIERREZ BAEZ CARLOS - 01937581S el día 07/07/2021 con un certificado emitido

VºBº Cotutor

Fernando San José Martínez

Dpto. de Matemática Aplicada

ETSIAAB - Universidad Politécnica de Madrid

Firmado por SAN JOSE MARTINEZ FERNANDO - DNI 00791279X el día 07/07/2021 con un certificado emitido por AC Administración Pública

Madrid, 8 de julio de 2021

Para ti abuelo, todo lo bueno que me pasa es gracias a ti.
Gracias por seguir enseñándome a vivir día a día.
Todo lo importante que hice, hago y haré te lo dedicaré y esto no podía ser menos.
Sé que estas muy orgulloso de mi

AGRADECIMIENTOS.

Primero de todo, quiero agradecerle este trabajo al Manu de 17 años. Te escribo 5 años después, desde tu cama para decirte que ha sido la mejor decisión que podrías haber tomado. Gracias por darte esa nueva oportunidad y mostrarte tal como eres, has conseguido que mucha gente a tu alrededor con la que puedes ser tú, con la que no tienes que esconder tus rarezas y con la que eres infinitamente feliz y eso es mejor que cualquier logro vacío.

También quería agradecer a Carlos y a Fernando, porque si algo admiro de las personas es su capacidad para enseñar y ellos lo han hecho conmigo de sobra. Es impagable todo el tiempo que habéis invertido en enseñarme. Gracias por aguantarme este año tan difícil. No podría haber tenido unos tutores mejores. Me he sentido como en casa investigando en el departamento con vosotros.

Como se ha caracterizado mi vida academica siempre lo he dejado todo para el final (espero que esto no lo lea el tribunal ajajajaj), y bueno pues eso que son las 11:34 estoy sin dormir y tengo que entregarlo ya asique como me ha dicho Carlos voy a intentar poner todo mi corazoón y agradecer este trabajo a esas personas de las que me acuerde en estos cinco minutos que son las mas importantes en mi vida. Se aceptan quejas y si alguien quiere que le escriba algo mas bonito que me lo diga

A mis padres, mi ejemplo a seguir en la vida. No podría valorar todo el sacrificio que habéis hecho para que yo pueda estudiar. Gracias por todos los valores que me habéis enseñado y que me hacen ser quien soy hoy en día. Sois unos luchadores, aunque no me vaya de casa todavía, espero haceros sentir bastante orgullosos. Os quiero a los dos muchísimo.

A mi familia , en especial, a mi abuela por iluminarme los días desde tu sillón hasta cuando me llamas pesado me haces el hombre mas feliz del mundo, a mi hermana, por aguantarme como nadie taaaaaantos años, a mi tío Nacho, por haberme enseñado lo bonito que es aprender y conocer en especial a la ciencia , a mi madrina por la conexión tan especial que tenemos y su ayuda incondicional .

A mi familia de la universidad, gracias por haberme dejado ser yo mismo, han sido los mejores años de mi vida. A Perez y a Moi, al escudo le quedan muchas aventuras juntos, a Elia tu hermanito quiere que vayas con él muchísimas más veces a cualquier otra parte, a Lucía, gracias por esa complicidad nuestra que nos hace tan especial ya me has enseñado siempre muchas cosas pero ultimamente mas te quiero con locura , a Julian mi más fiel compañero de aventuras en la ETSIAAB, a Carmen por ayudarme en tantas ocasiones , a Esther por estar ahi desde el principio y siempre, a Male por ser mi mami y cuidarme tanto, a Alo por sus locuras y nuestras conversaciones profundas .

A mis amigos del barrio, a homeopatia, por que no hay nada más bonito que tener que oirme cantar y que aun así sigan queriendo estar conmigo, a Mariam y a Helena, por tantos años que llevamos juntos, a Susana, por ser la persona mas valiente, luchadora, a la que mas admiro y la que mas me parece una tia de puta madre.

También le quería agradecer a la cerveza por ponerle sabor a estos cinco años de locuras, te amo con locura. También gracias a Asocis la fiesta en donde más feliz soy, espero que vuelvas mi niña.

Queda un poco cutre terminar así pero cuento con que esto no lo lea mucha gente, quería agradecer también a Eurovision por ponerle la banda sonora a la redacción de este trabajo durante tantos días y noches.

Y bueno ya se han acabao los cinco minutos que le puedo dedicar a esto, en resumen, os quiero a todos los que hayais tenido el honor de leer esto.

Índice

Índice de figuras

Índice de cuadros

Lista de abreviaturas

AN: Ácido Nucleico.

AR: (*Autoregressive*). Autorregresivo.

ARI: (*Adjusted Rand Index*). Índice de Rand Ajustado.

ARIMA: (*Autoregressive Integrated and Moving Average*). Autorregresivo Integrado y de Medias Móviles.

ARMA: (*Autoregressive and Moving Average*). Autorregresivo y de Medias Móviles.

ARN: Ácido Ribonucleico.

ADN: Ácido Desoxirribonucleico.

cirARN: ARN circular.

DFA: (*Detrended Fluctuation Analysis*). Análisis de Fluctuaciones sin Tendencia.

ECDP: *European Centre for Disease Prevention and Control.*

K-NN: (*K- Nearest Neighbors*). K-Vecinos más próximos.

log: logaritmo.

MA: (*Moving Average*). Medias móviles.

mill: millones.

MPL: (*Mean Path Length*). Distancia media más corta.

mARN: ARN mensajero.

NCBI: *National Center for Biotechnology Information.*

pb: pares de bases.

NN: (*Neural Network*). Red Neuronal.

e.e.m: Error estándar de la media.

VG: (*Visibility Graph*). Grafo de Visibilidad.

Resumen

The analysis of time series has resulted in being particularly useful in several areas of knowledge and related studies (such as economic forecasting, budgetary analysis, yield processing), for which a significant number of tools have been developed. Because of the advance of sequencing techniques, there are databases with millions of sequences of nucleotides with hardly any characterisation. It exits an algorithm which can convert nucleotide sequences into time series. The present piece of research frames a study of time series and transformed networks from nucleotides sequences. We also investigate differents methods, based on machine learning and deep learning, which could characterize the sequences through several mathematical parameters.

There are many parameters of time series that we are going to study in this research. Firstly, we studied the correlation of time series, using the Detrented Fluctuation Analysis method. Then we parameterize the time series to describe some processes, such as autoregressive processes, processes of moving averages and integrated approaches. Finally, these time series were transformed into networks in the third place to study and quantify other associated properties: networks connectivity and the small world effect.

We examined the resulting parameters obtaines from the DNA sequences and also their applicability as classifiers for several groupings: content in introns; the realm the organism; the type of Nucleic Acid (DNA or RNA); functionality (codifying or structural), and genetic compartment where the sequence is located.

Furthermore, we implemented different classification methods based on Machine Learning and Deep Learning. These two methods will classify according to the parameters of time series and networks present in this piece of research. First, we grouped the sequences by an unsupervised method, k-means. Then, the sequences were grouped by a monitored method, K-Nearest Neighbors. Per cent success rates increased between 50 % and 70 % in relation to the previous process. Finally, an Artificial Neural Network was trained with the before-mentioned parameters. This network could successfully characterise up to 95 % of the sequences for certain features.

Prospects resulting from this study aimed to increase the sequences to enhance the confidence level when classifying. It is also focused on studying other properties of the networks and time series, such as the fractal properties. Moreover, the number of features to be predicted with these types of parameters can also be increased.

CAPÍTULO 1. INTRODUCCIÓN.

La vida está escrita con palabras de 4 letras. Estas palabras son las secuencias de ADN de las células y las letras son los nucleótidos que lo forman: adenina (A), tímina (T), guanina (G) y citosina (C). La combinación de estas cuatro letras es un formato único y aparentemente básico, pero contiene la información para que un amplio rango de fenómenos biológicos ocurran.

La secuenciación de ADN estudia la composición del ADN y la información que contiene esta composición. Las primeras tecnologías de secuenciación fueron desarrolladas por Sanger et al. [1] en 1977. En los últimos años la llegada de las nuevas tecnologías de secuenciación (*Next generation DNA sequencing*) ha acelerado y abaratado la capacidad para secuenciar esta molécula [2]. Este crecimiento ha permitido a que haya bases de datos genéticos como Gene del National Center for Biotechnology Information (NCBI) [3] que contiene más de 30 mill. de secuencias, de las cuales muchas se encuentran sin caracterizar. En los próximos años, nuevos dispositivos de secuenciación aparecerán produciendo un aumento vertiginoso de la información genética de la que dispondremos, lo que requerirá el desarrollo de nuevos métodos y herramientas capaces de analizar esta gran cantidad de información [4]. Este es uno de los grandes retos de la bioinformática.

El gran progreso de la biología molecular en los últimos años nos ha permitido entender gran parte de la estructura y funciones de los genomas. La biología molecular, tradicionalmente, se ha centrado en el estudio de las partes del genoma conocidas como 'codificantes' que son las que intervienen en la producción de proteínas. Pero esta parte del genoma cubre menos del 10 % del genoma de muchos eucariotas. La parte restante del genoma, conocida como 'no codificante', ha sido considerada durante mucho tiempo un estorbo sin ningún significado biológico. El descubrimiento de ciertas funciones regulatorias en estas partes del genoma han cambiado este punto de vista [5]. Esto unido a los rápidos avances en la secuenciación de secuencias nucleotídicas han llevado a muchos grupos a llevar a cabo investigaciones sobre la distribución de los cuatro nucleótidos en genomas completos o grandes fragmentos que contengan estas partes 'no codificantes', en busca de códigos y reglas ocultas [6]. Entendemos códigos ocultos como partes del genoma o subsecuencias que tienen un papel importante en la función del ADN y que no sea la de codificación de proteínas. Para el estudio de las distribuciones y la presencia de códigos ocultos se requieren análisis matemáticos.

Una herramienta matemática ampliamente conocida y utilizada son las series temporales. Están ampliamente estudiadas y su uso se extiende a múltiples ramas como la economía donde se utiliza por ejemplo para predecir comportamientos a partir de valores pasados [7]. Aparentemente la idea de serie temporal unida a secuencia de ADN parece inconexa, pero esto cambia con el novedoso trabajo de Peng et al. [8] donde utiliza herramientas matemáticas para explicar acontecimientos biológicos. En dicho trabajo presentan un algoritmo para crear una serie temporal a partir de una secuencia de nucleótidos y estudiar propiedades de estas series temporales. Nosotros utilizaremos en nuestra investigación herramientas provenientes de este trabajo.

Se describió en este trabajo y en trabajos posteriores [9] que las regiones 'no codificantes' presentaban correlaciones de largo alcance, mientras que las secuencias 'codificantes' mantienen correlaciones de corto alcance, en otras palabras estas regiones se comportaban igual que secuencias formadas al azar. Este trabajo pretende seguir con esta línea de investigación pero aumentando el área de análisis: i) el estudio de otros parámetros matemáticos relacionados con las series temporales y ii) ampliar la investigación de estos parámetros matemáticos en otras características del ADN.

En este trabajo se van a llevar a cabo varios estudios matemáticos de las series temporales asociadas a una secuencia de nucleótidos. A parte del cálculo de la correlación, vamos a calcular unos parámetros que aproximan las series temporales a unos procesos conocidos[7]. También convertimos nuestras series temporales en grafos gracias a un algoritmo de transformación conocido como Grafo de Visibilidad [10]. Un grafo esta formado por una serie de nodos o puntos, y conexiones entre esos nodos, que se denominan vértices. Cuantificaremos las siguientes propiedades del grafo con el fin de caracterizar las secuencias: grado de conectividad del grafo y propiedad del mundo pequeño [10].

Por tanto, analizamos la viabilidad de ampliar el uso de otros parámetros, a parte del presentado en el trabajo de Peng [8], para caracterizar nuestras secuencias de nucleótidos.

Estos nuevos parámetros no solo los vamos a cuantificar en series temporales asociadas a secuencias de regiones 'codificantes' o 'no codificantes', vamos a ampliar el espectro de características. El porcentaje de regiones 'no codificantes' ha ido variando con la evolución, tiene sentido que

el estudio de estos parámetros en secuencias de organismos mas evolucionados como las eucariotas sea distinto al de secuencias de organismos más arcaicos como los procariotas. También vamos a ver la diferencia de estos parámetros en secuencias con funciones estructurales como los ribosomas o secuencias genómicas.

Otra característica que tendremos en cuenta en nuestro trabajo es la localización de la secuencia de ADN. Según la teoría endosimbiótica, orgánulos de las células eucariotas como la mitocondria y el cloroplasto contienen material genético porque provienen de organismos procariotas antiguos [11].

Es por esto que nos parece interesante el estudio de los parámetros en diferentes localizaciones. Por tanto consideramos los siguientes grupos de clasificación del trabajo: contenido en intrones, reino al que pertenece la secuencia, tipo de Ácido Nucleico, codificación del Ácido Nucleico y la localización.

Si estos parámetros realmente pueden describir características del ADN, al final del trabajo vamos a buscar la implementación de varios métodos de clustering que puedan caracterizar las secuencias mediante los parámetros asociados a sus series temporales. Las series temporales podrían ayudar a la caracterización de la abundante información generada mediante la secuenciación.

Los objetivos de este trabajo son :

1. Implementación de algoritmos matemáticos para el análisis de series temporales y grafos. Validación de los programas mediante la comparación con los resultados de los artículos en los que las herramientas son presentadas [8].

2. Conseguir una base de datos de secuencias suficientemente amplia.

3. Caracterización de secuencias genéticas mediante herramientas matemáticas ya programadas y estudiadas.

4. Evaluar la posibilidad de usar los diferentes parámetros de la investigación como clasificadores de secuencias.

5. Implementación de diferentes métodos de clasificación basados en el aprendizaje automático y aprendizaje profundo. Estos métodos clasificaran en función de los parámetros de series temporales y grafos presentados en el trabajo. Observar la bondad del ajuste entre clasificación esperada de estos métodos y la clasificación real.

CAPÍTULO 2. MATERIALES Y MÉTODOS.

2.1. Datos de las secuencias

Para este trabajo se utilizaron 316 secuencias de nucleótidos, obtenidas de las bases de datos de Gene [3] y Nucleotide [12] del National Center for Biotechnology Information (NCBI)[13]. Para su clasificación hemos seguido el criterio utilizado por la propia base de datos.

Estas 316 secuencias son de diferentes organismos, esparcidos a lo largo del espectro filogenético, por ejemplo: *Escherichia Coli*, *Homo Sapiens*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Mus musculus*, *Danio Rerio*, etc...

La elección de las secuencias no se pudo realizar de forma aleatoria. Esta búsqueda se realizó para que todas las características estuvieran representadas mediante un número mínimo de secuencias. Debido a la baja presencia de algunos grupos en las bases de datos sería difícil hacer una buena selección de secuencias de forma aleatoria. Además hay que tener en cuenta que las secuencias de cualquier base de datos están ya sesgadas por los intereses de secuenciación. Todo esto hace que nuestro trabajo tenga un limite exploratorio. Para una mayor confianza estadística sería necesario tomar una muestra más amplia.

Las secuencias escogidas para realizar este trabajo presentan una gran varianza en longitud de secuencia. La más pequeña, que es la secuencia de rARN 5S de *Chlorobium phaeobacteroides* contiene 109 pb (pares de bases), la más grande es el genoma completo del fago λ de *E.coli* contiene 97050 pb; el 75 % tienen una longitud de secuencia entre 109 pb y 11260 pb.

En la Figura 1 podemos ver un histograma con la distribución de las longitudes de las secuencias.

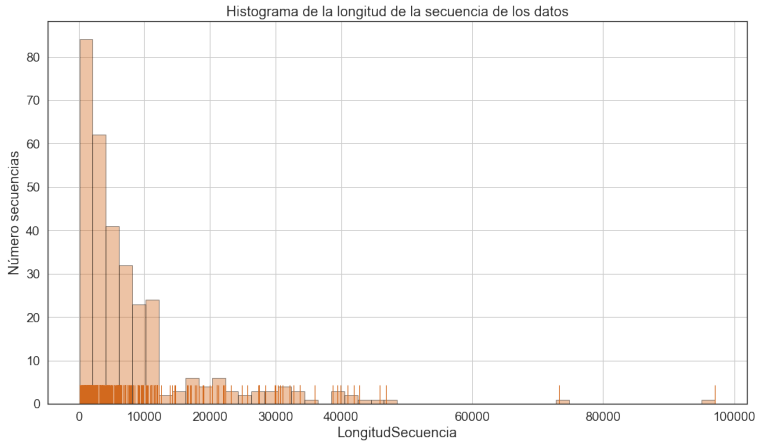


Figura 1: Histograma de la longitud de las secuencias. Se puede observar que la gran mayoría de los datos se encuentran entre 0 y 10000 pb

Se agruparon estas secuencias en diferentes grupos de clasificación según estaban caracterizadas en la base de datos del NCBI:

- **Contenido de Intrones.**
  Los intrones son regiones del ADN que son transcritos a ARN, pero luego son eliminados del transcrito maduro antes de la traducción [14]. En este trabajo hemos estudiado secuencias que contenían intrones y secuencias que no los contenían.

- **Reino al que pertenece el organismo de la secuencia estudiada.**
  Según la clasificación de Robert Whittaker [15], una de las más aceptadas, los seres vivos se dividen en 5 reinos. El reino Animal, reino Vegetal, reino Fungi, reino Protista y el reino Mónera.

  Para la realización de este trabajo hemos estudiado secuencias de los 5 reinos. En el reino Mónera se encuentran tanto Bacterias como Archaeas, pero nuestro trabajo consta solo de secuencias de Bacterias para el grupo Mónera. Además, aunque generalmente no sean considerados seres vivos hemos trabajado con secuencias de nucleótidos de Virus.

- **El tipo de Ácido Nucleico.**
  Hemos clasificado las secuencias según si eran de ADN o si eran de ARN. También las hemos clasificado según su topología, si eran secuencias lineales o si eran circulares. Los ARN circulares (cirARNs) son secuencias circulares de ARN que provienen del pre-mARN y que sufren un splicing alternativo y unen mediante un enlace covalente sus extremos[16]. No se ha podido analizar un grupo de cirARNs, ya que no se han podido encontrar secuencias de este tipo en la base de datos de NCBI.

- **Secuencias codificantes o estructurales.**
  Algunas secuencias contienen información genética para sintetizar proteínas. En este estudio a esas secuencias, y a los mensajeros que provienen de secuencias de este tipo, se les denominan codificantes ('protein coding'). El resto de secuencias que no pertenecen a este grupo son secuencias estructurales, por ejemplo, ARNs ribosómicos.

- **Localización del Ácido Nucleico.**
  Hay numerosos compartimentos genéticos donde se pueden localizar los Ácidos Nucleicos en los diferentes tipos de organismos. En este trabajo las diferentes localizaciones que vamos a estudiar son el Citoplasma, el Núcleo, el Cloroplasto, la Mitocondria, la Cápsida y los Plásmidos. El genoma central de las bacterias (nucleoide) [17] no se encuentran en un orgánulo membranoso como si ocurre en eucariotas (núcleo). Las secuencias pertenecientes a los genomas centrales tanto de eucariotas como de procariotas se encuentran agrupados en el grupo de Núcleo.

En el Cuadro 1 podemos ver los diferentes grupos para cada característica, el número de secuencias que contiene cada grupo y su longitud media.

Característica	Grupo	N	longitud media (pb)
Intrones	Sin	209	5144
	Con	107	14797
Reino	Animal	160	11207
	Mónera	37	2402
	Fungi	27	3382
	Vegetal	50	4264
	Protista	14	1984
	Virus	28	7932
Tipo AN	ADN lineal	163	9737
	ARN lineal	133	5506
	ADN circular	20	5728
Codificación del AN	Codificante	279	8378
	Estructural	37	2461
Localización del AN	Núcleo	121	10724
	Plásmido	14	2905
	Mitocondria	20	4002
	Cloroplasto	18	2283
	Citoplasma	115	6553
	Cápsida	28	7932

Cuadro 1: Resumen de los grupos de estudio de cada característica.

2.2. Algoritmo de transformación de secuencias genéticas a series temporales

El objeto de estudio de este trabajo son las secuencias de nucleótidos, vamos a estudiarlas, transformándolas en series temporales.

Una serie temporal es el resultado de observar los valores de una variable a lo largo del tiempo. Se suelen tratar en intervalos regulares (por ejemplo un día, un año...) para facilitar su tratamiento, asignando los valores a instantes temporales concretos.

Una serie temporal puede estar formada por la medición de una o varias variables. Sus valores son irrepetibles debido a su carácter histórico. El estudio de series temporales nos puede

ayudar a interpretar eventos ya ocurridos o pronosticar su comportamiento en el futuro.

Para llevar transformar las secuencias de nucleótidos en series temporales vamos a utilizar un algoritmo de transformación [8].

Estas series se conocen como *DNA walks* o caminos de ADN y son una forma de representar la secuencia de forma gráfica. Este camino esta definido por una regla de mapeo binaria. El caminante de este camino, empieza en 0 y sube un paso hacia arriba si al avanzar un paso por la cadena de ADN se encuentra con una base pirimidínica y un paso hacia abajo si se encuentra con una base púrica.

Este camino está formado por todos los desplazamientos netos, $y(l)$, de todos los pasos l del camino. El desplazamiento se calcula con la Fórmula 1.

$$y(l) = \sum_{i=1}^{l} u(i) \tag{1}$$

El desplazamiento neto $y(l)$ se calcula como el sumatorio de todos los pasos $u(i)$ entre $i = 0$ hasta $i = l$. Los pasos $u(i)$ tendrán valor $+1$ si en la posición i de la cadena de ADN se encuentra una base pirimidínica (Tímina (T), Citosina (C) o Uracilo (U)) y tendrá un valor -1 si en la posición i de la cadena de ADN se encuentra una base púrica (Adenina (A) o Guanina (G)).

Podemos ver una representación de los desplazamientos netos, $y(l)$, para todos los pasos l en la Figura 2. Es la serie temporal transformada a partir de la secuencia de ADN del Adenovirus tipo 2.

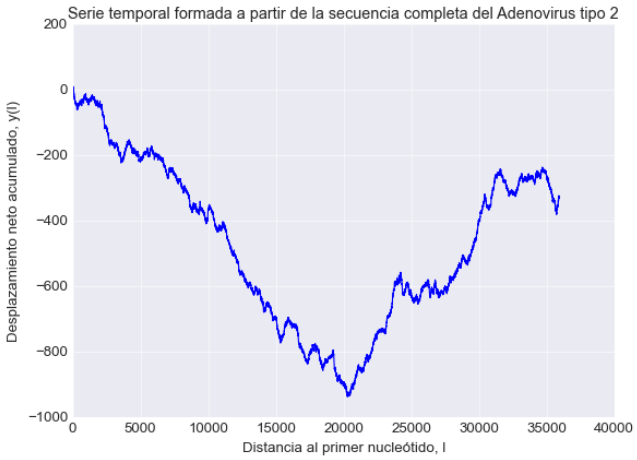


Figura 2: Serie temporal transformada con el algoritmo descrito previamente a partir de la secuencia completa del Adenovirus tipo 2.

2.3. Herramientas matemáticas para el estudio de las series temporales

2.3.1. Herramientas clásicas

Las series temporales pueden ser estudiadas mediante herramientas clásicas que las aproximaban a diferentes procesos. Estos procesos son útiles para observar la dependencia de los valores de una serie temporal de su pasado. Hemos analizado 4 tipos de procesos [7]:

1. **Proceso autorregresivos, AR(p).**
   En un proceso AR(p), los sucesivos valores de una serie temporal dependen linealmente de una cantidad finita de los valores anteriores. Esta cantidad será el parámetro p del proceso. Un valor del proceso z_t es función de los p valores anteriores $f(z_{t-1}, z_{t-2}, ..., z_{t-p})$, pero también depende de las innovaciones a_t que son la nueva información que se añade al proceso en cada instante. La innovación a_t es una variable aleatoria que sigue una distribución normal de media cero y varianza constante. Por ejemplo la formula que describe un proceso autorregresivo de primer orden, AR(1) sería: $z_t = c + \phi z_{t-1} + a_t$ donde c y $-1 < \phi < 1$ son constantes a determinar, a_t es la innovación y z_{t-1} es el valor anterior al valor que estamos observando.

Los procesos que se describen como autorregresivos tienden a tener una memoria más larga que otro tipo de procesos de series temporales [7]. Entendiendo memoria larga con que el valor actual esta relacionado con más valores de su pasado que en otro tipo de procesos de series temporales, como los de medias móviles, MA(q).

2. **Proceso de medias móviles, MA(q).**
   Este es un proceso que a diferencia de los procesos autorregresivos el valor actual z_t ya no depende del valor anterior o de una cantidad finita de valores anteriores, sino que depende de una cantidad finita de innovaciones pasadas. Un proceso de primer orden de este estilo dice que $\tilde{z_t} = a_t - \theta \cdot a_{t-1}$ donde $\tilde{z_t} = z_t - \mu$, siendo μ la media del proceso; a_t son las innovaciones que siguen un proceso de ruido blanco con varianza constante y θ es la dependencia del valor actual de la serie de la última innovación ocurrida. Este proceso lo podemos generalizar en un proceso cuyo valor no solo dependa de la última innovación, sino de las q ultimas, obteniendo un proceso medias móviles de orden q, MA(q). En contraste con los procesos autorregresivos su memoria es más corta, es decir el valor actual depende de menos observaciones pasadas.

3. **Proceso autorregresivo de medias móviles, ARMA(p,q).**
   Es probable que la serie temporal tenga propiedades de memoria tanto a largo plazo como a corto plazo, por lo que nos interesa estudiar un modelo que estudie ambos tipos de memoria. Este es el modelo ARMA(p,q) que mezcla los procesos autorregresivos con los de medias móviles.

4. **Proceso autorregresivo integrado de medias móviles, ARIMA(p,d,q).**
   Hemos estado asumiendo que todos los procesos anteriores son estacionarios. De hecho la mayoria de series no son estacionarias, su media va variando a lo largo del tiempo, por eso tenemos que estudiar estas series con unos procesos conocidos como integrados. El termino d del modelo ARIMA denota el número de veces que la serie se diferencia de un proceso estacionario, ARMA(p, q).

A la hora de calcular estos parámetros se han buscado los parámetros que maximizan los valores de AIC, AICc y BIC del modelo [18]. Estos criterios son diferentes medidores de la calidad relativa de un modelo estadístico para el conjunto de datos que tenemos.

2.3.2. Herramientas avanzadas

Vamos a realizar una medida de la correlación matemática de las series temporales obtenidas a partir de las secuencias de nucleótidos. Realizaremos el estudio de la fluctuación ($F(l)$) con el método Análisis de Fluctuaciones sin Tendencia, DFA, por su expresión en inglés *Detrended Fluctuation Analysis* [8].

La fluctuación F para un valor de distancia l se puede calcular como la diferencia entre la media del valor cuadrado del desplazamiento y el cuadrado del valor medio del desplazamiento después de l pasos,

$$F^2(l) = \overline{[\triangle\, y(l)]^2} - \overline{[\triangle\, y(l)]}^2 \tag{2}$$

los desplazamientos $\triangle\, y(l)$ son la diferencia entre dos valores de desplazamiento neto separado por una distancia l,

$$\triangle\, y(l) = y(l_0 + l) - y(l_0) \tag{3}$$

Las barras en la fórmula 2 nos indica la media para una distancia l fija, moviendo todas las posiciones de la serie temporal l_0 posibles, que se encontraran en el intervalo $l_0 \in [1, N - l]$. El valor de N es la longitud de la secuencia, que será igual al número de valores de la serie temporal.

Podemos aproximar la función $F(l)$ a una función potencial de la siguiente manera,

$$F(l) \sim l^{\alpha} \tag{4}$$

El calculo de α nos puede describir principalmente dos tipos de comportamiento en relación a la correlación de la serie temporal:

- El caso de $\alpha = \frac{1}{2}$, esto nos indica que la secuencia de nucleótidos es completamente aleatoria. No existe correlación.
- El caso de $\alpha > \frac{1}{2}$, esto nos indica que la secuencia de nucleótidos presenta una correlación de largo alcance.

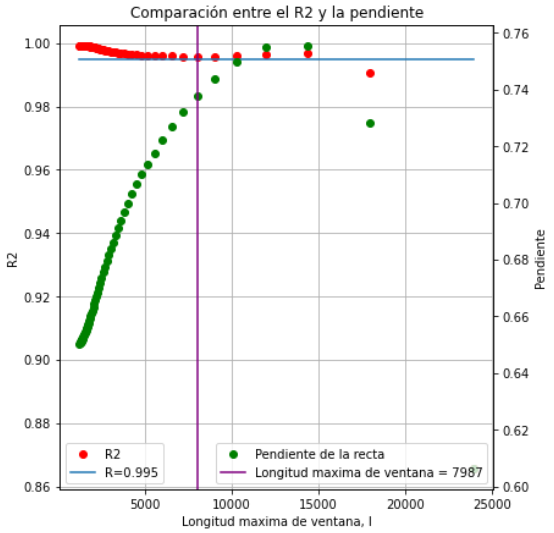
Para poder calcular este valor α de cada serie temporal hemos hecho un programa que calcula los valores de $F(l)$ a partir de los valores de desplazamiento neto $y(l)$ aplicando la fórmula 2.

Ajustamos los valores logarítmicos de $F(l)$ y l a una recta de regresión, donde la pendiente (α) de la recta será el parámetro que estemos buscando. A partir de aquí nos referiremos a este parámetro como Peng-α.

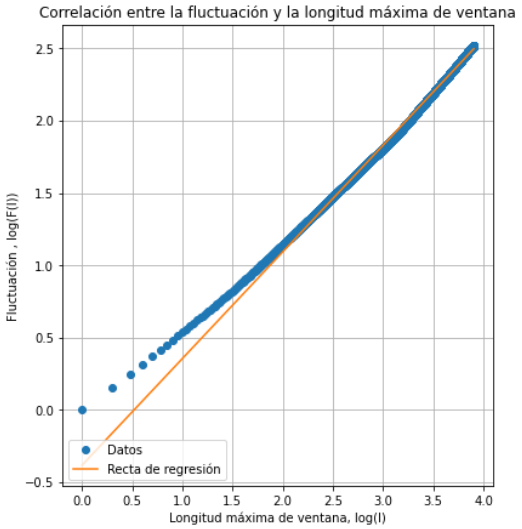
Este parámetro sirve para calcular una medida cuantitativa de la correlación para grandes distancias a lo largo de la cadena del ADN, dependiendo cual de los dos comportamientos previamente descritos presente.

Se ha visto que para tamaños de ventana muy amplios ($l > 1000$) el error estadístico aumenta y el R^2 disminuye [8]. Este efecto lo ponemos ver en la Figura 3. El tamaño máximo de l que vamos a utilizar en el trabajo es el que nos asegure un ajuste con un valor mínimo de $R^2 = 0{,}995$.

Una vez seleccionado el valor máximo de l, calcularemos nuestro parámetro Peng-α como podemos observar en la Figura 3b.



(a) Comparación entre el ajuste de la recta, R^2, la pendiente, α y el tamaño máximo de ventana, l.



(b) Recta de regresión entre los datos de fluctuación, F(l) y el tamaño de ventana, l. Estimación de Peng-α.

Figura 3: Ejemplo del calculo de la fluctuación en la serie temporal de la secuencia de la Figura 2.En la Figura 3a vemos el tamaño máximo de ventana que podemos usar para el cálculo de la fluctuación que nos asegura un ajuste en la Figura 3a con un R^2 determinado. La Figura 3b nos muestra el ajuste entre el logaritmo del tamaño de ventana ($log(l)$) y el logaritmo de la fluctuación ($log(F(l))$) con un $R^2 = 0{,}995$, la pendiente de esta recta es el parámetro Peng-α.

2.4. Algoritmo de transformación de series temporales a grafos. Grafo de Visibilidad.

Podemos ver las series temporales con un grafo y estudiar sus propiedades. Para poder transformar nuestra serie temporal en un grafo, vamos a implementar algoritmo de transformación conocido como Grafo de Visibilidad, VG, por su expresión en inglés *Visibility Graph* [10].

Cada valor de la serie temporal asociada a la secuencia de nucleótidos, presentará un nodo asociado, por lo que, cada grafo tendrá el mismo numero de nodos que número de nucleótidos tenga la secuencia. Es decir el grafo tendrá N nodos. Todos los grafos de este trabajo van a tener una serie de propiedades en común: no son direccionados (no hay dirección en las conexiones), los nodos (l_i) al menos van a estar conectados con los nodos inmediatamente anterior (l_{i-1}) y posterior

(l_{i+1}), excepto si es el primero que solo estará conectado con el segundo; y si es el último que solo estará conectado con el penúltimo.

El resto de conexiones estarán regidos por el criterio de visibilidad. Dos nodos están conectados si podemos trazar una linea recta entre ellos sin que esta linea sea cortada. Esto lo podemos ver gráficamente en la Figura 4.

Matemáticamente hablando dos nodos $a(l_a, y_a), b(l_b, y_b)$ cumplen el criterio de visibilidad y por tanto están conectados, si todos los puntos $c(l_c, y_c)$ que se encuentren entre ellos $(l_a < l_c < l_b)$ cumplen la fórmula 5.

$$y_c < y_b + (y_a - y_b) \cdot \frac{l_b - l_c}{l_b - l_a} \tag{5}$$

Con este criterio obtendremos todas las conexiones entre nodos, que serán bidireccionales. Si a ve a b, entonces b ve a a. Esta algoritmo mide la visibilidad en función de la geometría de la serie temporal.

El programa que hemos implementado a partir de los valores de desplazamiento neto de la serie temporal, crea un grafo con tantos nodos como nucleótidos tenga la secuencia y con las propiedades previamente descritas. Añade todas las conexiones del grafo, comprobando el criterio de visibilidad entre todos los puntos $a(t_a, y_a)$ y $b(t_b, y_b)$ del grafo.

En la Figura 4 podemos ver un ejemplo gráfico de como funciona el algoritmo de visibilidad.

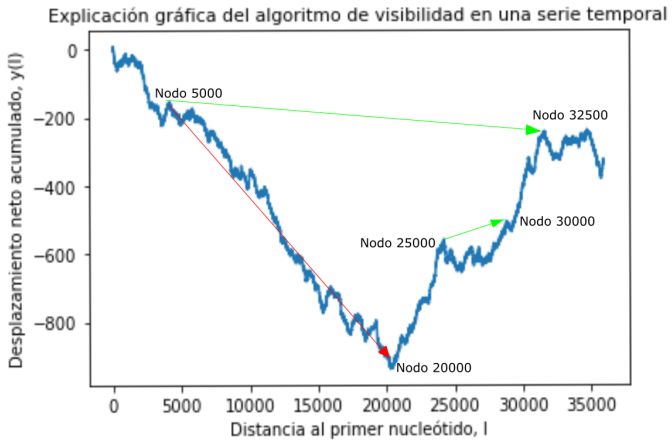


Figura 4: Ejemplo gráfico del algoritmo de visibilidad. Entre el nodo 5.000 y el nodo 20.000 hay numerosos puntos que impiden su visibilidad directa por lo que esos dos nodos no estarán conectados en el grafo. Esto no sucede, por ejemplo, entre los nodos 5.000 y 32.500 y entre los nodos 25.000 y 30.000 que sí estarán conectados en el grafo.

2.5. Propiedades asociadas a los grafos.

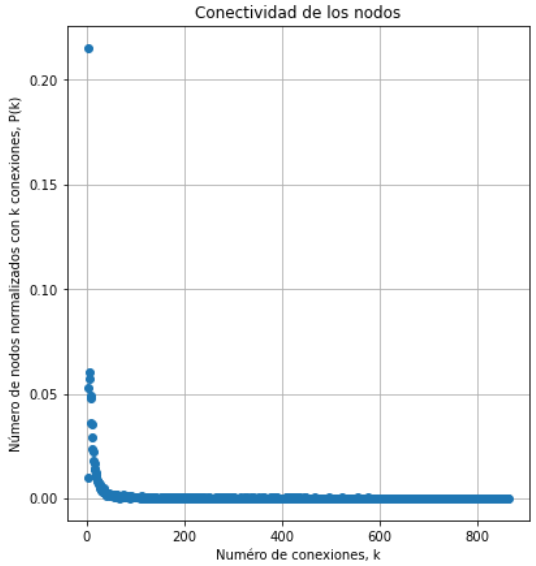
Hay numerosas propiedades que tienen los grafos. La primera de ellas será la de conectividad del grafo y la segunda será la propiedad del mundo pequeño [10], que cuantificaremos mediante el cálculo de la ruta media más corta entre nodos.

2.5.1. Grado de conectividad del grafo.

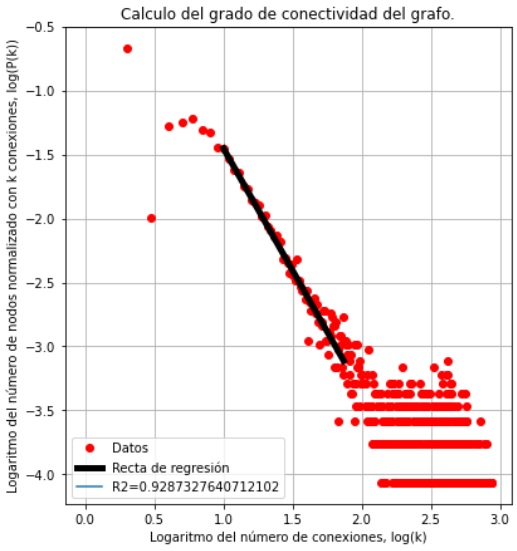
Cada nodo de nuestro grafo estará conectado a otros nodos, mínimo 2, como norma general. El grado de conectividad del grafo será una medida de su distribución de conexiones, $P(k)$. Para ver el grado de conectividad de nuestro grafo, vamos a calcular cuántos nodos con n conexiones tenemos. Cuando n sea pequeño como norma general tendremos muchos nodos, y cuando n vaya aumentando el número de nodos con ese numero mayor de conexiones ira disminuyendo. Para ver este efecto de forma numérica nos fijamos en la Figura 5b donde hay muchos nodos con pocas conexiones (aproximadamente $\frac{1}{3}$ de los nodos tienen entre 2 y 4 conexiones), hay algunos nodos que presentan mayor número de conexiones (aproximadamente $\frac{1}{3}$ de los nodos tienen entre 5 y 15 conexiones) y hay pocos nodos que presentan un número grande de conexiones (aproximadamente $\frac{1}{3}$ de los nodos tienen entre 16 y 875 conexiones). Obtendremos una función que aproximamos a

una potencial (Figura 5a), así pues, como en casos anteriores, representaremos en doble escala logarítmica, donde aproximaremos a una recta (Figura 5b, cuya pendiente será nuestro grado de conectividad del grafo. Llamamos a este parámetro VG.

El programa que hemos hecho para calcular esto crea una lista con el número de conexiones y crea otra lista que en la misma posición va guardando cuantos nodos hay con esas conexiones, mientras recorre el grafo. Una vez ha terminado de recorrer el grafo, en doble escala logarítmica lo ajusta a una recta de regresión. En la Figura 5b vemos un ejemplo del grafo de la secuencia del gen de BMCC1 de *Homo sapiens*.



(a) Frecuencia normalizada de nodos con k conexiones.



(b) Ajuste de recta de regresión del gráfico de la izquierda en escala doble logarítmica. Estimación de VG.

Figura 5: Estimación del grado de conectividad del grafo del gen BMCC1 de *Homo sapiens*. Podemos ver en la Figura 5a la estimación de VG. Para un ajuste mejor de la recta de regresión no tenemos en cuenta los puntos que forman lineas horizontales. Estas líneas horizontales se forman porque cuando el número de conexiones k aumenta, solo hay 1,2,3 o 4 nodos que presentan ese número de conexiones.

2.5.2. Propiedad del mundo pequeño.

Otra propiedad que vamos a estudiar de los grafos obtenidos, tiene que ver con el fenómeno del mundo pequeño [19], esto ocurre cuando dos nodos del grafo no son vecinos entre sí, y sin embargo, el número de conexiones que los separa es pequeño. Por ejemplo en la Figura 6 la longitud media del camino más corto entre nodos es inferior a 5. Esto ocurre, como pudimos observar en el la Figura 5a, debido a la presencia de algunos nodos con una alta densidad de conexiones que permiten reducir el numero de interconexiones que separan los nodos más distanciados. Para cuantificar este fenómeno estudiamos la variación de la longitud media del camino más corto entre todos los nodos a medida que aumentamos el tamaño de nuestro grafo.

La longitud media mas corta entre nodos, $MPL(n)$ crece proporcionalmente al logaritmo del número de nodos n en la red. Es decir $MPL(n) \propto log(n)$.

Se crea un grafo a partir de las conexiones entre nodos. También se seleccionan varios valores de n para los que se crean subgrafos con el número n de nodos y calcula para cada subgrafo la longitud media más corta entre nodos, $MPL(n)$.

Ajustamos los datos de $log(MPL(n))$ y $log(n)$ a una recta de regresión, cuya pendiente será el parámetro que cuantifica nuestra propiedad del mundo pequeño, a partir de aquí este parámetro será conocido como MPL.

Para encontrar la ruta más corta entre dos nodos hemos utilizado un método conocido como búsqueda en anchura[20]. Es un método para grafos no direccionados como el nuestro, escogimos este por ser rápido y eficiente, algo muy importante debido al gran tamaño de nuestros grafos. No teníamos los recursos computacionales suficientes para calcular este parámetro en secuencias con

una longitud mayor a 7300 pb. El grupo de secuencias donde se ha podido calcular este parámetro consta de 209 secuencias que es un 66,14 % del total de secuencias.

En la Figura 6 vemos un ejemplo con el gen de la cromato reductasa de *Pseudomonas Putida*.

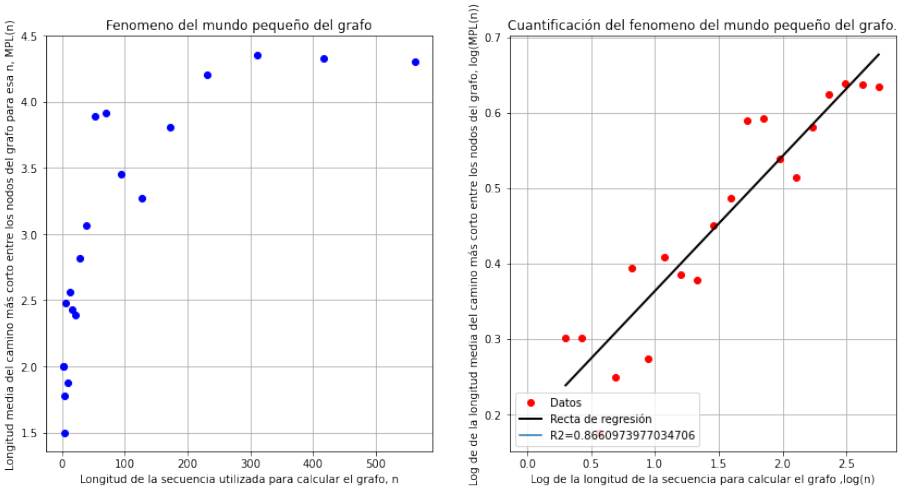


Figura 6: Fenómeno del mundo pequeño. En el gráfico de la izquierda podemos observar que cuando empieza a aumentar la longitud de la secuencia, el MPL aumenta rápidamente, ya que es muy fácil encontrar nodos muy separados, pero llega un momento que es muy difícil encontrar nodos que estén separados por más de 5 nodos, aunque el tamaño del grafo haya aumentado considerablemente. En el gráfico de la derecha podemos observar los valores en doble escala logarítmica. El valor de la pendiente de la recta de regresión que obtenemos será nuestro parámetro MPL

2.6. Análisis de los resultados

Los análisis estadísticos se realizaron con el software Python. El principal análisis estadístico realizado en este trabajo ha sido la comparación entre dos o más conjuntos de secuencias.

2.6.1. Normalidad

Los análisis que comparan medias como t-Student y ANOVA necesitan que la distribución de los datos que se están comparando sigan una distribución Gaussiana. Para evaluar si nuestros datos seguían este tipo de distribución hemos realizado test de normalidad de Shappiro-Wilkins [21] con un nivel de significación del 0.05.

Cuando algún grupo de secuencias no se distribuía significativamente de forma normal, hemos eliminado las valores atípicos o 'outliers' hasta que el nivel de significación de la distribución subía por encima del 0.05.

2.6.2. Homocedasticidad

Los análisis que comparan medias como t-Student y ANOVA asumen que la distribución de datos presenta homocedasticidad, es decir, que la varianza de los errores es constante a lo largo de las observaciones. Para evaluar si nuestros datos siguen este tipo de comportamiento hemos realizado un test de Levene [22] con un nivel de significación del 0.05.

Cuando los resultados de una subpoblación no presentaban homocedasticidad, hemos actuado de la misma manera que con la normalidad, hasta que hemos obtenido un nivel de significación superior al 0.05.

2.6.3. Comparación de las medias de 2 o más poblaciones

Una vez nuestras distribuciones de los resultados de las diferentes poblaciones separadas por parámetros y características siguen una distribución normal y presentan homocedasticidad, pasamos a encontrar diferencias significativas en sus medias.

En el caso de que comparemos 2 poblaciones, por ejemplo las poblaciones con y sin intrones, vamos a utlizar un test t de Student [23], donde la hipótesis nula H_0 dice que las medias de los dos grupos a comparar sean iguales. Utilizamos un nivel de significación del 0.05.

En el caso de que comparemos más de 2 poblaciones, por ejemplo las poblaciones separadas por reino, primero vamos a realizar un test ANOVA [24], donde la hipótesis nula H_0 es que todas las medias de los grupos son iguales con un nivel de significación del 0,05. En el caso de rechazar esta hipótesis nula, pasamos a realizar un test Tukey [25] para el análisis Post-Hoc, que compara todas las medias de los grupos 2 a 2 con el mismo nivel de significación que el test ANOVA.

2.7. Clasificación de los datos

Hemos utilizado diferentes métodos de inteligencia artificial con el fin de poder caracterizar secuencias a partir de los parámetros del trabajo.

2.7.1. Aprendizaje automático no supervisado (k-medias).

El primero de todos estos métodos de clustering es el *k-means* o k-medias, es un método no supervisado. Este tipo de algoritmos busca patrones en los datos de los parámetros sin tener en cuenta los datos de la caracterización. Lo único que tenemos que especificar al algoritmo es el número de grupos que queremos dividir nuestros datos.

Este método va a asignar a cada punto un grupo en función de la estructura de los datos y el número de grupo que le indiquemos. Para poder medir la efectividad de clasificación de este método vamos a comparar los datos obtenidos con los datos esperados. Para medir cuanto se ajustan estos grupos a los grupos que nosotros teníamos hemos utilizado el estadístico Índice de Rand Ajustado (ARI) [26], el Índice de Rand mide el porcentaje de acierto de la clasificación. $RI = \frac{a+b}{a+b+c+d}$ donde a es el número de verdaderos positivos en la clasificación y b el número de verdaderos negativos en la clasificación, por tanto, $a + b$ medida del número de aciertos entre los datos obtenidos y los datos esperados y donde $a + b + c + d$ es el número de datos posibles, siendo c y d los valores de falsos negativos y falsos positivos. El Índice de Rand Ajustado (ARI) normaliza numerador y denominador teniendo en cuenta los valores esperados cuando las dos particiones se hacen al azar.

2.7.2. Aprendizaje automático supervisado (k-Vecinos más próximos).

El segundo método de clustering que vamos a utilizar es un algoritmo supervisado, es decir que tenemos etiquetados con las características nuestro conjunto de entrenamiento. El conjunto de entrenamiento será del 70 % de los datos frente al 30 % del conjunto de validación. Es un algoritmo muy sencillo de implementar y que funciona muy bien en conjunto de datos pequeños como el de nuestro trabajo.

Este tipo de algoritmo funciona midiendo la distancia entre los elementos a clasificar y el resto de elementos del conjunto de entrenamiento. Después selecciona los k elementos más cercanos (la función de distancia que usamos es la euclidiana). Entre los k elementos más cercanos se decide la categoría del elemento a clasificar en función de la característica que defina a la mayoría de los k elementos más cercanos.

En nuestro trabajo hemos usado un k de 11, es esencial elegir un número impar para poder desempatar ciertos valores. En la Figura 7 vemos un ejemplo utilizando los parámetros MA(q) y Peng-α para clasificar mediante un K-NN si las secuencias presentan o no intrones.

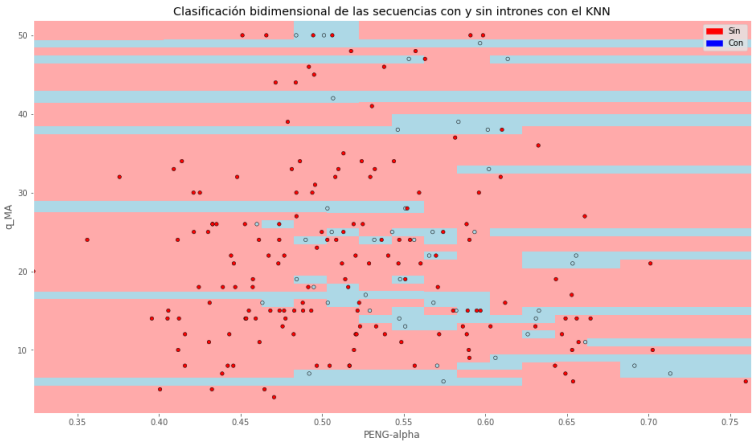


Figura 7: Ejemplo de clasificación bidimensional mediante un K-NN. El eje X contiene el valor del parámetro Peng-α y el eje Y contiene el valor q del modelo MA. A partir del conjunto de entrenamiento se calcula el fondo del gráfico mediante la distancia a los K-Vecinos más próximos. En la zona azul hay más K-vecinos más próximos de secuencias con intrones y en la zona roja hay más K-vecinos más próximos de secuencias sin intrones. El conjunto de validación está formado por lo puntos cuyo color representa la característica real de la secuencia.

En el fondo de la Figura 7 podemos observar que las zonas que han sido entrenadas para ser secuencias con intrones y secuencias sin intrones según la distancia a los elementos del conjunto de entrenamiento, los puntos del gráfico son los puntos del conjunto de validación. Para este ejemplo, al comparar las características reales de las secuencias y las características esperadas por el algoritmo K-NN hemos obtenido un porcentaje de acierto del 81 %.

Este concepto lo podemos ampliar hasta el número de parámetros que queramos, obtendremos un espacio de dimensión n divido en zonas. Cada zona estará clasificada en un determinado grupo de la clasificación de la característica que estemos estudiando. Esto solo lo podemos visualizar cuando haya un espacio de dimensión 3 como máximo. Hemos calculado todos los K-NN para todas las combinaciones de los 10 parámetros del trabajo, teniendo en cuenta que si utilizamos MPL el número de secuencias se vera reducido. Para cada K-NN analizaremos el porcentaje de acierto del K-NN sobre el conjunto de validación.

2.7.3. Aprendizaje profundo. Clasificación mediante redes neuronales.

Una red neuronal son modelos computacionales que están inspirados en las conexiones cerebrales. Las redes neuronales son entrenadas para ser capaces de relacionar ciertos valores de entrada con unos valores de salida. Se caracterizan porque las relaciones entre los valores de entrada y de salida no son lineales [27]. Nuestros valores de entrada van a ser los parámetros de las series temporales y grafos asociadas a las secuencias de nucleotidos y nuestros valores de salida serán la característica.

Hemos creado 10 neuronas distintas, dos para cada grupo de clasificación. El esquema general de estas redes neuronales lo podemos ver en la Figura 8.

Le presentamos las características a nuestra red neuronal en codificación One-hot, o lo que es lo mismo, tendremos un vector lleno de 0, excepto en la posición asignada al grupo correspondiente que se encontrará con un valor 1 [28]. Por ejemplo, para el contenido en intrones, el vector (1,0) significaría secuencia sin intrones y (0,1) secuencia con intrones.

Tendremos 2 redes neuronales por cada característica, una de estas tendrá nueve entradas en la primera capa (ya que estaremos analizando nueve parámetros) y la otra tendrá diez entradas en la primera capa (además de los nueve parámetros anteriores utilizaremos MPL).

Nuestro modelo esta ajustado para que se utilice como optimizador 'Adam' y la métrica para evaluar sea la precisión (*accuracy*). La función de perdida para las características de contenido de intrones y codificación del Ácido Nucleico es 'binary_crossentropy', ya que solo presentan dos grupos por característica. Para el resto la función de perdida utilizada ha sido la de 'categorical_crossentropy', ya que presentan mayor números de grupos por característica.

La capa de salida tendrá tantas neuronas como grupos haya en cada característica, por

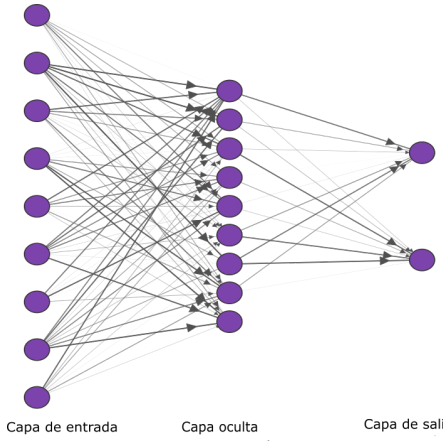


Figura 8: Esquema de la Red Neuronal multicapa. La capa de entrada presenta una función de activación de tipo *relu* o rectificadora, la capa oculta tiene una función de activación de tipo *relu* y la capa de salida tiene una función de activación sigmoide. Estas funciones de activación han sido escogidas para aumentar el porcentaje de acierto. La función sigmoide nos asegura en la capa de salida que nuestra salida de red se encontrará entre 0 y 1 y es fácil de mapear a cualquier probabilidad de clase 1.

ejemplo en los reinos serian 6. Cada neurona de esta capa de salida nos calcula la probabilidad de que los parámetros que hayamos introducido en la primera capa sean del grupo asociado a esa neurona. El grupo asignado será el que presente mayor probabilidad en su neurona.

Una vez entrenada la red, al introducir los parámetros asociados a una secuencia, la primera neurona nos da un valor de 0.001 y la segunda de 0.999. Esto quiere decir que la red ha calculado que hay una probabilidad del 99.9 % de que esos parámetros pertenezcan a una secuencia de AN con intrones.

Una vez entrenada la red con nuestro conjunto de entrenamiento calculamos el porcentaje de acierto con nuestro conjunto de validación.

Para elegir los conjuntos de validación y de entrenamiento, hemos utilizado la validación cruzada, ya que no teníamos un número muy elevado de secuencias. Una técnica que otorga robustez a los resultados cuando no tenemos un elevado número de datos [29]. La validación cruzada consiste en repetir la creación de la red neuronal para distintos conjuntos de entrenamiento y de validación. El conjunto de validación será siempre del 10 % pero iremos rotando este porcentaje a lo largo de todos los datos, primero cogeremos el primer 10 % de los datos, luego los datos entre el 10 % y el 20 %, y así sucesivamente hasta llegar al final. Por tanto obtendremos 10 conjuntos de validación distintos. La red será entrenada con el 90 % de datos restantes. Tendremos 10 porcentajes de acierto, el porcentaje de acierto de la red para cada característica será la media de estos 10 porcentajes.

2.8. Herramientas computacionales

Las diferentes herramientas descritas previamente han sido programas en ambientes de *Python* [30] y *RStudio* [31].

Para el tratamiento con secuencias en formato FASTA hemos utilizado la librería *Bio* de *Python* [32], para el análisis de los modelos que describen las series temporales hemos utilizado la función *auto.arima* de *RStudio* [33], para el tratamiento y estudio de grafos hemos utilizado la librería *NetworkX* de *Python* [34], para el uso de métodos de clustering de aprendizaje automático hemos utilizado la librería *Sklearn* de *Python* [35], y para la creación y entrenamiento de Redes Neuronales hemos utilizado las librerías *Keras* y *TensorFlow* de *Python* [36].

Para algunos cálculos muy pesados computacionalmente, se han optimizado los programas implementando estructuras de C en Python [37]. Todos los programas han corrido en un servidor del Departamento de Matemática Aplicada de la UPM.

CAPÍTULO 3. RESULTADOS Y DISCUSIÓN.

Comenzamos el apartado de resultados del trabajo presentando el Cuadro 2 donde hacemos un resumen estadístico de todos los parámetros previamente explicados:

Método	Parámetro	N	$\bar{x}$	$\hat{\sigma}^2$	mín	máx	$\hat{R^2}$
AR	p	316	0.047	0.147	0	4	-
MA	q		19.158	124.038	2	50	
ARMA	p		0.041	0.084	0	3	
	q		15.601	160.266	0	50	
ARIMA	p		2.206	12.913	0	20	
	d		1.370	0.234	1	2	
	q		1.313	1.759	0	6	
Peng-α			0.551	0.007	0.323	0.796	0.991
VG		312	-1.498	0.044	-2.000	-0.912	0.883
MPL		191	0.173	0.003	0.021	0.329	0.723

Cuadro 2: Tabla resumen estadístico sobre los parámetros del trabajo. N representa el número de secuencias para las que hay calculado cada parámetro, $\bar{x}$ representa la media de cada parámetro, $\hat{\sigma}^2$ representa la varianza de cada parámetro, mín y máx representan lo valores mínimos y máximos de cada parámetro, el $\hat{R^2}$ representa la media de los R^2 usados para el ajuste de cada recta en el cálculo de los parámetros Peng-α, VG y MPL.

Los programas que ejecutan los algoritmos descritos en las Secciones 2.2, 2.3, 2.4 y 2.5 se pueden encontrar en el siguiente repositorio de código de GitHub.

Transformamos las 316 secuencias en series temporales mediante el algoritmo descrito en el Capítulo 2.2.

Se valido el correcto funcionamiento de los programas antes de realizar el calculo a todo el grueso de secuencias. En el caso del parámetro Peng-α, se corroboro que nuestro programa arrojaba los mismos resultados en las mismas secuencias que se presentaban en ese trabajo [8]. Con respecto al resto de parámetros, los programas que calculaban estos parámetros se validaron mediante gráficos donde nos asegurábamos del correcto funcionamiento de los mismos.

3.1. Análisis de los parámetros del trabajo en los grupos de estudio.

3.1.1. Peng-α.

Calcularemos el parámetro Peng-α de las series temporales, utilizando el método DFA como hemos explicado en el Capítulo 2.3.2.

Vamos a estudiar los resultados del parámetro Peng-α para las 316 secuencias del trabajo y su posible utilización como clasificador de diferentes grupos. Para ello realizaremos test de medias para estudiar si hay diferencias significativas entre las medias de los grupos.

Podemos observar en la Figura 9 un histograma con los resultados del parámetro Peng-α.

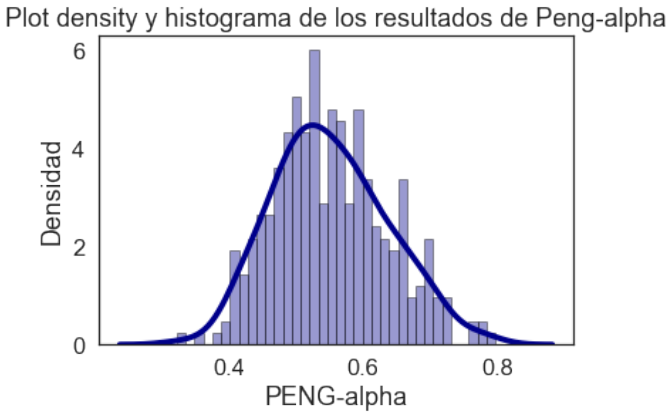


Figura 9: Histograma con los resultados del parámetro Peng-α de las 316 secuencias del trabajo. Podemos ver representada la curva de distribución sobre el histograma en un color más oscura. Los datos se distribuyen siguiendo una distribución normal ($p - valor = 0{,}052$).

La media para todas las secuencias de este parametro es 0.551. Observamos que esta media no varía en función de la longitud de la secuencia; al dividir el conjunto de secuencia en función de la longitud. No obstante, la varianza sí cambia al variar la longitud de secuencia. Como podemos observar en la Figura 10. A partir de una longitud de secuencia de 10000 pb, señalado con la recta vertical, la varianza disminuye. por tanto, cuanto mayor sea la longitud de secuencia, especialmente a partir de este tamaño, más robusta sera la estimación del Peng-α.

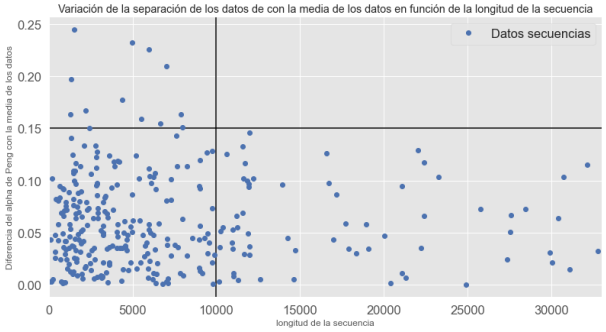


Figura 10: Distancia entre el valor Peng-α y el valor de la media en función de la longitud de la secuencia (pb).

A continuación vamos a presentar los resultados de este parámetro para los diferentes grupos en los que se ha clasificado el trabajo y si pueden ser utilizados como un clasificador.

Contenido de intrones.

Queremos comprobar si el parámetro Peng- α es capaz de resultar útil para decidir si una secuencia contiene o no intrones. Para ello, realizamos un test t de Student [Sección 2.6.3] para verificar si las medias de dicho parámetro son significativamente diferentes. Disponemos de 107 secuencias con intrones y 192 sin ellos. Verificamos las hipótesis de normalidad [Sección 2.6.1] (con $p - values$ de 0.061 (con intrones) y de 0,086 (sin intrones)) y la hipótesis de homocedasticidad [Sección 2.6.2] ($p - valor = 0{,}155$) para ambos grupos de secuencias. El análisis estadístico de los resultados para este parámetro se puede ver en el Cuadro 3:

Contenido de intrones	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Sin intrones	192	0.5206	0.0057	$2{,}315 \cdot 10^{-12}$
Con intrones	107	0.5813	0.0042	

Cuadro 3: Resultados del test t de Student que compara las medias del parámetro Peng-α en la clasificación por contenido de intrones. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test de medias t de Student.

El test t de Student arroja un $p-valor = 2{,}315 \cdot 10^{-12}$ por lo tanto comprobamos que las medias son significativamente diferentes para un nivel de confianza de 0,05. Este parámetro puede funcionar como clasificador para esta característica, ya que presenta un nivel de significación muy bajo comparado con la referencia del 0.05 que utilizamos en el trabajo.

Podemos ver en la Figura 11 representados los histogramas de los grupos con y sin intrones para este parámetro.

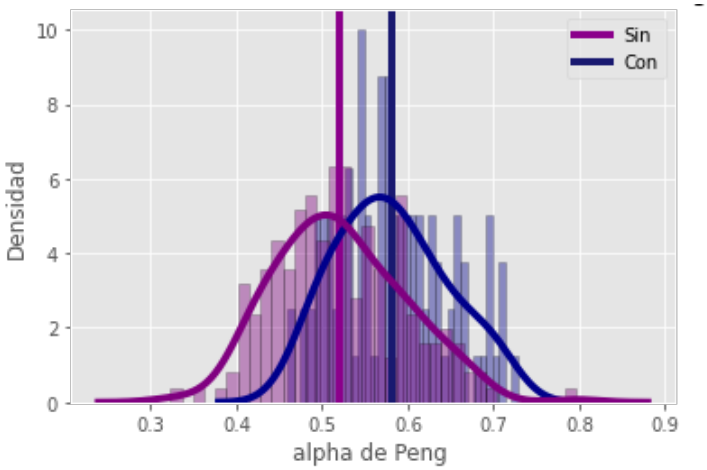


Figura 11: Histogramás de los resultados del parámetro Peng-α en función del contenido en intrones. La barras verticales más oscuras representan las medias poblaciones de cada grupo. También se encuentran dibujadas las curvas de distribución sobre los histogramas de cada grupo.

En trabajos previos ya se estudió el valor de este mismo parámetro en grupos de secuencias con intrones y sin intrones [8]. En ese trabajo se concluyó que las medias de este parámetros en ambos grupos eran distintas, para un grupo de 10 secuencias sin intrones la media de Peng-$\alpha \pm (2 \cdot e.e.m) = 0{,}50 \pm 0{,}01$ y para un grupo de 14 secuencias con intrones la media de Peng-$\alpha \pm (2 \cdot e.e.m) = 0{,}61 \pm 0{,}03$. Nuestros resultados son para un grupo de 192 secuencias sin intrones la media de Peng-$\alpha \pm (2 \cdot e.e.m) = 0{,}52 \pm 0{,}01$ y para un grupo de con intrones la media de Peng-$\alpha \pm (2 \cdot e.e.m) = 0{,}58 \pm 0{,}01$. Nuestros datos con un mayor número de secuencias para cada grupo presentan el comportamiento descrito en el trabajo de Peng [8], las secuencias con intrones presentan correlación de largo alcance (Peng-$\alpha > 0{,}5$) mientras que las secuencias sin intrones presentan correlación de corto alcance (Peng-$\alpha = 0{,}5$).

Reino al que pertenece la secuencia de Ácido Nucleico.

Queremos comprobar si el parámetro Peng- α es capaz de resultar útil para decidir el reino del organismo al que pertenece la secuencia de Ácido Nucleico. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.277 (Animal), de 0,076 (Mónera), de 0.189 (Fungi), de 0.288 (Vegetal), de 0.138 (Protista) y de 0.141 (Virus)) y la hipótesis de homocedasticidad ($p-valor = 0{,}493$) para todos los grupos de secuencias.

El resultado del test ANOVA lo podemos ver en el Cuadro 4

Reino	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Animal	160	0.578603	0.0051	$2{,}11 \cdot 10^{-11}$
Mónera	36	0.475746	0.0035	
Fungi	27	0.524364	0.0061	
Vegetal	50	0.5472	0.0069	
Protista	14	0.5051	0.0095	
Virus	25	0.5227	0.0051	

Cuadro 4: Resultados del test ANOVA que compara las medias del parámetro Peng-α en la clasificación por reino. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

El test ANOVA arroja un $p-valor = 2{,}11 \cdot 10^{-11}$, por lo que asumimos que las medias no son significativamente iguales. Tras el test ANOVA haber rechazado que las medias eran iguales hemos hecho un test de Tukey para el análisis post-Hoc para identificar las medias de que subconjuntos son significativamente iguales. Podemos ver el resultado de este test en la Figura 12.

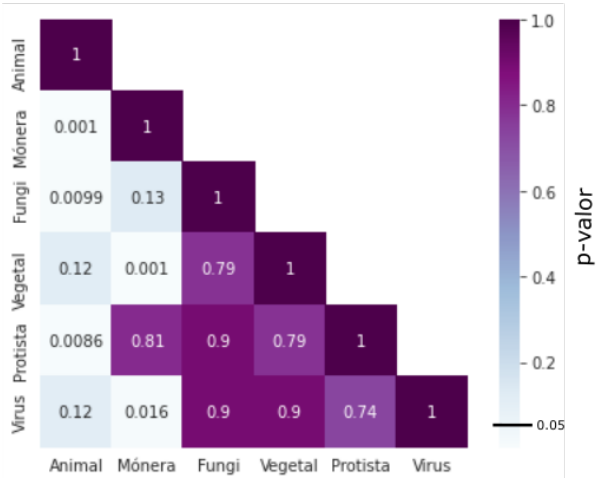


Figura 12: Test Tukey para el parametro Peng-α de los grupos clasificados por su reino. Podemos ver los $p-valores$ para cada par de grupos, el color del fondo del cuadrado será más oscuro cuanto mayor sea $p-valor$.

En este caso las medias diferentes entre pares de grupos ($p - valor < 0{,}05$) son: Animal y Mónera (0.001); Animal y Fungi (0.0099); Animal y Protista (0.0086); Mónera y Vegetal (0.001) y Mónera y Virus (0.0157). El parámetro sí que puede servir para clasificar las secuencias en estos casos. De los pares de grupos que presentan diferencias significativa en las medias podemos observar que hay diferencias entre los grupos de eucariotas (Vegetal, Animal, Fungi) y el grupo de procariotas (Mónera). Realizamos un test de Student para ver si hay diferencias significativas en las medias de los grupos de secuencias de eucariotas y de procariotas. Se cumplen las hipótesis de normalidad y homocedasticidad. El test de medias arroja como resultado un $p-valor = 2 \cdot 10^{-13}$, por tanto este parámetro presenta medias significativamente distintas para los dos grupos y podría servir para clasificarlos. Esto puede estar relacionado con las diferencias en los porcentajes de regiones codificantes y no codificantes entre las secuencias de eucariotas y procariotas [6].

Tipo de Ácido Nucleico.

Queremos comprobar si el parámetro Peng- α es capaz de resultar útil para decidir si una secuencia es de ARN o ADN, y si esta secuencia es lineal o circular. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p - valores$ de 0.965 (ADN lineal), de 0,258 (ADN circular) y de 0.056 (ARN lineal) y la hipótesis de homocedasticidad ($p-valor = 0{,}118$) para todos los grupos de secuencias.

El resultado del ANOVA y el resumen estadístico del parámetro en estos grupos se encuentra en el Cuadro 5

Tipo de AN	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
ADN lineal	163	0.5542	0.0067	0,0395
ADN circular	20	0.5749	0.0096	
ARN lineal	118	0.5348	0.0051	

Cuadro 5: Resultados del test ANOVA que compara las medias del parámetro Peng-α en la clasificación por tipo de ácido nucleico. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

El test ANOVA resulta con un $p-valor = 0{,}0395$ que las medias no son significativamente iguales entre estos grupos, al realizar el análisis post-Hoc, cuando comparamos por pares los grupos obtenemos que sus medias si son todas significativamente iguales ($p-valor > 0{,}05$). Este parámetro no presenta diferencias en las medias para estos grupos clasificados en función de su tipo de AN, no sirve como clasificador.

Codificación del Ácido Nucleico.

Queremos comprobar si el parámetro Peng- α es capaz de resultar útil para decidir si una secuencia es codificante o estructural. Para ello, realizamos un test t de Student para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.134 (Codificante) y de 0,114 (Estructural)) y la hipótesis de homocedasticidad ($p-valor = 0{,}052$) para ambos grupos de secuencias.

Podemos ver el resultado del test t de Student en el Cuadro 6.

Codificación del AN	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Codificante	256	0.546	0.0049	$4{,}592 \cdot 10^{-13}$
Estructural	34	0.4697	0.0027	

Cuadro 6: Resultados del test t de Student que compara las medias del parámetro Peng-α en los grupos de secuencias codificantes y secuencias estructurales. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test de medias t de Student.

El test resulta con un $p-valor$ muy pequeño en comparación al 0.05 que usamos de límite, por tanto podemos decir que las medias son significativamente distintas. Este parámetro lo podemos utilizar como clasificador.

Localización del Ácido Nucleico

Queremos comprobar si el parámetro Peng- α es capaz de resultar útil para decidir la localización del Ácido Nucleico en las células o en capsidas. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.336 (Núcleo), de 0,981 (Plásmido), de 0.069 (Mitocondria), de 0.712 (Cloroplasto), de 0.423 (Citoplasma) y de 0.1 (Cápsida)) y la hipótesis de homocedasticidad ($p-valor = 0{,}1$) para todos los grupos de secuencias.

El resumen estadístico de este parámetro en los distintos grupos y el resultado del test ANOVA lo podemos ver en el Cuadro 7.

Localización	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Núcleo	121	0.5693	0.0051	0,00054
Plásmido	13	0.4937	0.0024	
Mitocondria	14	0.5763	0.0101	
Cloroplasto	18	0.4826	0.0052	
Citoplasma	115	0.5525	0.0072	
Cápsida	26	0.5215	0.0049	

Cuadro 7: Resultados del test ANOVA que compara las medias del parámetro Peng-α en la clasificación por localización. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

Tras resultar el test ANOVA con un $p - valor = 0{,}00054$, las medias presentan diferencias significativas. Este parametro puede utilizarse como clasficador. Despúes del test ANOVA realizamos un analisi post Hoc con un test de Tukey que compara las medias de los grupos 2 a 2, el resultado de este test lo podemos ver en el Cuadro 13.

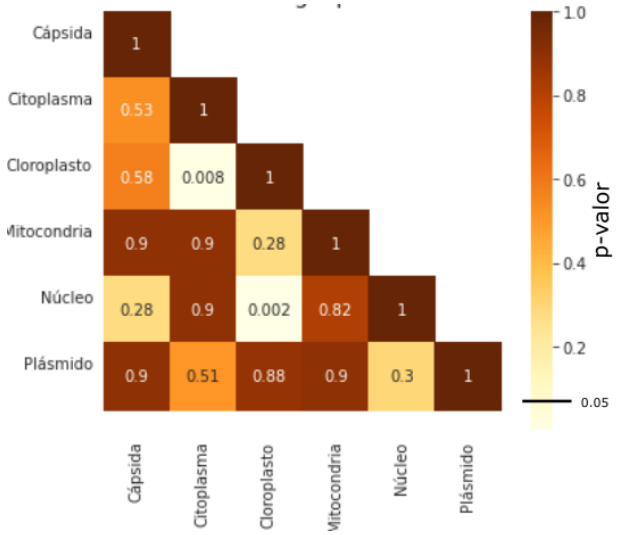


Figura 13: Test tukey para el parámetro Peng-α en los grupos clasificados por su localización. Podemos ver los $p-valores$ para cada par de grupos, el color del fondo del cuadrado será más oscuro cuanto mayor sea $p-valor$.

Las medias significativamente distintas fueron las siguientes: Citoplasma y Cloroplasto (0.0085); Cloroplasto y Núcleo (0.0021). Este parámetro nos puede sirve para observar diferencias en las medias de los pares de grupos anteriormente mencionados.

3.1.2. Grado de conectividad del grafo o VG.

Transformamos las series temporales que teníamos en grafos, mediante el algoritmo explicado en el Capítulo 2.4.

Calcularemos el parámetro VG, cuantificando el grado de conectividad de los grafos obtenidos, como hemos explicado en el Capítulo 2.5.1

Vamos a estudiar los resultados del parámetro VG para las 316 secuencias del trabajo y su posible utilización como clasificador de diferentes grupos. Para ello realizaremos test de medias para estudiar si hay diferencias significativas entre las medias de los grupos.

Podemos observar en la Figura 14 un histograma con los resultados del parámetro VG.

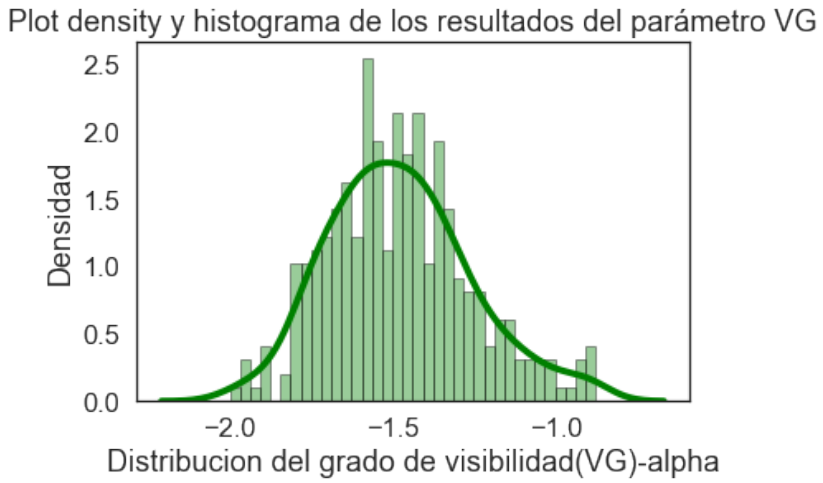


Figura 14: Histograma con los resultados del parámetro VG de las 316 secuencias del trabajo. Podemos ver representada la curva de distribución sobre el histograma en un color más oscura. Los datos se distribuyen siguiendo una distribución normal ($p - valor = 0{,}058$).

A continuación vamos a presentar los resultados de este parámetro para los diferentes grupos en los que se ha clasificado el trabajo y si pueden ser utilizados como un clasificador.

Contenido de intrones.

Queremos saber si el parametro VG puede resultar útil para decidir si una secuencia presenta intrones o no. Para ello realizaremos un test t de Student para verificar si las medias de los grupos son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.078 (con intrones) y de 0,160 (sin intrones)) y la hipótesis de homocedasticidad ($p - valor = 0{,}254$) para ambos grupos de secuencias.

El análisis estadístico de los resultado para este parámetro se puede ver en el Cuadro 3:

Contenido de intrones	N	$\bar{x}$	$\hat{\sigma}^2$	$p - valor$
Sin intrones	207	-1.4934	0.0472	0.507
Con intrones	105	-1.477	0.036	

Cuadro 8: Resultados del test t de Student que compara las medias del parámetro VG en la clasificación por contenido de intrones. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test de medias t de Student.

El nivel de significación es 0.507, muy superior al límite establecido del 0,05. Las diferencias en las medias de este parámetro en los grupos con intrones y sin intrones no son significativas. El parámetro no sirve para diferenciar esta característica.

Reino al que pertenece la secuencia de Ácido Nucleico.

Queremos comprobar si el parámetro VG es capaz de resultar útil para decidir el reino del organismo al que pertenece la secuencia de Ácido Nucleico. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p - valores$ de 0.277 (Animal), de 0,076 (Mónera), de 0.189 (Fungi), de 0.288 (Vegetal), de 0.138 (Protista) y de 0.141 (Virus)) y la hipótesis de homocedasticidad ($p - valor = 0{,}493$) para todos los grupos de secuencias.

El resumen estadístico de este parámetro en los distintos reinos lo podemos ver en el Cuadro 9.

Reino	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Animal	157	-1.443	0.0339	0,001632
Mónera	37	-1.549	0.0506	
Fungi	27	1.513	0.0497	
Vegetal	50	1.5464	0.0388	
Protista	12	-1.6134	0.0105	
Virus	28	-1.4866	0.0768	

Cuadro 9: Resultados del test ANOVA que compara las medias del parámetro VG en la clasificación por reino. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

Tras haber resultado el test ANOVA con un $p-valor = 0{,}001$, concluimos que las medias no son significativamente iguales. Hemos realizado un análisis post-Hoc con un test de Tukey. Con el resultado de este test podemos decir que todas las medias son significativamente iguales ($p-valor > 0{,}05$), menos entre el reino Vegetal y el reino Animal (0.0243). El parámetro no sirve para clasificar esta característica.

Tipo de Ácido Nucleico.

Queremos comprobar si el parámetro VG es capaz de resultar útil para decidir si una secuencia es de ARN o ADN, y si esta secuencia es lineal o circular. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.095 (ADN lineal), de 0,558 (ADN circular) y de 0.425 (ARN lineal) y la hipótesis de homocedasticidad ($p-valor = 0{,}05$) para todos los grupos de secuencias.

En el Cuadro 10 presentamos los resultados del test ANOVA, junto a un análisis estadístico del parámetro en los diferentes grupos.

Tipo de AN	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
ADN lineal	161	-1.5087	0.0421	0,1155
ADN circular	15	-1.5481	0.0122	
ARN lineal	133	-1.466	0.0443	

Cuadro 10: Resultados del test ANOVA que compara las medias del parámetro VG en la clasificación por tipo de ácido nucleico. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

El test ANOVA resulta que las medias de los 3 grupos son significativamente iguales. Por tanto podemos decir que este parámetro tampoco sirve para clasificar las secuencias según su tipo de ácido nucleico.

Codificación del Ácido Nucleico.

Queremos saber si este parámetro puede ser útil para clasificar si una secuencias es estructural o codificante. Para ello realizamos un test t de Student. Las distribuciones de este parámetro en las 2 poblaciones sí que siguen el criterio de normalidad. Esto no sucede con la homocedasticidad. No hay homogeneidad en las varianzas. Para que el test de Levene diera un resultado significativo tendríamos que eliminar mas del 70 % de las secuencias codificantes. Las varianza de la población de codificantes es más del doble que la varianza de la población de secuencias estructurales, 0.042 y 0.0179 respectivamente. Las medias de este parámetro son -1.4992 para secuencias codificantes y -1.4803 para secuencias estructurales, estas medias no parecen distintas pero no podemos comprobarlo debido a la imposibilidad de realizar un test estadístico. No consideramos que este parámetro sirva como clasificador.

Localización del Ácido Nucleico

Queremos comprobar si el parámetro VG es capaz de resultar útil para decidir la localización del Ácido Nucleico en las células o en capsidas. Para ello, realizamos un test ANOVA para verificar

si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p - valores$ de 0.052 (Núcleo), de 0,736 (Plásmido), de 0.05 (Mitocondria), de 0.17 (Cloroplasto), de 0.912 (Citoplasma) y de 0.51 (Cápsida)) y la hipótesis de homocedasticidad ($p - valor = 0{,}15$) para todos los grupos de secuencias.

Los resultados del test de ANOVA y el resumen estadístico lo podemos observar en el Cuadro 11.

Localización	N	$\bar{x}$	$\hat{\sigma}^2$	$p - valor$
Núcleo	119	-1.4923	0.0366	0,89451
Plásmido	14	0.4937	0.0024	
Mitocondria	14	-1.5236	0.0749	
Cloroplasto	19	-1.527	0.039	
Cápsida	28	-1.4866	0.0768	
Citoplasma	115	-1.4693	0.0434	

Cuadro 11: Resultados del test ANOVA que compara las medias del parámetro VG en la clasificación por localización. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p - valor$ que arroja el test ANOVA.

El test ANOVA que dice resulta que las medias son significativamente iguales. Por el parámetro no sirve para separar a las secuencias en esta característica.

3.1.3. Camino medio más corto entre nodos o MPL.

Este parámetro se calcula cuantificando la propiedad del mundo pequeño como se ha explicado en el Capítulo 2.5.2.

Vamos a estudiar los resultados del parámetro MPL. Para este parámetro, debido a la carga computacional, solo se ha podido calcular en secuencias con una longitud menor a 7300 pb. Tenemos un total de 191 secuencias en las que hemos podido calcular este parámetro. Vamos a analizar su posible utilización como clasificador de diferentes grupos. Para ello realizaremos test de medias para estudiar si hay diferencias significativas entre las medias de los grupos.

Podemos observar en la Figura 15 un histograma con los resultados del parámetro MPL.

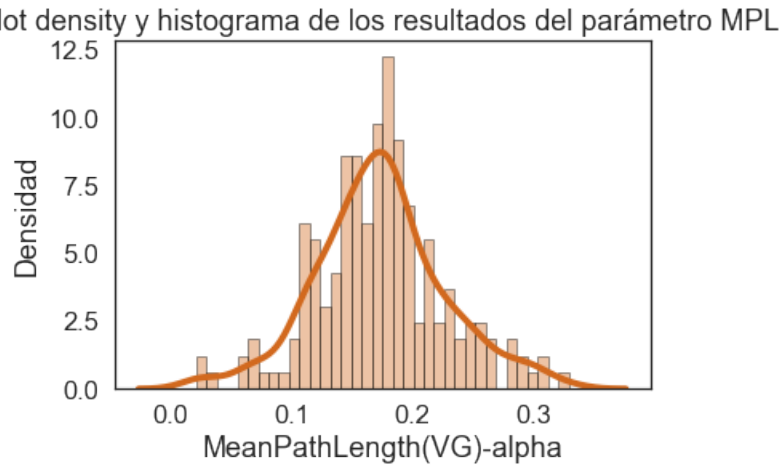


Figura 15: Histograma con los resultados del parámetro MPL de las 191 secuencias del trabajo. Podemos ver representada la curva de distribución sobre el histograma en un color más oscura. Los datos se distribuyen siguiendo una distribución normal ($p - valor = 0{,}073$).

A continuación vamos a presentar los resultados de este parámetro para los diferentes grupos en los que se ha clasificado el trabajo y si pueden ser utilizados como un clasificador.

Contenido de intrones.

Queremos saber si el parametro MPL puede resultar útil para decidir si una secuencia presenta intrones o no. Para ello realizaremos un test t de Student para verificar si las medias de los grupos son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.837 (con intrones) y de 0,1 (sin intrones)) y la hipótesis de homocedasticidad ($p-valor = 0{,}893$) para ambos grupos de secuencias.

El resultado del test t de Student y el análisis estadístico de los resultados para este parámetro se puede ver en el Cuadro 12:

Contenido de intrones	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Sin intrones	142	0.175	0.00235	0.491
Con intrones	44	0.169	0.00243	

Cuadro 12: Resultados del test t de Student que compara las medias del parámetro MPL en la clasificación por contenido de intrones. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test de medias t de Student.

El test arroja que la diferencia en las medias de este parámetro en los dos grupos no es significativa. El parámetro no sirve para diferenciar ambos grupos.

Reino al que pertenece la secuencia de Ácido Nucleico.

Queremos comprobar si el parámetro MPL es capaz de resultar útil para decidir el reino del organismo al que pertenece la secuencia de Ácido Nucleico. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.22 (Animal), de 0,3 (Mónera), de 0.266 (Fungi), de 0.118 (Vegetal), de 0.08 (Protista) y de 0.424 (Virus)) y la hipótesis de homocedasticidad ($p-valor = 0{,}053$) para todos los grupos de secuencias.

Podemos ver el resultado de este test ANOVA y un resumen estadístico de los parámetros en el Cuadro 13.

Reino	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Animal	70	0.1679	0.0032	0,00043
Mónera	36	0.1907	0.0048	
Fungi	25	0.1904	0.0051	
Vegetal	43	0.1629	0.0022	
Protista	14	0.178	0.0028	
Virus	19	0.2357	0.0052	

Cuadro 13: Resultados del test ANOVA que compara las medias del parámetro MPL en la clasificación por reino. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

Una vez el resultado del test ANOVA nos dice que no todas las medias son iguales, hemos hecho un analisis post-Hoc con un test de Tukey. Podemos ver el resultado de este test en el Cuadro 16.

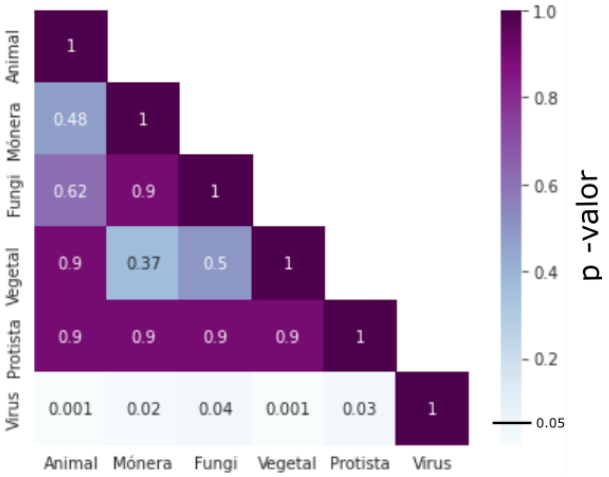


Figura 16: Test Tukey para el parámetro MPL de los grupos clasificados por su reino. Podemos ver los $p-valores$ para cada par de grupos, el color del fondo del cuadrado será más oscuro cuanto mayor sea $p-valor$.

Podemos ver en el Cuadro 16 que los grupos que presentan diferencias significativas en su medias son : Animal y Virus (0.001); Mónera y Virus (0.024); Fungi y Virus (0.0438); Protista y Virus (0.0274) y Vegetal y Virus (0.001).

Este parámetro destaca por que la media entre el grupo de secuencias de Virus y el resto de grupos es significativamente distinto. Si comparamos las medias del grupo de Virus con la media de un grupo formado por el resto de reinos mediante un t de Student (ambas distribuciones cumplen las hipótesis de normalidad y homocedasticidad), obtenemos un nivel de significación del test del 0.0004 muy inferior comparado con el límite del 0.05. Este parámetro diferencia muy bien las secuencias de Virus y las del resto de reinos.

Tipo de Ácido Nucleico.

Queremos comprobar si el parámetro MPL es capaz de resultar útil para decidir si una secuencia es de ARN o ADN, y si esta secuencia es lineal o circular. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.295 (ADN lineal), de 0,139 (ADN circular) y de 0.06 (ARN lineal) y la hipótesis de homocedasticidad ($p-valor = 0{,}223$) para todos los grupos de secuencias.

Podemos ver el resultado del test ANOVA y un resumen estadístico de los parámetros en el Cuadro 14.

Tipo de AN	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
ADN lineal	99	0.1857	0.0037	0,0049
ADN circular	16	0.1404	0.0032	
ARN lineal	83	0.1673	0.0026	

Cuadro 14: Resultados del test ANOVA que compara las medias del parámetro MPL en la clasificación por tipo de ácido nucleico. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

Una vez resultado el test ANOVA con un $p-valor < 0{,}05$ lo que significa que las medias no son significativamente iguales, el análisis post-Hoc con el test de Tukey da como resultado que los grupos que no son significativamente iguales son el ADN circular y el ADN lineal con un $p-valor = 0{,}0097$. Es el parámetro que mas diferencias en las medias presenta para los grupos de secuencias separadas por el tipo de AN.

Codificación del Ácido Nucleico.

Queremos comprobar si el parámetro MPL es capaz de resultar útil para decidir si una secuencia es codificante o estructural. Para ello, realizamos un test t de Student para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.06 (Codificante) y de 0,09 (Estructural)) y la hipótesis de homocedasticidad ($p-valor = 0{,}08$) para ambos grupos de secuencias.

El resultado del t de Student y el análisis estadístico del parámetro lo podemos ver en el Cuadro 15.

Contenido de intrones	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Sin intrones	158	0.177	0.0028	0.3935
Con intrones	33	0.1706	0.0017	

Cuadro 15: Resultados del test t de Student que compara las medias del parámetro MPL en los grupos de secuencias codificantes y secuencias estructurales. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test de medias t de Student.

El test resulta con un $p-valor = 0{,}394$, lo que nos dice que las medias no son significativamente iguales, no podemos utilizar este parámetro como clasificador.

Localización del Ácido Nucleico

Queremos comprobar si el parámetro MPL es capaz de resultar útil para decidir la localización del Ácido Nucleico en las células o en cápsidas. Para ello, realizamos un test ANOVA para verificar si las medias de dicho parámetro son significativamente diferentes. Verificamos las hipótesis de normalidad (con $p-valores$ de 0.073 (Núcleo), de 0,231 (Plásmido), de 0.202 (Mitocondria), de 0.196 (Cloroplasto), de 0.147 (Citoplasma) y de 0.18 (Cápsida)) y la hipótesis de homocedasticidad ($p-valor = 0{,}073$) para todos los grupos de secuencias.

Podemos ver el resultado de este test ANOVA y un resumen estadístico de los parámetros en el Cuadro 16.

Localización	N	$\bar{x}$	$\hat{\sigma}^2$	$p-valor$
Núcleo	67	0.1783	0.0022	0,00027
Plásmido	12	0.2014	0.0067	
Mitocondria	16	0.1964	0.0024	
Cloroplasto	18	0.169	0.0046	
Cápsida	18	0.2268	0.004	
Citoplasma	71	0.1597	0.0029	

Cuadro 16: Resultados del test ANOVA que compara las medias del parámetro MPL en la clasificación por localización. N corresponde al número de secuencias de cada grupo, $\bar{x}$ es la media de cada uno de estos grupos, $\hat{\sigma}^2$ es la varianza de los datos de cada grupo. También podemos observar el $p-valor$ que arroja el test ANOVA.

Tras haber resultado el test ANOVA con $p-valor = 0{,}00027$, las medias de los grupos no son significativamente iguales. Realizamos un test de Tukey para comparar las medias de los grupos por pares. Podemos ver el resultado de este test en la Figura 17

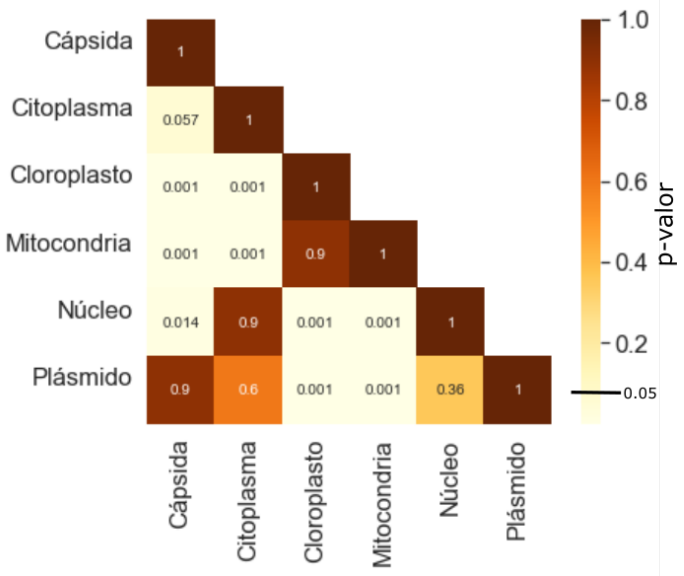


Figura 17: Test Tukey para el parámetro MPL de los grupos clasificados por su localización. Podemos ver los $p-valores$ para cada par de grupos, el color del fondo del cuadrado será más oscuro cuanto mayor sea $p-valor$.

Las medias de los grupos que son significativamente distintas ($p-valor < 0{,}05$) son: Cápsida y Cloroplasto (0.001); Cápsida y Mitocondria (0.001); Cápsida y Núcleo (0.0142); Citoplasma y Mitocondria (0.001); Citoplasma y Mitocondria (0.001); Cloroplasto y Núcleo (0.001); Cloroplasto y Plásmido (0.001); Mitocondria y Núcleo (0.001); Mitocondria y Plásmido (0.001).

La mayoría de las medias de los diferentes grupos son significativamente diferentes. El parámetro, por tanto, sirve para diferenciar las secuencias por su localización.

3.1.4. Modelos AR, MA, ARMA y ARIMA.

Se estudiaron los diferentes parámetros que aproximaban las series temporales a los procesos explicados en el Capítulo 2.3.1.

Vamos a estudiar los resultados de los parámetros de modelos autorregresivos y medias móviles de las 316 secuencias mediante un diagrama de cajas, que podemos ver en la Figura 18.

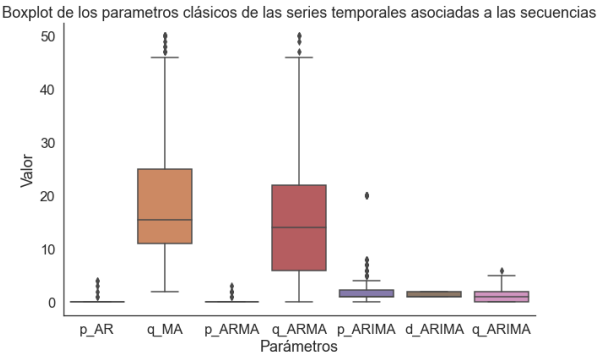
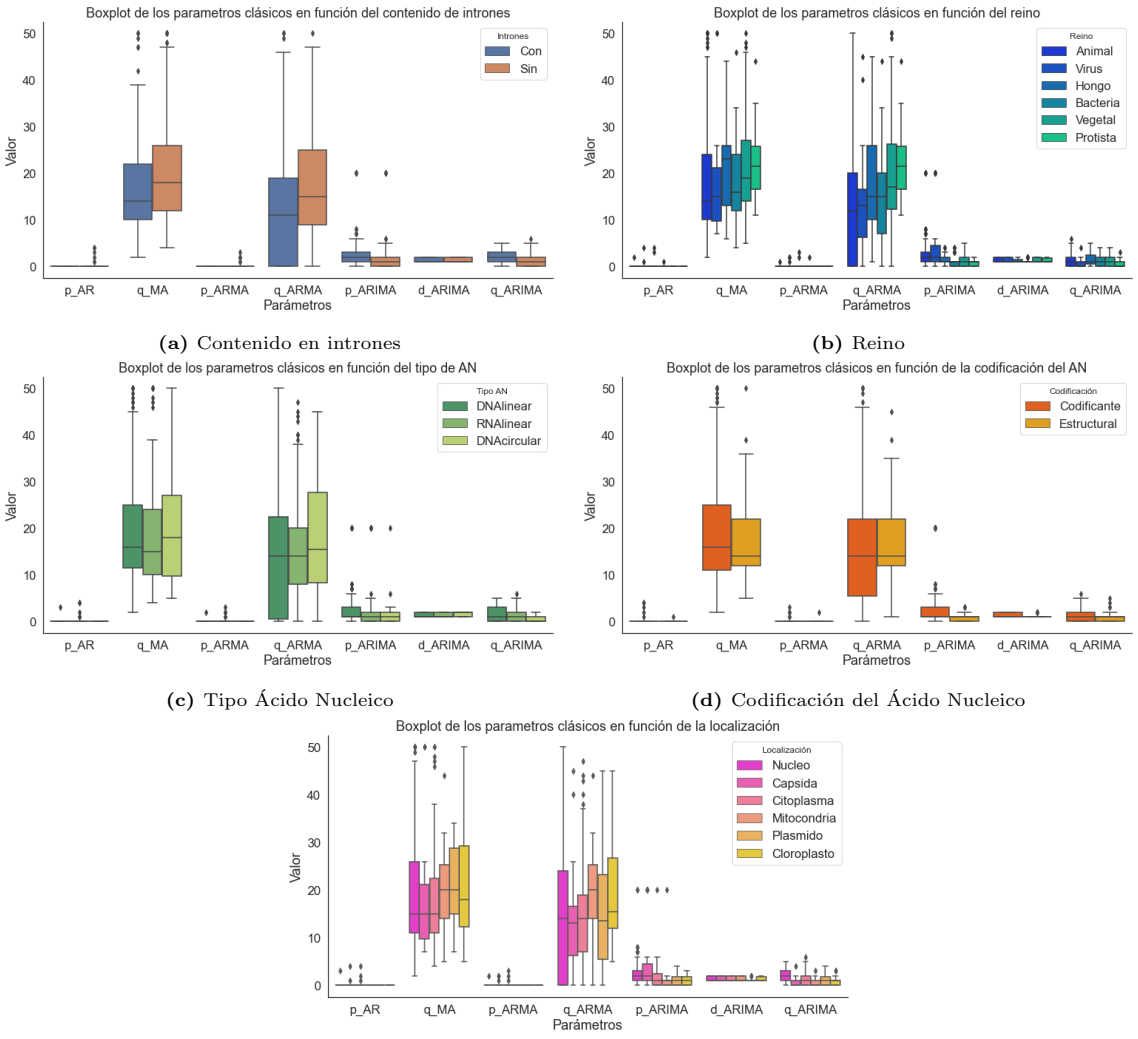


Figura 18: Diagrama de cajas de los parámetros clásicos para modelos de series temporales de todas las secuencias del trabajo. En este diagrama de cajas podemos ver las distribuciones de los diferentes resultados. Las cajas que aparecen coloreadas están delimitadas por el valor del primer cuartil y del tercer cuartil. La línea horizontal que se encuentra dentro de la caja corresponde al valor de la mediana.

Estos parámetros toman valores enteros por tanto no podemos realizar el análisis de las medias que hemos venido haciendo con el resto de parámetros. Pero de la Figura 18 podemos destacar, que las series temporales asociadas a las secuencias de nucleótidos destacan por su comportamiento de medias móviles (representado por los valores q de MA, ARMA) frente al comportamiento

autorregresivo (representado por los valores p de AR, ARMA). No presentan comportamientos autorregresivos, por tanto un valor de la secuencia no influye en el valor del siguiente. En los procesos integrados (ARIMA) los valores p, dyq parecen tener un comportamiento parecido, casi nulo. Las series temporales no parecen presentar un proceso integrado.

Los diferentes grupos de clasificación no parecen presentar diferencias en las medias de estos parámetros. Aparentemente todos los grupos mantienen este comportamiento de medias móviles. Todo esto lo podemos observar en la Figura 19.



(a) Contenido en intrones

(b) Reino

(c) Tipo Ácido Nucleico

(d) Codificación del Ácido Nucleico

(e) Localización del Ácido Nucleico

Figura 19: Diagrama de cajas de los parámetros clásicos de modelos autorregresivos y de medias móviles para los diferentes grupos de clasificación del trabajo. En este diagrama de cajas podemos ver las distribuciones de los diferentes resultados. Las cajas que aparecen coloreadas están delimitadas por el valor del primer cuartil y del tercer cuartil. La línea horizontal que se encuentra dentro de la caja corresponde al valor de la mediana.

3.2. Caracterización de secuencias mediante clustering no supervisado.

El metodo no supervisado que vamos a utilizar para la clasificación es el k-medias, que se encuentra explicado en la Sección 2.7.1.

El primer problema es encontrar el número óptimo de grupos.

Para obtener el número óptimo de grupos para nuestros datos, hemos representado el número de clusters posibles frente a una puntuación de bondad de ajuste, que es el inverso del BIC (*Bayesian Information Criterion*) [18].

El número óptimo de clusters es el mínimo en el que esta puntuación se ha estabilizado, la estimación de este punto es arbitraria.

En la Figura 20 podemos observar el ajuste de este criterio BIC para el clustering con todos los parámetros del trabajo. Al variar el conjunto de parámetros con los que hacemos el clustering no supervisado, el numero de clusters óptimos también varia.

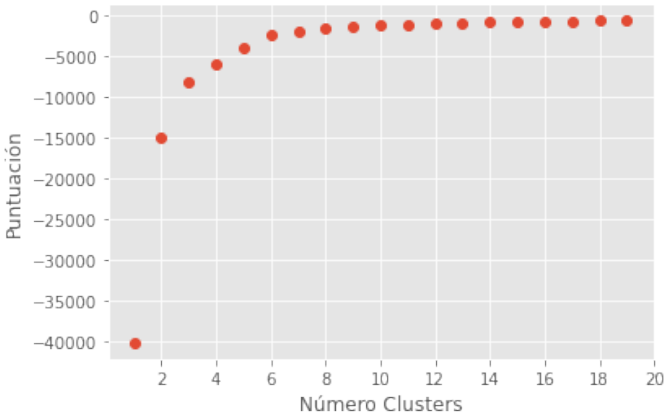


Figura 20: Determinación del número de k-grupos en el clustering para todos los parámetros. Podemos ver como a partir de dos k-grupos la puntuación sube considerablemente, luego hasta 6 k-grupos sigue subiendo considerablemente. Después de 6 se mantiene casi estable por eso hemos determinado que el número óptimo de k-grupos está entre 2 y 6.

Hemos realizado este método de clustering para todas las 1022 combinaciones sin repetición de todos los parámetros del trabajo. Para todas las combinaciones de parámetros hemos estimado que el número óptimo de clusters se encuentra entre 2 y 6.

Hay que destacar que cuando queríamos introducir el parámetro MPL teníamos que reducir el número de entradas a 209, por lo comentado en la Sección 2.5.2.

Evaluamos si el clustering que hace cada combinación de parámetros y su número óptimo de clusters, coincide con los grupos de características que tenemos, para ello hemos calculado el ARI 2.7.1.

Por ejemplo, cuando hemos utilizado todos los parámetros obteníamos 6 como número óptimo de clusters. Comparamos estos 6 grupos con las características que tenemos que presentan 6 grupos, en el caso de nuestro trabajo la clasificación por localización y por reino. Para la bondad de ajuste entre los 6 grupos no supervisados y los 6 grupos de ambas características calculamos el ARI.

El número de clusters utilizado en cada caso ha sido el número de clases de cada característica, por ejemplo en el caso del Reino el número de clases ha sido 6.

Habiendo realizado el k-medias con cada una de las 1022 combinaciones sin repetición de todos los parámetros, y calculado el ARI resultante de compararlo con la distribución esperada para cada una de las características, el mayor ARI que obtenemos es 0.13. Únicamente se han obtenido 4 combinaciones de parámetros con ARI $> 0{,}1$. Es notable resaltar que estas combinaciones de parámetros tenían en común al parámetro q de ARIMA. Esto puede ser una señal que este parametro es un buen clasificador. Estas combinaciones presentaban $k = 2$ como número óptimo de clusters y la característica que se comparaba para el calculo de ARI era el contenido en intrones.

El resto de conjuntos de parámetros, no clasificaban ninguna característica con una puntuación ARI $> 0{,}04$. Estos valores ARI cercanos a 0 nos indican que el clustering no supervisado no presenta relación con los valores de las características esperadas.

El método de clustering no supervisado k-medias no parece un método adecuado para caracterizar las secuencias de nucleótidos a partir del uso de parámetros de series temporales y de grafos.

3.3. Caracterización de secuencias mediante K-Vecinos más próximos.

Al igual que en el caso del clustering no supervisado, hemos realizado este método de clustering para todas las combinaciones de parámetros (exceptuando el MPL) y todas las características.

En el Cuadro 17 podemos ver los resultados obtenidos de esta forma.

Característica	Porcentaje de acierto (en %)	Mejor combinación de parámetros para K-NN
Contenido en intrones	73.4	q_MA, q_ARIMA, d_ARIMA
Reino	50.7	Peng-α, q_MA, q_ARMA
Tipo AN	59.5	p_AR, d_ARIMA, q_ARIMA
Codificación del AN	91.1	Peng-α, VG, q_MA, p_ARIMA, q_ARIMA
Localización del AN	48.1	Peng-α, VG, q_ARMA, p_ARIMA, q_ARIMA

Cuadro 17: Resultados de clustering mediante K-NN. En la columna de la derecha podemos observar la combinación de parámetros que han obtenido el porcentaje de acierto máximo.

Podemos observar en el Cuadro 17 destaca la presencia en todas las características de al menos un parámetro de medias móviles (q). Destaca en concreto el del modelo ARIMA ya que se encuentra en 4 de 5 características. Es un buen clasificador para este método, al igual que pasaba con el método de k-medias no supervisado. Por regla general los métodos de clustering k-medias y K-NN son más precisas cuando utilizan parámetros clásicos, en vez de avanzados.

Hemos repetido estos resultados pero introduciendo las combinaciones con el parámetro MPL y reduciendo el número de secuencias. En el caso del contenido de intrones, el tipo de AN y la codificación del AN al introducir este parámetro el porcentaje de acierto aumentaba entre un 4-10 % pero en ninguna de las mejores combinaciones aparecía MPL, por tanto este aumento se debe a la disminución en el numero de secuencias. En el caso de las características restantes el porcentaje no aumentaba al introducir el nuevo parámetro. Por tanto podemos decir que a la hora de clasificar nuestras secuencias mediante este método, el parámetro MPL no mejora el resultado.

Se puede observar una mejoría considerable en la clasificación de todas las características al utilizar un algoritmo supervisado respecto al algoritmo no supervisado.

3.4. Caracterización de secuencias de nucleótidos mediante redes neuronales.

Hemos utilizado las redes neuronales (NN) presentadas en la Sección 2.7.3.

Los porcentajes de acierto de las diferentes redes neuronales podemos observarlos en el Cuadro 18.

Característica	NN(316) sin MPL	NN(209) con MPL	NN(209) sin MPL
Contenido en intrones	92.21	91.48	90.52
Reino	79.84	68.79	67.46
Tipo de AN	78.86	72.63	72.22
Codificación del AN	95.74	90.18	80.534
Localización del AN	77.11	72.15	70.19

Cuadro 18: Resultados de las diferentes redes neuronales para la clasificación de características. Los resultados están todos expresados en %. Entre paréntesis se encuentran el número de secuencias utilizadas para la Red Neuronal.

Los porcentajes de acierto de la red neuronal están todos por encima del 70 %. El uso de redes neuronales aumenta el porcentaje de acierto respecto al K-NN en todas las características.

La primera columna representa los resultados de la red neuronal construida con todas las secuencias y con 9 parámetros. No podemos introducir el décimo parámetro (MPL de nuestro trabajo ya que no solo se encuentra calculado en alguna de las 316 secuencias.

Para poder meter el parámetro MPL, tenemos que reducir el numero de secuencias a 209, todos los valores de porcentaje de acierto bajan entre 1-9 %.

Este descenso es provocado por la reducción del numero de secuencias principalmente ya que la tasa de acierto en la red neuronal con las 209 secuencias y sin el parámetro MPL no varía con respecto a cuando utilizamos el parámetro. Exceptuando el caso de la codificación del AN donde sí observamos una gran diferencia (en torno a 10 %) a la hora de clasificar.

CAPÍTULO 4. CONCLUSIONES.

1. El estudio de las series temporales asociadas a las secuencias de nucleótidos son herramientas útiles para el estudio de las secuencias de ácidos nucleicos.

2. Se valida el resultado del artículo que inspira principalmente este trabajo [8], que dice que las secuencias con intrones presentan correlación de largo alcance (Peng-$\alpha > 0.5$) y las secuencias sin intrones no presentan correlación (Peng-$\alpha = 0.5$).

3. El parámetro Peng-α es el parámetro que presenta mayor diferencias significativas entre las medias a la hora de comparar las secuencias con y sin intrones. Este parámetro también es el que presenta mayor diferencias significativas entre las medias a la hora de comparar las secuencias codificantes y estructurales.

4. En la clasificación de secuencias por reinos, el parámetro Peng-α separa bien las secuencias de reinos eucariotas y las secuencias del reino procariota. La distribución del parámetro MPL en Virus presenta diferencias significativas en la media comparadas con las medias de MPL en secuencias de todos el resto de los reinos.

5. En la clasificación de secuencias por tipo de ácido nucleico, no hay ninguno de los parámetros estudiados presenta medias significativas entre los grupos, a pesar de que el test ANOVA resulta estadísticamente significativo. El parámetro que mejor clasifica para este grupo es el MPL.

6. Para clasificar las secuencias por su localización en la célula, el parámetro que presenta diferencias más significativas en las medias es MPL.

7. Las series temporales asociadas a secuencias de nucleótidos presentan un comportamiento más parecido a procesos de medias móviles que procesos autorregresivos.

8. El método de clustering no supervisado k-medias no es un buen algoritmo de agrupamiento para separar las secuencias en ninguna de las características estudiadas en este trabajo.

9. El método de clustering supervisado k-Vecinos más próximos mejora los resultados de agrupamiento respecto al k-medias. Destacan en las combinaciones de parámetros que mejor clasifican las características con este método, los parámetros clásicos, especialmente el parámetro q de ARIMA.

10. Las redes neuronales son el mejor método de predicción de características a partir de los parámetros derivados de la serie temporal asociadas a secuencias de nucleótidos. Separa de forma correcta más del 90 % entre secuencias con y sin intrones y entre secuencias codificantes y estructurales. La red neuronal separa de forma correcta más del 75 % del resto de características.

11. Destacaría que todo el proceso del trabajo nos ha permitido avanzar hacia una herramienta basada en el calculo de diferentes herramientas matemáticas y el uso de algoritmos de clasificación de aprendizaje profunda. Una secuencia de nucleótidos puede ser caracterizada en los diferentes grupos de clasificación del trabajo con esta herramienta.

CAPÍTULO 5. BIBLIOGRAFÍA.

[1] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[2] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[3] National Center for Biotechnology Information Bethesda (MD): National Library of Medicine (US). Gene [internet], 2004. Disponible en Internet: https://www.ncbi.nlm.nih.gov/gene/, consultado por última vez el 18/06/2021.

[4] Mehdi Kchouk, Jean-Francois Gibrat, and Mourad Elloumi. Generations of sequencing technologies: from first to next generation. *Biology and Medicine*, 9(3), 2017.

[5] Graham Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature methods*, 11(3):294–296, 2014.

[6] Marcello Buiatti, C Acquisti, G Mersi, and P Bogani. The biological meanings of dna correlations. In *Fractals in Biology and Medicine*, pages 235–245. Springer, 2002.

[7] Daniel Peña. *Análisis de series temporales, Alianza Editorial, Madrid.* 2005.

[8] Chung-Kang Peng, Sergej V Buldyrev, Ary L Goldberger, Shlomo Havlin, Francesco Sciortino, Michael Simons, and H Eugene Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168–170, 1992.

[9] Nidhal Bouaynaya and Dan Schonfeld. Nonstationary analysis of coding and noncoding regions in nucleotide sequences. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):357–364, 2008.

[10] Lucas Lacasa, Bartolo Luque, Fernando Ballesteros, Jordi Luque, and Juan Carlos Nuno. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975, 2008.

[11] Lynn Sagan. On the origin of mitosing cells. *Journal of Theoretical Biology*, 14(3):225–IN6, 1967.

[12] National Center for Biotechnology Information Bethesda (MD): National Library of Medicine (US). Nucleotide [internet], 1988. Disponible en Internet: https://www.ncbi.nlm.nih.gov/nuccore/, consultado por última vez el 18/06/2021.

[13] National Center for Biotechnology Information Bethesda (MD): National Library of Medicine (US). National center for biotechnology information (ncbi)[internet], 1988. Disponible en Internet: https://www.ncbi.nlm.nih.gov/, consultado por última vez el 23/06/2021.

[14] Walter Gilbert. Why genes in pieces? *Nature*, 271(5645):501–501, 1978.

[15] Robert H Whittaker. New concepts of kingdoms of organisms. *Science*, 163(3863):150–160, 1969.

[16] Yaqin Cai, Xiaocan Lei, Zhuo Chen, and Zhongcheng Mo. The roles of cirrna in the development of germ cells. *Acta Histochemica*, 122(3):151506, 2020.

[17] Subhash C Verma, Zhong Qian, and Sankar L Adhya. Architecture of the escherichia coli nucleoid. *PLoS Genetics*, 15(12):e1008456, 2019.

[18] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[19] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

[20] Scott Beamer, Krste Asanovic, and David Patterson. Direction-optimizing breadth-first search. In *SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pages 1–10. IEEE, 2012.

[21] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486, 2012.

[22] Howard Levene. Robust tests for equality of variances. in 'contributions to probability and statistics: essays in honor of harold hotelling'.(eds i olkin, sg ghurye, w hoeffding, wg madow, hb mann) pp. 278–292, 1960.

[23] Tae Kyun Kim. T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6):540, 2015.

[24] Manuel Terrádez and Angel A Juan. Análisis de la varianza (anova). *Catalunya: Universidad Oberta de Catalunya*, 2003.

[25] Maksim A Terpilowski. scikit-posthocs: pairwise multiple comparison tests in python. *Journal of Open Source Software*, 4(36):1169, 2019.

[26] Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386, 2004.

[27] Jure Zupan. Introduction to artificial neural network (ann) methods: what they are and how to use them. *Acta Chimica Slovenica*, 41:327–327, 1994.

[28] Allen Chieng Hoon Choong and Nung Kion Lee. Evaluation of convolutionary neural networks modeling of dna sequences using ordinal versus one-hot encoding method. In *2017 International Conference on Computer and Drone Applications (IConDA)*, pages 60–65. IEEE, 2017.

[29] Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

[30] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[32] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[33] Rob J Hyndman, Yeasmin Khandakar, et al. Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software*, 27(3):1–22, 2008.

[34] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[37] Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith. Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39, 2010.