UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

GRADO EN BIOTECNOLOGÍA

DEPARTAMENTO DE MATEMÁTICA APLICADA

Herramientas bioinformáticas para el estudio del ADN: caracterización de genes, estudio filogenético y detección de secuencias funcionales

TRABAJO FIN DE GRADO

Autor: Roberto Molina Ballesteros

Tutor: Carlos García-Gutiérrez Báez

Junio de 2023

UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Técnica Superior De

Ingeniería Agronómica, Alimentaria y de Biosistemas

GRADO DE BIOTECNOLOGÍA

HERRAMIENTAS BIOINFORMÁTICAS PARA EL ESTUDIO DEL ADN: CARACTERIZACIÓN DE GENES, ESTUDIO FILOGENÉTICO Y DETECCIÓN DE SECUENCIAS FUNCIONALES.

TRABAJO FIN DE GRADO

Roberto Molina Ballesteros

MADRID, JUNIO 2023

Tutor: Carlos García-Gutiérrez Báez

Dpto. de Matemática Aplicada, UPM

TITULO DEL TFG- HERRAMIENTAS BIOINFORMÁTICAS PARA EL ESTUDIO DEL ADN: CARACTERIZACIÓN DE GENES, ESTUDIO FILOGENÉTICO Y DETECCIÓN DE SECUENCIAS FUNCIONALES

Memoria presentada por Roberto Molina Ballesteros para la obtención del título de Graduado en Biotecnología por la Universidad Politécnica de Madrid

Fdo: Roberto Molina Ballesteros

VºBº Tutor

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

Madrid, 26 de junio de 2023

Agradecimientos

A mis padres, puesto que son quienes me han dado la confianza y libertad para llegar a ser quien soy. Aunque me voy lejos, creo que son buenos pasos.

A mi *pandilla*, qué aunque haya sido el último en llegar, me han hecho sentir que siempre he estado. Solo siento no poder hacer llegar mejor lo que habéis sido para mí.

A mis amigos, los que habéis visto y aguantado a más *Robertos*. Conocerme hoy es conocer un poco de todos vosotros.

A Sofía, por pasar de odiarme ser a quién me ha tenido que aguantar todo este proceso. No creo que exista mejor compañía hasta las 6 de la mañana.

Y a Carlos, que ha sido quien se ha encargado de tejer todo esto. Gracias por darme libertad para nadar, pero también mantener mis pies en el suelo.

Índice

Índice de cuadros

Índice de figuras

Lista de abreviaturas

A: Adenina.

ADN: Ácido Ribonucleico.

ARN: Ácido Desoxirribonucleico.

ARNm: ARN mensajero.

ARNr: ARN ribosómico.

C: Citosina.

$\mathbf{C95(n)}$: Cuantil 95 de D_{KL}.

$\mathbf{C_{iso}}$: Coeficiente isoperimétrico.

$\mathbf{D_0}$: Dimensión generalizada para q=0.

$\mathbf{D_1}$: Dimensión generalizada para q=1.

$\mathbf{D_{-80,80}}$: Rango del espectro multifractal calculado entre q=-80 y q=80.

DDBJ: (*DNA Data Bank of Japan*). Banco de ADN de Japón[1].

DGNE: Distancia Geodésica No Enraizada.

$\mathbf{d_H}$: Distancia de Hausdorff.

$\mathbf{dim_{box}(S)}$: Dimensión *box counting*.

$\mathbf{D_{KL}}$: Divergencia de Kullback-Leibler.

$\mathbf{d_S}$: Función Mínima Distancia Media.

$\mathbf{d_{Sm}}$: Función Mínima Distancia Media Media.

$\mathbf{d_{SM}}$: Función Mínima Distancia Media Máxima.

EMBL: (*European Molecular Biology Laboratory*). Laboratorio Europeo de Biología Molecular[2].

G: Guanina.

H: Entropía de Shannon.

K2: Modelo Kimura 2-parámetros.

NCBI: (*National Center for Biotechnology Information*). Centro Nacional para la Información Biotecnológica[3].

$\mathbf{Max(n)}$: Máximo de D_{KL}.

$\mathbf{P}$: Distribución de nucleótidos de una secuencia de ADN.

$\mathbf{P_{chr}}$: Distribución del cromosoma.

$\mathbf{P_{iso}}$: Distribución del *isochore*.

$\mathbf{P_{obs}}$:Distribución observada.

$\mathbf{P_S}$: Distribución simulada.

RF: Robinson-Foulds.

RFG: Robinson-Foulds generalizado.

T: Timina.

T92: Modelo Tamura 3-parámetros.

TN93: Modelo Tamura y Nei.

$\mathbf{V(n)}$: Varianza de D_{KL}.

Summary

In this study, bioinformatics tools for DNA sequence analysis were developed. The main objective was to explore existing tools and propose new uses for them.

The starting point of this work was the Nandy representation, which transforms a nucleotide chain into a set of points in the plane, according to the order of the bases that compose it. This representation was devised in order to highlight the local relative abundance of nucleotides and to differentiate between regions with large variations.

This work has explored two new applications: the geometric and multifractal characterization of the representations and their application to the study of phylogeny. The parameters obtained from the characterization have been related to different characteristics of the sequences, for example, the presence of introns or the kingdom to which they belong. Also, methods have been proposed to estimate phylogenetic relationships between individuals through geometric comparison of DNA sequences of homologous genes. This was attempted in other studies, but new approaches have been taken.

For the localization of non-coding sequences with functionality in long DNA strands, a probabilistic approach is proposed. The starting hypothesis is that, in vertebrates, the distribution of nucleotides in zones which are not under selection pressure, have constant relative base frequencies.

For the development of this project, a database of 10,400 DNA sequences from different organisms, used both in sequence characterization and phylogenetic study, was prepared.

The results obtained show the potential of these tools, both in the field of phylogeny and in the localization of functional sequences, which can complement the existing ones.

Capítulo 1. Introducción y objetivos

La cantidad de secuencias disponibles en internet es enorme y se encuentra en constante crecimiento, repartida en numerosas bases de datos de libre acceso como NCBI[3], EMBL[2] o DDBJ[1]. Tan solo NCBI cuenta con más de 240 millones de secuencias, a abril de 2023[4], y las técnicas de secuenciación masiva han acelerado el crecimiento del número de estas.

Las herramientas de análisis de ADN *in silico* son programas y algoritmos que permiten realizar análisis exhaustivos del ADN sin la necesidad de llevar a cabo más experimentos de laboratorio, los cuales resultan costosos y laboriosos. Aunque ya existen un gran número de herramientas de este tipo, es necesario seguir desarrollando nuevas formas de estudiar el ADN. Una alternativa a la invención de herramientas es intentar dar nuevas utilidades a otras ya existentes.

La finalidad de este trabajo es encontrar nuevos métodos para estudiar el ADN, proponiendo ideas y empleando técnicas ya existentes pero con otros enfoques.

1.1. Distintas aplicaciones para la representación de Nandy de 1994

En su trabajo de 1994[5] Nandy propuso un método para representar cadenas de ADN como una sucesión de puntos en el plano según la secuencia de nucleótidos **(Figura 1)**.

Esta representación entra dentro de lo que se denominan métodos de paseo bidimensional[6], pero la peculiaridad de la propuesta de Nandy es que coloca en el eje horizontal las bases púricas, y en el vertical, a las pirimidínicas.

Esta herramienta ha sido empleada con diversos fines, como estudiar la relación entre especies a través de sus genes[6], y ha inspirado nuevas formas de representar el ADN, como la representación de Huffman[7] o la representación de Yau[8].

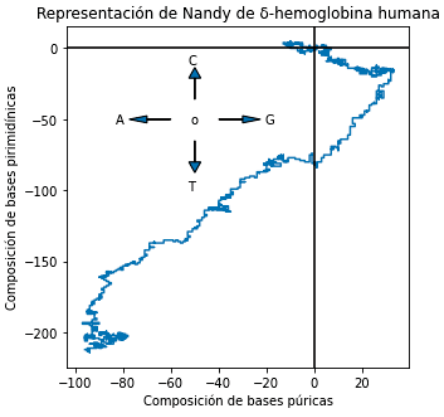


Figura 1: Representación de Nandy de la secuencia de DNA de la subunidad delta de la hemoglobina humana (NCBI *Reference Sequence*: NG_063112.2). Se ha empleado el paquete `Matplotlib`.

En este trabajo se han explorado aplicaciones alternativas para la representación de Nandy, con el fin de encontrar nuevas herramientas bioinformáticas para el análisis de secuencia de ADN.

1.1.1. Caracterización de representaciones de secuencias de ADN

Si se tiene en cuenta la cantidad de veces que un punto es visitado durante la representación de Nandy, se obtiene una distribución. De este modo se hace posible caracterizarlas tanto geométricamente como distribucionalmente.

Considerando la representación como un conjunto geométrico, se pueden explorar características como su circularidad y su dimensión fractal.

El conjunto de puntos obtenido puede ser encerrado por su correspondiente envoltura convexa **(Figura 2)**. Así, a la representación de cada secuencia de ADN le corresponde un único polígono convexo. Según la distribución de nucleótidos, el polígono puede parecerse más o menos a una circunferencia, lo que se entiende como circularidad.

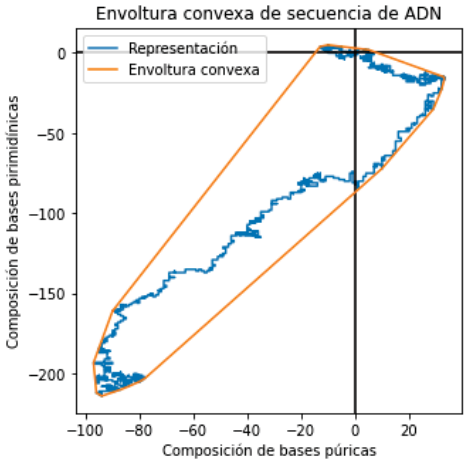


Figura 2: En azul: Representación de Nandy de la secuencia de DNA de la subunidad delta de la hemoglobina humana (*NCBI Reference Sequence*: NG_063112.2). En naranja: la envoltura convexa correspondiente a la secuencia

Los fractales son objetos pertenecientes al campo de la geometría que presentan irregularidades a cualquier escala[9]. Estos conjuntos a menudo describen fenómenos de la naturaleza de forma más precisa que las figuras pertenecientes a la geometría clásica[9]. Se caracterizan por presentar autosimilitud a distintas escalas, que en el caso de las representaciones de Nandy, es de tipo estadística **(Figura 3)**.

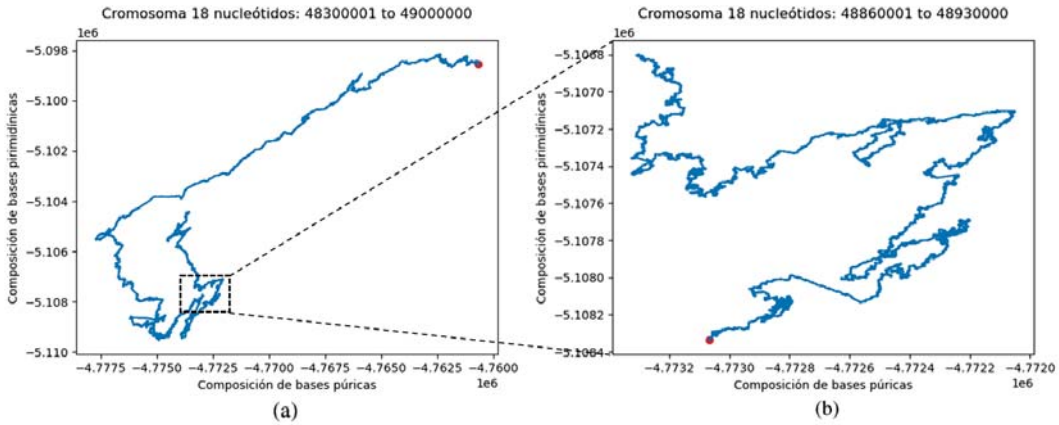


Figura 3: (a) En azul: representación de Nandy de un fragmento de 700 mil nucleótidos del cromosoma 18 humano (NCBI *Reference Sequence*: NC_000018.10), desde el nucleótido 48300001 al 49000000. En rojo: el primer nucleótido. (b) En azul, la representación de 7 mil nucleótidos pertenecientes a la imagen (a), que van desde el 48860001 al 48930000. En rojo: el primer nucleótido de la secuencia.

Mandelbrot, en su artículo "*How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension*"[10], fue quien describió por primera vez estos objetos. Del mismo modo que la costa de Inglaterra, las representaciones de Nandy mantienen autosemejanza estadística en un rango de escalas. Los fractales son caracterizados a través de la dimensión fractal.

Como se ha explicado más arriba en este apartado, es posible obtener una distribución a partir de la representación de Nandy **(Figura 4)**. De esta manera, las representaciones pueden estudiarse como multifractales. Mientras que los fractales se refieren a un conjunto, los multifractales lo hacen a una medida[11]. Además, mientras que los fractales clásicos tienen una dimensión fractal única, las distribuciones multifractales se caracterizan por tener un espectro de dimensiones fractales asociadas a diferentes subconjuntos de su soporte. La teoría de los multifractales se aplica en diversos campos, como la física de los sistemas complejos, la geofísica, la meteorología o la teoría del caos[12]. Ayuda a comprender y modelar fenómenos que exhiben heterogeneidad y variabilidad a diferentes escalas.

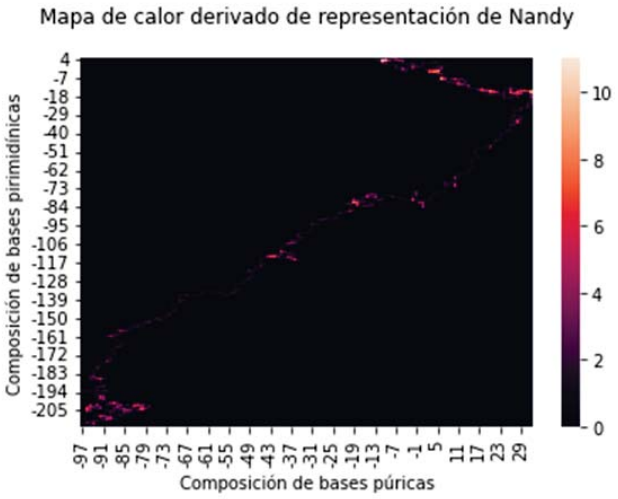


Figura 4: Mapa de calor derivado representación de Nandy de la secuencia de DNA de la subunidad delta de la hemoglobina humana (NCBI *Reference Sequence*: NG_063112.2[3]. El gráfico se realizó mediante el paquete `Seaborn`.

1.1.2. La filogenia

La filogenia es una rama de la biología encargada del estudio de las relaciones evolutivas entre diferentes individuos o especies, basándose en la premisa de que todos los organismos de la Tierra tienen un antecedente común[13]. Dentro de los distintos enfoques que se toman para su estudio, la filogenia molecular es el más extendido y validado [14]. Para establecer relaciones de parentesco entre especies, se comparan varias secuencias de un mismo elemento que se encuentre conservado entre los individuos, como puede ser un gen. Principalmente suele hacer uso de las distancias o similitudes entre secuencias de ADN, de ARN o de proteínas.

Los métodos basados en distancias dependen de una fase de alineamiento en la que las secuencias que se acompasan, con el fin de resaltar las zonas de similitud. El alineamiento múltiple de secuencias requiere de métodos heurísticos dada su complejidad[15]. En función del grado de similitud de las secuencias, se determina una distancia genética, que se emplea para crear árboles filogenéticos.

La representación de Nandy ofrece una nueva manera de abordar este problema, a partir de la similitud geométrica de las representaciones de las secuencias de ADN **(Figura 5)**. Estos métodos pueden complementar a las técnicas moleculares basadas en distancias genéticas, descritas más arriba.

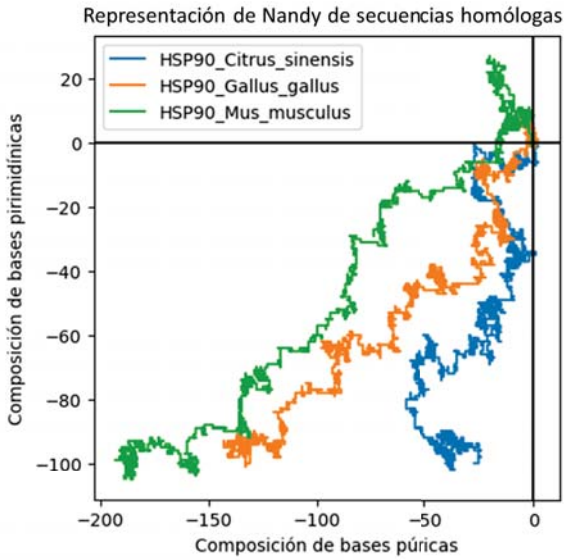


Figura 5: Representación de Nandy conjunta para 3 secuencias de genes homólogos. En azul: gen correspondiente a *Citrus sinensis* (NM_001288915.1). En naranja: gen de *Homo sapiens* (NM_001017963.3). En verde: gen de *Mus musculus* (lcl|NM_010480.5_cds_NP_034610.1_1).

El objetivo de esta parte del TFG es encontrar 2 nuevas aplicaciones a la representación de Nandy:

- **Objetivo 1:** La caracterización geométrica y multifractal de las representaciones de Nandy de secuencias de genes. Se ha medido su circularidad y dimensión fractal; así como su *espectro de Rényi*, explicado en el **Apartado 2.2.1**. Se estudió si estos parámetros dependen de factores biológicos como el grupo filogenético, la presencia de intrones o la proteína para la que codifica.

- **Objetivo 2:** La creación de árboles filogenéticos mediante la comparación entre los conjuntos de puntos obtenidos a través de la representación de Nandy. Se han estudiado 3 funciones distintas de similitud entre secuencias, comparándolas con los árboles que se obtienen tratando los mismos datos con técnicas basadas en distancias genéticas y se han evaluado sus limitaciones.

1.2. Localización de secuencias funcionales

La detección de secuencias no codificantes con función es uno de los principales retos de la biología hoy en día[16], y se han desarrollado procedimientos para ello, como la Inmunoprecipitación de la Cromatina acoplada a Secuenciación[17] (CHIP-seq). Sin embargo, estos métodos son eficaces para detectar algunos elementos reguladores, como potenciadores o *enhancers*, pero existen otros más difíciles de hallar, o incluso puede que existan mecanismos cuya naturaleza es todavía desconocida. Por ello se necesita seguir explorando nuevas aproximaciones a este problema.

El ADN de los organismos eucariotas se divide principalmente en dos tipos: ADN codificante (los exones) y ADN no codificante, dividido en intrones y zonas intergénicas[18]. Los exones portan el contenido genético para las proteínas, su información es transcrita a ARNm, quien posteriormente es traducido a una secuencia de aminoácidos. Aunque los intrones no codifiquen para proteínas, tienen funciones reguladoras esenciales para la propia síntesis proteica. Los intrones también son transcritos junto a los exones, pero son eliminados durante maduración del ARNm por un mecanismo denominado *splicing*, gracias a secuencias que se encuentran en los propios intrones. Este proceso puede ocurrir de distintos modos, de manera que de un mismo ARNm se obtienen distintas variantes de un gen, lo que se denomina *splicing* alternativo. Esta función y otras como la regulación de la propia transcripción del gen[19], la degradación de ciertos ARNm, o la exportación desde el núcleo[20], se encuentran reguladas por el propio intrón.

En cuanto al ADN intergénico, en un comienzo, como no codifica proteínas y no se conocía su función, fue denominado "ADN basura"[16]. Con el tiempo, la finalidad de algunas de las regiones de este ADN basura ha sido descubierta, como la presencia de secuencias reguladoras de la transcripción, ARN reguladores de cadena larga o pseudogenes[19]. Gran cantidad del ADN intergénico está compuesto por secuencias cortas repetidas, como las presentes en los telómeros o los microsatélites, o por secuencias que realmente no tienen función. Por lo tanto, el reto consiste en distinguir el ADN intergénico funcional del que realmente podemos denominar como "ADN basura"[16].

Los organismos empaquetan su genoma en una o varias moléculas de ADN, conocidas como cromosomas. Algunos autores definen los cromosomas de los vertebrados como mosaicos de *isochores*[21]. Estos son segmentos de gran longitud, mayor a 3×10^5 nucleótidos, con un contenido de GC relativamente homogéneo. Se relacionan con las bandas de los cromosomas y muestran que la heterogeneidad en la distribución de los nucleótidos varía según la escala con la que se mire. Si bien la teoría de los *isochores* ha perdido fuerza, puesto que las familias de *isochores* se solapan, lo que vuelve difícil su

clasificación[22], transmiten la idea de que, a gran escala, se tiende a la homogeneidad en la distribución de nucleótidos. Los exones que se encuentran en estas regiones pueden tener una distribución de nucleótidos distinta a la del *isochore*, pero su efecto sobre el global de la distribución es despreciable debido a su relación de tamaño.

El ADN está compuesto por las bases ACTG y en principio no hay razón para esperar una mayor abundancia de AT que CG en el genoma de un organismo. Sin embargo, ya desde el descubrimiento de la Ley de Chargaff se observó que es esto no ocurre así[23]. Este desequilibrio en la frecuencia relativa de las bases es causado por múltiples razones. Por un lado, en bacterias existe un sesgo en el proceso de mutación que lleva a un mayor contenido de AT, y hay evidencia de que esto ocurre en todos los clados de procariotas[24]. Por otro lado, existe un sesgo que lleva a un mayor aumento de CG en mamíferos relacionado con la recombinación[25]. También, las condiciones físicas en las que se desarrolle un organismo influyen, por ejemplo, algunas arqueas y bacterias extremófilas que viven a altas temperaturas poseen un mayor contenido en GC[26]. Esto es debido a que estas dos bases son capaces de establecer 3 puentes de hidrógeno entre ellas, a diferencia de AT, que solo pueden 2, así son capaces de mantener el ADN naturalizado a mayor temperatura[27]. Por último, entre otras causas, los procesos de selección tienen un papel importante en la composición de nucleótidos. Si la presencia de una base aumenta el *fitness*, la selección natural tenderá a conservar el cambio, y, por el contrario, si su presencia lo perjudica, será eliminado. Aunque, a priori, no existen bases que proporcionen mayor *fitness* que otras, se ha observado que casi la mitad de los exones se componen de secuencias con bajo contenido de GC[18]. En cambio, si una posición en el genoma no está sometida a una presión de selección, ocurre la deriva genética: con el tiempo el nucleótido que ocupe su lugar puede ser cualquiera, y esto está influido por los sesgos en los procesos de mutación[28]. Si esto ocurre sobre un conjunto de posiciones consecutivas, sus nucleótidos tenderán a distribuirse aleatoriamente a medida que transcurren las generaciones desde que la presión de selección dejó de actuar sobre esa zona. Hasta ahora no existe evidencia de fuertes variaciones en las ratios de mutación para segmentos de ADN mayores a 100 pares de bases en zonas no sujetas a una presión de selección[25].

El tercer objetivo de este TFG fue el estudio de la distribución de nucleótidos de zonas funcionales en zonas intergénicas, denominado como **objetivo 3**. La hipótesis de la que se partió es que, si una zona del cromosoma no se encuentra sometida a una presión de selección, su distribución de nucleótidos tenderá a la del *isochore* en el que se halla. Por el contrario, si una zona del cromosoma sí está sometida a una presión de selección, por tanto sí es funcional, su distribución podrá ser diferente o no.

La representación de Nandy refleja la distribución de nucleótidos de una secuencia, por lo que puede convertirse en una herramienta útil para visualizar esta idea. Como se observa en la **Figura 6**, el cromosoma mostrado tiene una distribución de nucleótidos homogénea a gran escala, ya que la representación parece una línea recta. Al ampliar la representación, aparecen zonas que muestran una distribución similar a la del cromosoma, mientras que otras no.

Encontrar una herramienta basada en la geometría de las representaciones puede ser complicado, por lo que se ha optado por un enfoque probabilístico.

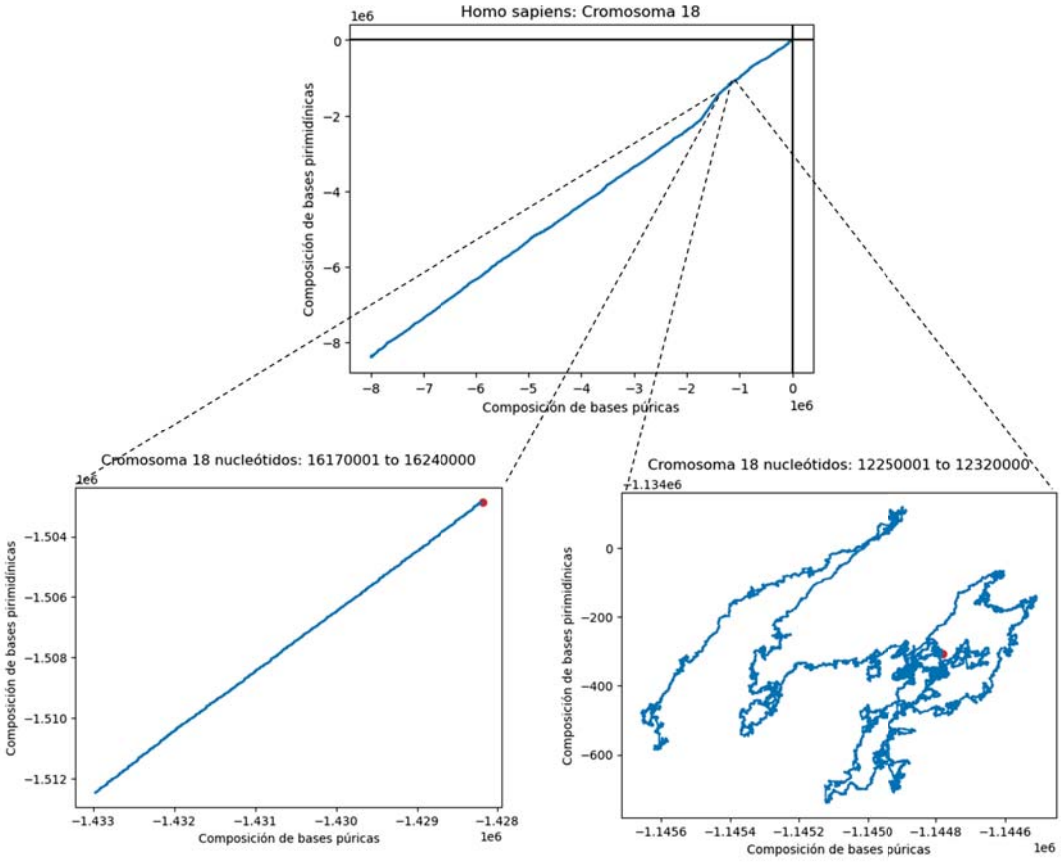


Figura 6: Arriba, representación de Nandy del cromosoma 18 de *Homo sapiens* (NC_000018.10). Abajo izquierda. En azul, representación de un fragmento de 70 mil bases correspondientes a un segmento del centrómero. En rojo, el primer nucleótido.Abajo izquierda. En azul, representación de un fragmento de 70 mil bases correspondiente a una zona con genes. En rojo, el primer nucleótido.

Capítulo 2. Materiales y métodos

2.1. Base de datos

Para los **objetivos 1 y 2** del TFG, se requirió de un gran número de secuencias. Se descargaron 10400 secuencias de la sección de genes de NCBI[3]. Se clasificaron las secuencias según diversos criterios, expuestos más abajo. Con el fin obtener un número tan alto de secuencias y poder organizarlas, se diseñaron programas en `Python` que automatizasen el proceso.

Para el **objetivo 1**, se tomaron secuencias homólogas de genes presentes en diferentes grupos filogenéticos:

- Secuencias que codifican para el **citocromo B** (CYTB), un componente esencial de la cadena de transporte de electrones, que desempeña un papel clave en la producción de energía en las células. Se encuentra presente en una amplia variedad de organismos, incluyendo procariotas y eucariotas, razón por la que es ampliamente empleado en filogenia[29].

- Secuencias que codifican para la proteína **P53**. Se encuentra en animales y se encarga de la detención del ciclo celular ante la detección de una mutación, muy relacionada con la aparición de tumores[30]. Es una proteína muy conservada en este grupo y, aunque se han encontrado proteínas en plantas con un origen común, solo se han tenido en consideración secuencias de animales[31].

- El ARN ribosomal (ARNr) se encuentra altamente conservado y se ha empleado durante mucho tiempo en la filogenia molecular de organismos eucariotas y procariotas, desde bacterias hasta animales y plantas[32]. Dentro de este grupo destaca el ARN ribosomal 16S (**ARNr 16S**), un componente de la subunidad menor del ribosoma. En eucariotas, la función del ARNr 16S ha sido sutituida por el ARNr 18S, sin embargo, el gen aún se conserva en la mitocondria y ha sido empleado en diversos estudios[33]. El grupo de secuencias tomadas contiene genes de los siguientes dominios y reinos: animales, arqueas, plantas, bacterias y protistas.

Las secuencias genómicas de *citocromo B* y *p53* formarán el grupo al que se le ha dado el nombre de **Grupo Genes**. Estos se compararán, puesto que se tratan de dos genes de distinta naturaleza: el *citocromo B* se encuentra tanto en procariotas como eucariotas, mientras que las secuencias de *p53* pertenecen únicamente al reino *Animalia*. Las secuencias de los 5 dominios y reinos de ARNr 16S,

formarán un grupo llamado **Grupo ARNr**. Su interés reside en que son secuencias homólogas de grupos taxonómicos muy diferentes. Se podría haber tomado otro gen conservado entre reinos y dominios, como el propio *citocromo B*, pero a diferencia de este, se trata de un gen que no codifica para proteínas, por lo que su evolución no se encuentra condicionada por el código genético. Además, el ARNr 16S de los eucariotas se encuentra en la mitocondria, por lo que el proceso de endosimbiosis que dio lugar a estos orgánulos[34] ha podido tener un impacto en las características que se van a estudiar.

Se tomaron también grupos de secuencias compuestas por genes cuya característica en común es su pertenencia a un reino o dominio:

- Se descargó un conglomerado de secuencias genómicas codificantes de genes de los siguientes dominios y reinos: animales, plantas y bacterias. El conjunto está formado por secuencias de ADN genómico, acotadas en un tamaño de secuencia entre 800 y 1300 bases, por las razones expuestas en el **Apartado 2.2.1**. Este grupo se denominó como **Grupo Conglomerado** y no comparte secuencias con el **Grupo ARNr 16S**, puesto que solo contiene secuencias codificantes para proteínas.

- Se obtuvo un conglomerado de secuencias de ARNm de animales, en un rango de 800 a 1300 bases. A este grupo se le denominó como **Grupo ARNm**.

Los tamaños de las secuencias descritas anteriormente se muestran en los **Cuadros suplementarios A.1, A.2, A.3 y A.4**.

A excepción de las secuencias de bacterias, arqueas y ARNm de animales, todas las secuencias pueden contener intrones. El objetivo de obtener este conjunto de secuencias es formar grupos homogéneos organizados por grupo taxonómico, naturaleza del ácido nucleico (genómica, ARNm) o tipo de gen, para poder comparar los parámetros obtenidos a partir de la representación de Nandy.

Para el **objetivo 2** se emplearon 10 secuencias para cada uno de los siguientes 5 genes homólogos **(Cuadro suplementario A.5)** comúnmente empleados en filogenia:

- La **HSP90** (proteína de choque térmico 90), una proteína molecular de chaperona altamente conservada en todos los organismos eucariotas[32]. Esta proteína pertenece a la familia de HSP, que juega un papel crucial en la protección de las células contra el estrés y las perturbaciones ambientales.

- El **citocromo C**[35], una proteína hemoproteica que se encuentra en la mitocondria de las células eucariotas. Es un componente esencial del sistema de transporte de electrones que se encarga de

la generación de ATP durante la respiración celular. La secuencia de aminoácidos del citocromo C es altamente conservada en la mayoría de los organismos, lo que indica su importancia evolutiva y funcional.

- El ARN ribosomal 18S (**ARNr 18S**) se trata del gen eucariota análogo al ARN ribosomal 16S de procariotas. Ambos son componentes de la subunidad pequeña ribosomal, cumpliendo la misma función, y es muy empleado en filogenia[32].

- La enzima gliceraldehído-3-fosfato deshidrogenasa (**GAPDH**) es una proteína involucrada en el metabolismo de la glucosa en la célula, que evidencia algunos de los acontecimientos evolutivos más importantes de la historia[36].

- La histona **H3** es la histona más empleada en filogenia[37], ya que está codificada por un gen altamente conservado en los eucariotas, debido a su importancia en la compactación del DNA. Se encarga de mantener y regular la estructura de la cromatina y puede estar sujeta a modificaciones epigenéticas que regulen su comportamiento.

2.2. La representación de Nandy

En el trabajo de Nandy de 1994[5] se propone un método de representación de secuencias de ADN en el plano, a través del siguiente procedimiento: i) se parte del origen de coordenadas y el primer nucleótido de la secuencia, ii) se lee una base de la secuencia y se mueve una unidad en una de las direcciones de los ejes en función del nucleótido, y iii) se avanza a la siguiente base y se vuelve al paso ii), hasta que se hayan leído todas las posiciones de la cadena. Dependiendo de la base, el movimiento será: adenina, a la izquierda; guanina, a la derecha; timina, hacia abajo, y citosina, hacia arriba.

Durante la representación, es posible pasar por algunos puntos múltiples veces. Si se almacena esta información, se obtiene una distribución en el plano **(Figura 4)**. Se define como masa $\mu(P)$ de un punto P del plano a la cantidad de veces que pasamos por dicho punto dividido entre el total de nucleótidos de la secuencia.

2.2.1. Caracterización geométrica

Envoltura convexa y circularidad La envoltura convexa se entiende como el polígono convexo más pequeño posible que envuelva un conjunto[38] **(Figura 2)**. Como se explica en el **Apartado 1.1.1**,

para determinar la circularidad de las representaciones de Nandy de las secuencias se optó por emplear el coeficiente isoperimétrico[39] de su envoltura convexa, que se define como:

$$C_{iso} = \frac{4\pi A}{L^2} \, ,$$

donde L es el perímetro de la envoltura convexa y A el área que encierra. C_{iso} mide cuánto se parece un polígono a un círculo. Toma valores en el intervalo (0,1], valdrá 1 cuando se trate de una circunferencia[40]. Las envolturas con una forma más alejada de la circunferencia tendrán valores de C_{iso} más bajos.

Para calcular la envoltura geométrica del conjunto de puntos, se empleó la función `ConvexHull` del paquete `Scipy` en `Python`. Esta función proporciona el área y los vértices de la envoltura, pero no el perímetro. Para obtener este último, se creó una función propia que lo calculase a partir de los vértices que definen la envoltura.

Dimensión fractal Como se explica en el **Apartado 1.1.1**, la complejidad de los fractales se caracteriza a través de la dimensión fractal. Para estimar la dimensión fractal de un conjunto S, se ha recurrido a la dimensión de Minkowski–Bouligand, también denominada *box-counting*[41], que se define como:

$$dim_{box}(S) = \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log (1/\varepsilon)} \, ,$$

donde ε es el tamaño de la caja empleado, y $N(\varepsilon)$ el número de cajas necesarias para cubrir todo el conjunto dado ese tamaño de caja. Así, esta dimensión mide como escala $N(\varepsilon)$ con el tamaño ε cuando este tiende a 0. Las representaciones de Nandy se convirtieron en imágenes, y se les añadió un orlado para que su lado fuese potencia de 2. Se empleó la partición diádica, en la que se realizan divisiones sucesivas de tamaño $\varepsilon = 2^{-k}L$, donde L es el tamaño de la imagen y $k = 0, 1, 2, 3 \ldots$ hasta que el tamaño ε alcanzase el valor de 1. Para cada tamaño ε, se obtiene un total de 2^k cajas. Para cada ε, se contó el número $N(\varepsilon)$ de cajas que interseccionaban con la representación de Nandy. Al representar el logaritmo del inverso de ε frente al logaritmo de $N(\varepsilon)$, se obtiene una gráfica como la **Figura B.1**. Se observa que la función es una recta en un intervalo de tamaños, donde la pendiente es $dim_{box}(S)$. Se han descartado los primeros tamaños para el cálculo, puesto que el orlado añadido condiciona el escalado en las primeras etapas.

2.2.2. Análisis multifractal

Al conservar la información de cuántas veces se pasa por un punto en la representación de Nandy, se obtiene una distribución en el plano. Como se explica más arriba en este mismo apartado, se denomina masa del punto $\mu(P)$ a la cantidad de veces que se ha pasado por P dividido entre el número total de nucleótidos de la secuencia. Se obtiene de esta manera una distribución que se puede caracterizar mediante el análisis multifractal. Este estudio se realizó a través de las *dimensiones generalizadas*, D_q, o *espectro de Rényi*[42]:

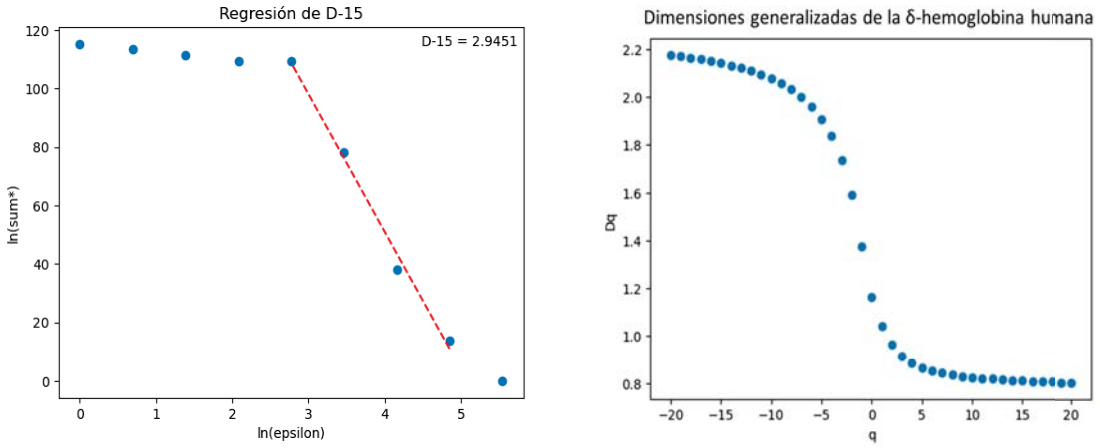
$$D_q \approx \frac{1}{q-1}\frac{\log \sum_{i=1}^{n(\varepsilon)} \mu_i(\varepsilon)^q}{\log \varepsilon} \quad \text{para} \quad q \neq 1, \quad \text{y}$$

$$D_1 \approx \frac{\sum_{i=1}^{n(\varepsilon)} \mu_i(\varepsilon) \log \mu_i(\varepsilon)}{\log \varepsilon}.$$

El método emplea particiones diádicas de la imagen de tamaños $\varepsilon = 2^{-k}L$, para $k = 0, 1, 2, \ldots$. Para cada k se obtuvieron un total de $n(\varepsilon) = 2^k$ cajas, donde la masa total de la caja i-ésima se denota como $\mu_i(\varepsilon)$. Las *dimensiones generalizadas* estudian las potencias q-ésimas de las masas $\mu_i(\varepsilon)$, con el tamaño de la partición ε, para un conjunto arbitrario de valores de q.

Es un cálculo análogo al de la dimensión *box-counting*. Se seleccionó un rango de q comprendido entre -20 y 20, con intervalos de 1. Se representó $\log \sum_{i=1}^{n(\varepsilon)} \mu_i(\varepsilon)^q$ frente a $\log \varepsilon$ para cada q, obteniendo que D_q es la pendiente de la gráfica dividida entre $(q-1)$. Para $q = 1$, se representa $\sum_{i=1}^{n(\varepsilon)} \mu_i(\varepsilon) \log \mu_i(\varepsilon)$ frente a $\log \varepsilon$, y la pendiente de la gráfica es D_1. Las distribuciones solo se comportan como multifractales en un rango de escalas, por lo que debe identificarse en el cálculo de cada D_q qué tamaños ε tomar **(Figura 7a)**. Esto se realizó con un programa propio que elimina puntos de los extremos de la regresión en función del grado del ajuste lineal.

Al representar (q, D_q), se obtiene una curva sigmoidal decreciente **(Figura 7b)** en caso de estar ante una distribución multifractal[43]. Cada valor de D_q contiene información sobre la distribución, siendo D_0 la dimensión fractal de Minkowski–Bouligand del soporte, D_1 la dimensión de la información o D_2 la dimensión de correlación.



(a) (b)

Figura 7: (a) Regresión lineal de D_{-15}, empleando los puntos donde se mantiene la linealidad. Sum* es $\log \sum_{i=1}^{n(\varepsilon)} \mu_i(\varepsilon)^q$ y D_{-15} es la pendiente de la recta dividida por (-15-1). (b) El *espectro de Rényi* calculado entre $q = -20$ y $q = 20$ de una determinada secuencia. Tanto (a) como (b) han empleado un fragmento del cromosoma 1 de *Homo sapiens* (NCBI *Reference Sequence*: ref|NC_000001.11|:124624088-124625552)

El *espectro de Rényi* se estabiliza para valores de q suficientemente grandes o pequeños. El rango del *espectro de Rényi*, denotado como $D_{-80,80}$, se debe calcular entre el q mínimo posible y el máximo, ya que se trata de una función decreciente en todo su dominio. Como se ve en la **Figura 7b**, con los valores de $q = -20$ y $q = 20$ las curvas se están estabilizando. Para mayor precisión en los cálculos de $D_{-80,80}$, se han empleado $q = -80$ y $q = 80$, puesto que exponentes con un valor absoluto mayor causan problemas de desbordamiento durante la ejecución del programa.

2.2.3. Tiempo de computación

Tanto el análisis fractal como multifractal presentan problemas en el tiempo de cómputo con los programas empleados. Se ha estimado el escalado del tiempo de computación de forma empírica para el equipo empleado, `Victus by HP laptop 16-d1xxx`, con un procesador `Intel Core i7` de 20 núcleos y 16GB de RAM. Como se observa en la **Figura 8**, el tiempo de procesamiento de una cadena crece de forma exponencial con el tamaño de la imagen asociada a su representación. La fórmula que aproxima el tiempo de computación $t(p)$ en función del tamaño del lado de la imagen en píxeles o nucleótidos p es la siguiente:

$$t(p) = e^{\frac{p-384,42}{128,14}} - 0,06$$

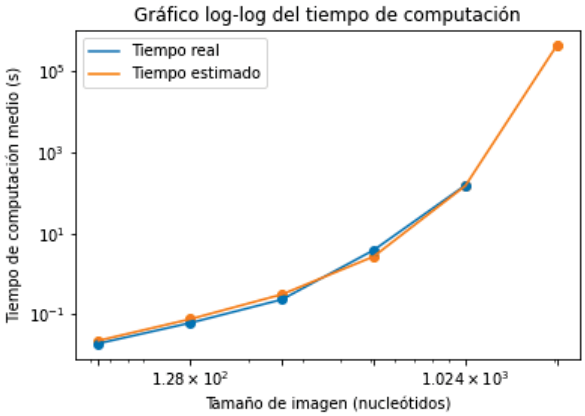


Figura 8: Azul: tiempo de computación medio para el cálculo de D_0 y D_1 para imágenes de distintos tamaños en el equipo `Victus by HP laptop 16-d1xxx`, con un procesador `Intel Core i7` de 20 núcleos y 16GB de RAM. Naranja: tiempo de computación estimado para este cálculo.

Se estima que el tiempo de cálculo de D_0 y D_1 de una imagen de 2048 píxeles será alrededor de 121 horas.

2.3. Herramientas para el estudio filogenético

Las herramientas utilizadas en el análisis filogenético pueden variar dependiendo de los datos disponibles, que suelen incluir secuencias de ADN, proteínas, características morfológicas o una combinación de las anteriores. La elección de las herramientas adecuadas es esencial para obtener resultados precisos y confiables.

La creación de árboles filogenéticos por métodos moleculares se realiza en varios pasos. Primero se mide la distancia genética entre las secuencias de ADN, ARN o proteínas empleadas. La matriz de distancias recoge el valor para cada pareja de clados que se empleen para construir el árbol. Tras esto, se agrupan los clados en función de esta distancia.

Se han empleado programas que implementan diversas funciones, tanto para el cálculo de distancias genéticas, modelos de sustitución óptimos, y creación, comparación y visualización de árboles filogenéticos.

- **MEGA11:** (Molecular Evolutionary Genetics Analysis version 11[44]) es uno de los programas más empleados en análisis filogenético. Se empleó la herramienta de alineamiento múltiple, tanto mediante el algoritmo de ClustalW[45] como de MUSCLE[46]. MEGA11 también permite calcular el mejor modelo de sustitución para un determinado conjunto de datos y crear árboles

filogenéticos empleando dichos modelos.

- **TREX[47]:** permite la creación de árboles filogenéticos por distintos métodos tomando como entrada una matriz de distancias.

- **Visual TreeCmp:** es una aplicación en línea[48], que emplea el programa de TreeCmp[49]. Permite comparar simultáneamente múltiples árboles con uno de referencia. Emplea hasta 19 métricas distintas de comparación de árboles filogenéticos, además de representarlos.

2.3.1. Similitud entre secuencias

Para determinar la similitud entre las secuencias descritas en el **Apartado 2.1**, se emplea por un lado la distancia genética, que es la herramienta que se utiliza normalmente en filogenia molecular, y por otro lado, las métricas que dependen de la representación de Nandy.

La **distancia genética** mide la diferencia entre el material genético de diferentes organismos. Existen distintas maneras de calcularla, pero normalmente se emplean métodos basados en procesos de Poisson[50]. Los algoritmos suelen considerar una ratio de heterogeneidad de mutación entre sitios, que sigue una función Gamma, que depende de lo que se conoce por el parámetro de Gamma denominado α[51]. Además, pueden considerarse sitios invariables. También tienen en cuenta cómo se tratan los *gaps* y la posición del nucleótido en el codón, si se trata de una secuencia codificante.

Se utiliza una matriz estocástica, que representa las tasas instantáneas de cambio de un nucleótido a otro. Esta matriz será distinta según el modelo de sustitución elegido. En el estudio, se emplearon los modelos de sustitución de Kimura 2-parámetros (K2)[52], Tamura 3-parámetros (T92)[53] y Tamura y Nei 1993 (TN93)[54].

Como ya se ha explicado en el **Apartado 2.2**, la representación de Nandy transforma las secuencias de ADN en conjuntos de puntos discretos en el plano, de manera que la similitud entre las secuencias se puede estimar de manera geométrica **(Figura 5)**. El **objetivo 2** del TFG consiste en evaluar tres funciones distintas que estimen las relaciones filogenéticas entre clados a partir de la similitud geométrica entre las representaciones de Nandy. En el estudio de Mizuta[55], se llegó a la conclusión de que la distancia del Taxi estima mejor las relaciones filogenéticas que la euclídea trabajando con esta representación. Esto se explica porque, a diferencia de la distancia euclídea, si se representan las 4 secuencias de un nucleótido posibles, todas son equidistantes. Por lo tanto, para medir distancias entre puntos, se empleó la distancia del Taxi.

Las funciones que se han evaluado son la distancia de Hausdorff[56], y dos funciones propuestas denominadas como d_{Sm} y d_{SM}.

Distancia de Hausdorff: Sean X y Y dos subconjuntos compactos del plano, la distancia de Hausdorff $d_H(X, Y)$ se define como:

$$d_H(X,Y) \ = \text{máx}\left\{d_1, d_2\right\}, \quad \begin{cases} d_1 = \sup_{x\in X}\{\text{ínf}_{y\in Y}\, d(x,y)\} \\ d_2 = \sup_{y\in Y}\{\text{ínf}_{x\in X}\, d(x,y)\} \end{cases},$$

siendo $d(x, y)$ la distancia entre dos puntos del plano.
Las dos funciones propuestas parten de una primera denominada Función Mínima Distancia Media:

$$d_S\ (X,\ Y) = \frac{1}{|X|}\sum_{x\varepsilon X} \text{mín}_{y\in Y}\ d(x,y)\ .$$

Esta se entiende como el promedio de la distancia a la que cada punto del conjunto X se encuentra del punto más cercano del conjunto Y. No se trata de una distancia puesto que, aunque sí cumple la condición de no negatividad, no cumple la de simetría. Las siguientes opciones sí cumplen la simetría:

- **Función Mínima Distancia Media Media**:

$$d_{Sm}(X,Y) = \frac{d_S(X,Y) + d_S(Y,X)}{2}$$

- **Función Mínima Distancia Media Máxima**:

$$d_{SM}(X,Y) = \text{máx}\{d_S(X,Y), d_S(Y,X)\}$$

No se ha comprobado que se cumpla la desigualdad triangular para ninguna de las tres funciones, por lo tanto no puede asegurarse que sean distancias en sentido estricto.

2.3.2. Árboles filogenéticos

Una vez obtenida la matriz de distancias, necesitamos emplear un método de agrupación. Se ha empleado el algoritmo *Neighbor-joining*, un método de agrupamiento de abajo arriba (aglomerativo), a menudo empleado en la creación de árboles filogenéticos[57]. Es, junto al UPGMA[58] (*unweighted pair group method with arithmetic mean*), uno de los métodos de construcción de árboles filogenéticos

basados en distancias genéticas más sencillos.

Tanto los árboles creados a partir de distancias genéticas como los de las funciones propuestas se han agrupado siguiendo un mismo método, puesto que se pretende evaluar únicamente la calidad de las métricas de similitud propuestas. Existen muchas aproximaciones para evaluar la calidad de los árboles creados con los árboles de referencia. Se han empleado distintas métricas que pueden complementarse. El **índice de Robinson-Foulds**[59], designado como RF, se trata de una distancia simétrica para calcular la diferencia entre dos cladogramas o árboles filogenéticos no enraizados (aunque es aplicable a enraizados). Únicamente se basa en la topología de estos, es decir, es independiente de la longitud de las ramas. El método tiene numerosos inconvenientes, como que se trata de una medida con sesgo, por eso, a menudo se emplea el **Índice Generalizado de Robinson-Foulds** (RFG), más consistente tanto teórica como prácticamente[60]. A menudo, los programas que emplean tanto RF o RFG, dividen entre dos el resultado, denotándose como RF(0.5) y RFG(0.5) respectivamente. Finalmente, la **Distancia Geodésica No Enraizada** es un método basado únicamente en la longitud de las ramas, y no en las relaciones topológicas del árbol. Se trata de calcular la suma total de la diferencia en la longitud de las ramas que conectan cada par de taxones para los dos árboles que se pretenden comparar.

2.4. Localización de secuencias funcionales

Para la localización de secuencias funcionales contenidas en largas secuencias de ADN, se empleó un enfoque probabilístico. Dado un fragmento de ADN, se denota su distribución de nucleótidos por $P = [p_A, p_T, p_C, p_G]$, donde p_A, p_T, p_C, p_G son la frecuencia relativa para cada uno de los nucleótidos ATCG.

Entropía de Shannon[61]: Mide la heterogeneidad o información promedio de una distribución. En el caso de la distribución de nucleótidos descrita como $P = [p_A, p_T, p_C, p_G]$, la entropía se calcula como:

$$H(P) = -\sum_{i \in I} p_i \ \log_2 p_i \ ,$$

con $I = \{A, T, C, G\}$.

La entropía de Shannon ha sido utilizada previamente para el estudio del ADN, que ha llevado a la creación de modelos predictivos[62]. Estos han mostrado gran utilidad, desde descubrir zonas con funciones hasta descubrir patrones en las secuencias de ADN.

La divergencia de Kullback-Leibler: es una medida asimétrica de similitud entre dos distribuciones de probabilidad, basado en el concepto de entropía de Shannon[63]. Dadas dos distribuciones P y Q, donde P es la distribución observada y Q a la que se espera que se ajusten los datos, la divergencia de Kullback-Leibler, que se denota como $D_{KL}(P||Q)$, se calcula como:

$$D_{KL}(P||Q) = \sum_{i \in I} p_i \, log\frac{p_i}{q_i} \, .$$

La divergencia de Kullback-Leibler no tiene límite superior en el caso general, pero sí en el caso discreto[64]. En la cuestión que aborda este TFG, $D_{KL}(P||Q)$ toma valores en el intervalo [0,2], donde en el caso de dos distribuciones idénticas, tendría valor 0, y para las distribuciones $P = [0{,}25, 0{,}25, 0{,}25, 0{,}25]$ y $Q = [1, 0, 0, 0]$, tiene un valor de 2.

2.4.1. Simulaciones

Para realizar los distintos experimentos ha sido necesario simular cadenas de ADN cuya distribución de nucleótidos provenga de una distribución determinada P. Estas se obtuvieron a través de un programa propio que emplea la librería `Numpy`. Se define como ronda de simulación la creación de secuencias de tamaños crecientes que siguen una distribución P. Los tamaños de secuencia se denominan n, y se realizan un total de N simulaciones para cada tamaño. De manera que una ronda puede tener el siguiente aspecto:

- Distribución de nucleótidos (P): [0.5000, 0.2500, 0.1250, 0.1250]
- Tamaños de secuencia (n): [1, 3, 10, 32, 100, 10^3, 10^4, 2×10^4, 3×10^4, 7×10^4]
- Número de simulaciones por cada tamaño n (N): 2×10^4

Capítulo 3. Resultados y discusión

El capítulo de resultados ha sido dividido en tres secciones, correspondientes a cada uno de los objetivos descritos en el **Apartado 1**.

3.1. Objetivo 1: caracterización de secuencias de DNA

La caracterización geométrica se realizó a través del C_{iso} y la dimensión fractal. La caracterización multifractal consistió en el cálculo de D_q, para q en el rango de -80 a 80, de la distribución de masa de la representación de Nandy de cada secuencia. Los parámetros que se consideraron en las comparaciones fueron D_0, D_1 y el rango del espectro multifractal ($D_{-80,80}$). Se caracterizaron las 10400 secuencias de la base de datos **(Apartado 2.1)**. Tanto los tamaños de las secuencias y sus medias, desviaciones típicas y rangos, quedan recogidos en los **Cuadros suplementarios A.1, A.2, A.3 y A.4**. Tras el cálculo de estos valores, se quiso comprobar lo siguiente:

- **[Experimento 1]**: Si para un mismo gen, existen diferencias significativas entre las características medidas entre reinos o dominios. Para este experimento se empleó el grupo de secuencias denominado **Grupo ARNr 16S**, compuesto por secuencias de ARNr 16S de animales, arqueas, plantas, bacterias y protistas.

- **[Experimento 2]**: Si las características medidas son significativamente diferentes entre grandes grupos filogenéticos como reinos o dominios. Para esto, se empleó el grupo de secuencias denominado **Grupo Conglomerado**, formado por secuencias de ADN genómico de animales, bacterias y plantas. Se ha acotado el tamaño de las secuencias a un rango entre aproximadamente 800 y 1350 nucleótidos, debido a las razones relacionadas con el tiempo de cómputo **(Apartado 2.2.1)**.

- **[Experimento 3]**: Si existe una diferencia significativa entre la caracterización de las secuencias de ADN genómico, pertenecientes al **Grupo Conglomerado**, y los ARNm, que forman parte del **Grupo ARNm**, puesto que estos últimos carecen de intrones. Como se explica en el **Apartado 1.2** la composición de nucleótidos difiere entre intrones y exones, razón por la que es de interés estudiar como repercute esto en las características medidas. Para ello, se han analizado secuencias de ADN genómico y ARNs de animales, de nuevo acotadas en el rango de 800 a 1350 nucleótidos.

- **[Experimento 4]**: Si se encuentran diferencias significativas entre las secuencias homólogas de genes distintos. Para ello se han empleado secuencias homólogas de *p53* y *cytb*, pertenecientes al **Grupo Genes** descrito en el **Apartado 2.1**.

En primer lugar, se realizaron las pruebas ANOVA para todas las características de los grupos del **Experimento 1** y **Experimento 2**, ya que son los únicos que comparan más de dos clases. Como se observa en el **Cuadro suplementario A.6**, las pruebas ANOVA obtuvieron un valor p inferior a 3.14E-10 para ambos experimentos y las 4 características. Tras esto se realizó una prueba de Tukey[65], que contrasta todas las parejas de clases del grupo, corrigiendo la significancia en función del número total de pruebas.

En el **Experimento 1**, se rechazó la hipótesis nula de igualdad de medias en todos los contrastes de características menos para las comparaciones de bacterias contra plantas y de plantas contra protistas (**Cuadro suplementario A.7**). Para rechazar las hipótesis, se empleó una Tasa de Error Global (FWER[66]) de 0.05. Las características D_0, D_1 y el rango se muestra a través de un diagrama de caja y bigotes en la **Figura 9**.

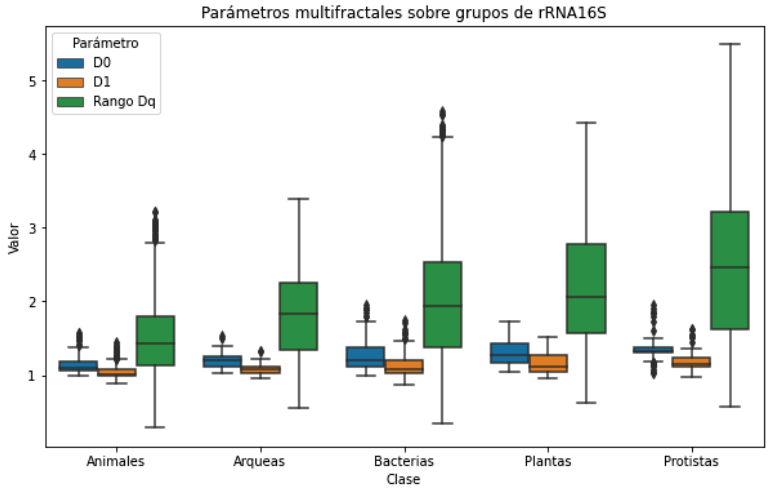


Figura 9: Diagrama de caja y bigotes de características D_0, D_1 y Rango D_q ($D_{-80,80}$), calculado entre D_{-80} y D_{80}, de las clases del **Experimento 1**. Representado con el paquete `Seaborn`

Como se observa en el **Cuadro suplementario A.1**, precisamente son las clases de plantas y protistas, junto con arqueas, las que tienen un menor número de secuencias, lo que apunta a que los resultados puedan no ser fiables. Por lo tanto, no se rechaza la hipótesis de que existe una diferencia entre los valores de coeficiente isoperimétrico, un carácter geométrico; y $D_{-80,80}$, un carácter distribucional, para las clases de bacterias y plantas, con valores p de 0.1939 y 0.5248 respectivamente. Para los contrastes

de plantas y protistas, tampoco se rechaza la hipótesis nula para las diferencias en $D_{-80,80}$, D_0 y D_1, con valores p de 0.1259, 0.087, 0.5968, respectivamente; de nuevo, carácteres tanto geométricos como distribucionales.

En el **Experimento 2**, tras la prueba de Tukey **(Cuadro suplementario A.8)**, se rechaza la hipótesis nula para todas las características de los contrastes de animales con bacterias (valor p menor a 10E-6 para todas las características) y animales con plantas (valor p menor a 10E-6 para D_0 y C_{iso}, y 0.0001 para D_1 y $D_{-80,80}$). Sin embargo, en el contraste de bacterias contra plantas, solo se rechaza para el coeficiente isoperimétrico (0.0466), mientras que no se rechaza para $D_{-80,80}$, D_0 y D_1 (0.2662, 0.7837, 0.6308), tal y como ocurría en el **Experimento 1**.

El **Experimento 3** y **Experimento 4**, al estar compuestos por dos clases cada uno, se han realizado pruebas T de Student en ambas. En el **Experimento 3**, en el contraste de características de secuencias de ADN genómico y ARNm de animales, solo se obtuvieron diferencias significativas en $D_{-80,80}$, con un valor p asociado de 4.22E-5 **(Cuadro suplementario A.9)**. Destaca la diferencia entre el valor p de D_0 con el de D_1, puesto que en el primer caso el valor es de 0.011, y en el segundo, 0.890. En el **Experimento 4**, todos los contrastes fueron significativos, tanto para D_0, D_1, $D_{-80,80}$ y C_{iso}, con valores p de 5.13E-27, 1.46E-30, 2.64E-7 y 5.10E-14 respectivamente, rechazándose la igualdad entre dichas características **(Cuadro suplementario A.10)**.

Tanto el **Experimento 1** y el **Experimento 2** apuntan a que tanto las características geométricas como multifractales sirven para distinguir el reino o dominio del organismo del que provenga la secuencia. Esto ocurre tanto si se estudian secuencias homólogas del mismo gen como si se toman genes diferentes. En ninguno de los dos experimentos se ha rechazado la hipótesis de igualdad de medias entre las características de D_0, D_1 y rango para las comparaciones entre las clases de bacterias y plantas. Esto puede indicar que las características estudiadas en estos dos grupos se parecen más entre sí que al resto, y no se puede reconocer a qué grupo pertenece la secuencia.

El **Experimento 3** solo ha mostrado diferencia significativa en $D_{-80,80}$. Se podría realizar en el futuro un estudio que emplee secuencias de ADN genómico con un contenido conocido de intrones, puesto que aquí no se ha cuantificado cuánto suponen del total, o un contraste entre exones e intrones. Como se explicó en el **Apartado 1.2**, la composición de nucleótidos es diferente en exones e intrones, lo que puede afectar a los parámetros que se han caracterizado.

Las pruebas empleadas se basan en premisas como la homogeneidad de varianzas o la normalidad. Como se observa en los **Cuadros suplementarios A.11**, los resultados de la prueba de Levene mues-

tran que no existe homogeneidad en las varianzas para la mayoría de las características medidas en los grupos. Así mismo ocurre con la normalidad, estudiada a través de la prueba Shapiro[67] **(Cuadro A.12)**. Por lo tanto, puede que los resultados no sean fiables.

3.2. Objetivo 2: estudio filogenético

Se tomaron las secuencias descritas en la base de datos del **Apartado 2.1**, que comprende 5 grupos de genes homólogos, cada uno formado por 10 secuencias.

Para crear árboles filogenéticos de referencia, cada conjunto fue introducido en MEGA11, donde se realizó su alineamiento múltiple. Al ser *citocromo C*, *gapdh*, *h3* y *hsp90* genes que codifican para proteínas, se empleó MUSCLE, mientras que para ARNr 18s se empleó ClustalW. Tras esto, se calculó el mejor modelo de sustitución nucleotídica, que se usó para la creación de árboles filogenéticos mediante Neighbor-joining. Todo esto se recoge en el **Cuadro suplementario A.13**.

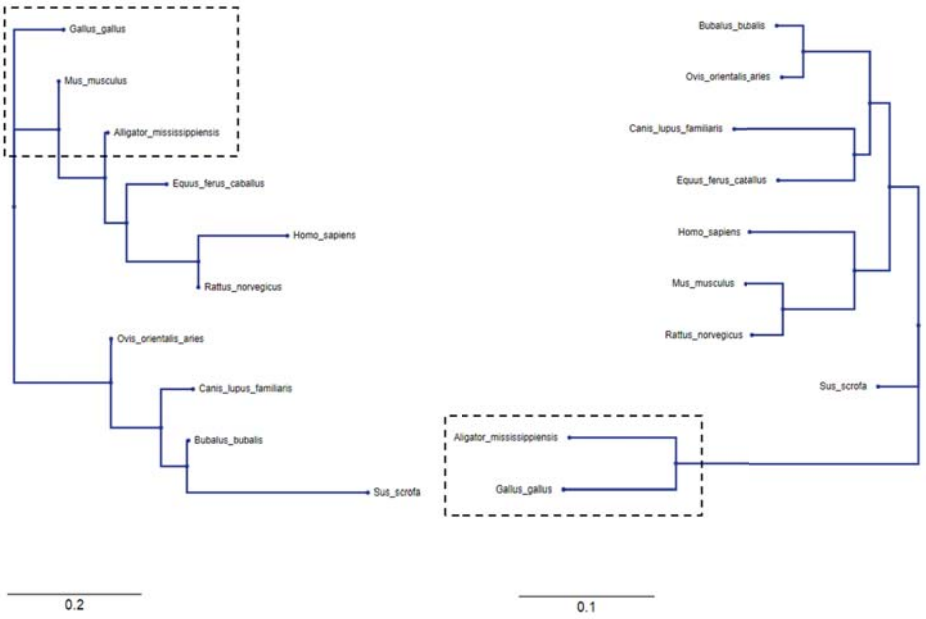


Figura 10: Árboles filogenéticos creados con 10 genes homólogos de *hsp90*. A la izquierda, distancias entre clados estimadas mediante d_{SM}, aplicada sobre conjunto de puntos obtenidos a partir de representación de Nandy. A la derecha, distancia entre clados usando la distancia genética sobre las secuencias. Los recuadros negros señalan las diferencias en la posición relativa de *Gallus gallus* y *Alligator mississippiensis*

Como se pretendía evaluar las métricas empleando un mismo tipo de agrupamiento, no se emplearon métodos que ofrecen mejores resultados como árboles de máxima verosimilitud, de máxima

parsimonia o basados en cadenas de Markov.

A partir de las representaciones de Nandy, se crearon matrices de distancias para cada grupo de genes, empleando la distancia de Hausdorff, d_{Sm} y d_{SM}. Posteriormente, se utilizó Trex para crear árboles mediante *Neighbor-joining* **(Figura 10)**.

Para comparar los árboles creados con la distancia de Hausdorff, d_{Sm} y d_{SM} con los árboles de referencia, se empleó Visual TreeCmp. Los resultados obtenidos se recogen en el **Cuadro 1**.

Cuadro 1: Se presentan los índices de árboles filogenéticos creados mediante los distintos métodos que derivan de la representación de Nandy. RFG (0.5) denota RFG divido entre dos y n el tamaño de las secuencias de ADN. Se dividió el resultado de las funciones propuestas entre 100 para poder ser leídas por Trex.

Métrica	**Gen**					**Índice**
	citocromo C	***gapdg***	***h3***	***hsp90***	**ARNr 18S**	
$\mathbf{d_{Sm}}$	15.0322	3.4799	4.1561	0.6557	0.8481	**RFG (0.5)**
	12.4181	2.5514	3.5451	0.4399	0.8059	**DGNE**
$\mathbf{d_{SM}}$	15.0241	0.7954	0.9382	0.9374	1.0986	**RFG (0.5)**
	12.6777	0.5461	0.7327	0.609	1.1663	**DGNE**
Hausdorff	16.5376	1.2185	1.4516	1.736	2.3317	**RFG (0.5)**
	13.5641	0.856	1.0738	1.0861	2.2916	**DGNE**
Media (n)	1632.60	1283.50	1082.50	1283.50	1905.10	**Media** (n)
Varianza (n)	2443374.93	7302.72	14300.06	7302.73	22345.88	**Varianza** (n)

Las comparaciones de árboles con peores índices asociados son los correspondientes al *citocromo C*, y esto puede deberse a que las secuencias empleadas tienen longitudes muy distintas, como se observa en su rango y varianza. Esto es un factor que afecta especialmente a la representación de Nandy, puesto que la inserción de muchos nucleótidos puede alejar zonas similares en el plano. Sin embargo, el resto de los grupos muestran índices 10 veces inferiores.

El árbol con mayor similitud observada es el árbol creado con d_{Sm} para *hsp90* **(Figura 10)**. Aún así, cuenta con numerosos errores, tal como el situar al ratón (*Mus musculus*) entre la gallina (*Gallus gallus*) y el aligátor (*Aligator mississippiensis*).

De las tres métricas empleadas, en promedio d_{SM} proporciona los mejores resultados. Sin embargo, para algunos de los grupos de árboles de los genes *hsp90*, ARNr 18S y, para uno de los índices únicamente, *citocromo C*, es d_{Sm} quien proporciona mejores resultados. Además de tener en cuenta los

índices, se examinaron visualmente los árboles debido a que, aunque los valores apunten a que una de las medidas ha dado mejor resultado, existen relaciones filogenéticas que tienen más importancia que otras. Cometer errores en el agrupamiento de clados que comparten categorías taxonómicas de mayor orden, como la clase, tiene más relevancia que hacerlo con categorías inferiores, como la familia.

Se examinaron los árboles correspondientes a ARNr 18S, que son prácticamente idénticos **(Figura B.2)**, pero con diferencias en la colocación de *Zea mays*, *Prosopis cineraria* y *Drosophila melanogaster*, dos plantas y un insecto. Por lo tanto, aunque d_{SM} obtuvo peores resultados para RFG(0.5) y DGNE, su árbol concuerda más con las relaciones filogenéticas de las especies empleadas.

En cuanto al gen *hsp90* también se observan diferencias en la colocación de los clados en los árboles de d_{Sm} y d_{SM} **(Figura B.3)**. Esto ocurre en la disposición de *Canis lupus*, *Ovis orientalis aries* y *Bubalus bubalis*. Los dos últimos clados pertenecen al mismo orden, y d_{SM} los ha colocado de una manera más adecuada, situando a *Ovis orientalis aries* y *Bubalus bubalis* en un grupo más bajo, especies que a diferencia de *Canis lupus*, comparten orden.

Las tres medidas de similitud cometen errores y agrupan peor los clados que los métodos basados en distancias genéticas. Incluso el árbol con mejores índices de entre todos los generados comete errores en la colocación de clados pertenecientes a un mismo orden o clase. Esto puede ser debido a diferentes razones:

- Las medidas se ven poco afectadas por las inserciones de una o pocas bases, un tipo de mutación de gran importancia dado que puede originar cambios en los marcos de lectura. Esto ocurre debido a que, en el plano, la inserción de una base solo supone el desplazamiento de una unidad. Sin embargo, se trata de un hecho con gran importancia biológica. De igual modo, no es posible detectar la aparición de codones de parada, o distinguir entre mutaciones sinónimas o no.

- Presenta un problema ante las grandes inserciones, ya que alejan zonas similares, además de aumentar la distancia entre los conjuntos de puntos. Por eso, en el caso del *citocromo C*, el resultado era mucho peor que en el resto, porque algunas de las secuencias triplicaban en longitud a sus homólogos. En cambio, la distancia genética sí permite trabajar con estas inserciones de gran tamaño.

- Los métodos basados en la representación de Nandy no hacen posible el alineamiento múltiple, lo que supone una gran ventaja para los métodos tradicionales[68].

Otra desventaja respecto a la distancia genética es que estas aportan información temporal sobre la evolución de las secuencias, haciendo posible hacer reconstrucciones temporales. En cuanto a las técnicas de *Boostrap*[69], no han sido empleadas en el trabajo, pero también se podrían realizar sobre las representaciones de Nandy.

3.3. Objetivo 3: localización de secuencias funcionales

Se partió del supuesto ya explicado en el **Apartado 1.2**, de que, si una zona del cromosoma no se encuentra sometida a una presión de selección, su distribución de nucleótidos P tenderá a la del *isochore*, que denotaremos como P_{iso}, en el que se encuentra. Se rechazará la hipótesis nula de que un fragmento de cromosoma no es funcional cuando su distribución nucleotídica P sea distinta a P_{iso}. Debido a la complejidad de tener que hallar en qué *isochore* se encuentra cada fragmento de ADN y cuál es su distribución, se empleó la P del propio cromosoma P_{chr} ($P_{iso} = P_{chr}$). Como se puede observar en la **Figura 6**, la pendiente de la representación del cromosoma es constante, a excepción de la zona del centrómero, señalada con un recuadro negro. Esto indica que la distribución de nucleótidos es homogénea en toda su extensión.

Se estudió como se comporta la distribución de nucleótidos en las zonas no sujetas a presión de selección según su tamaño. Se realizó una serie de simulaciones para distintas distribuciones.

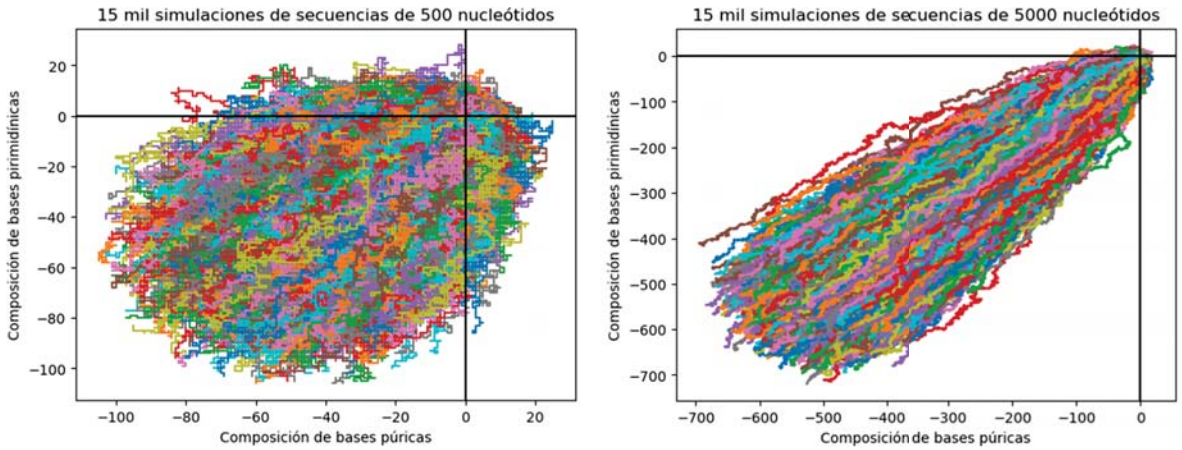


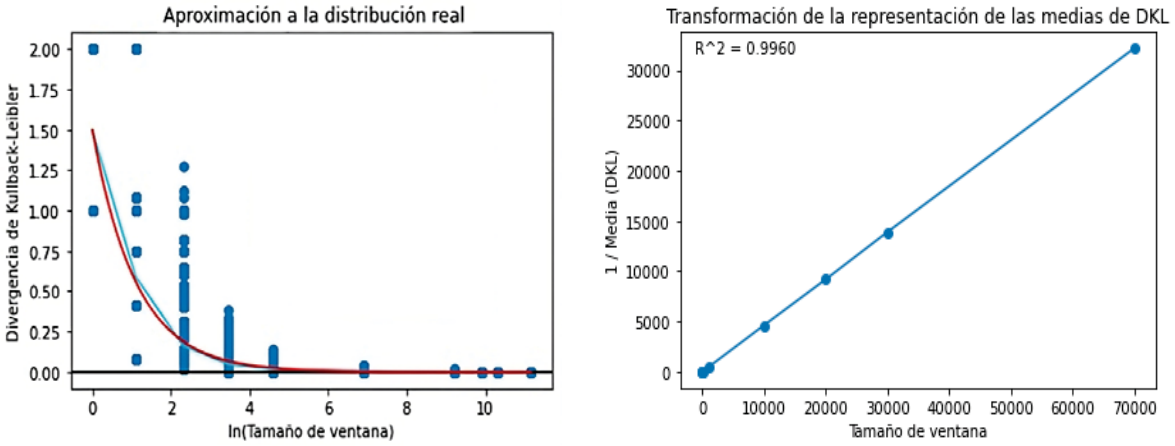
Figura 11: Representación de Nandy de las simulaciones de secuencias que siguen la $P_{chr} = [0{,}3003, 0{,}30195, 0{,}19720, 0{,}20055]$, donde P_{chr} es la distribución de nucleótidos del cromosoma 18 de *Homo sapiens*. A la izquierda, 15 mil simulaciones de 500 nucleótidos. A la derecha, 15 mil simulaciones de 5000 nucleótidos

A medida que se simulen secuencias más grandes, la distribución observada P_{obs} se aproximará a la empleada para las simulaciones, P_S. Esto se puede observar de forma gráfica gracias a la representación de Nandy **(Figura 11)**.

Para medir la aproximación de las distribuciones observadas P_{obs} a la P_S, se empleó la divergencia de Kullback-Leibler (D_{KL}) **(Apartado 2.4)**. Así se obtuvo la función de distribución de probabilidad de la D_{KL} de forma empírica **(Figura B.4)**.

Tras esto, se simularon secuencias de diferentes tamaños para determinar cómo se comportan distintos parámetros de las distribuciones de la D_{KL} en función del tamaño de las secuencias simuladas. Para las P_S descritas en el **Cuadro 11**, se generaron 5×10^3 secuencias para cada uno de los siguientes tamaños de cadena: $[1, 3, 10, 32, 100, 10^3, 10^4, 2 \times 10^4, 3 \times 10^4, 7 \times 10^4]$.

Debido a que los tamaños seleccionados no son del mismo orden, en la **Figura 12a** se representa D_{KL} frente al logaritmo del tamaño de la secuencia.



(a) (b)

Figura 12: Comportamiento de $D_{KL}(P_{obs}||P_S)$. En este caso P_S sigue la P de la ronda de simulación de $S_2([0{,}5000, 0{,}2500, 0{,}1250, 0{,}1250])$. (a) En azul oscuro: $\ln D_{KL}$ de cada secuencia simulada frente a su tamaño de secuencia. En azul claro: la curva que representa la media de los $\ln D_{KL}$. En rojo: ajuste de la media a $K/ln(n)$, donde K es una constante y n el tamaño de la secuencia. (b) Azul oscuro: el inverso de las medias de D_{KL} frente al tamaño de la secuencia. En azul claro: su ajuste lineal.

En la **Figura 12a**, se observó que $\bar{D}_{KL}(n) \approx \frac{K}{n}$. Esto se comprobó para las simulaciones del **Cuadro A.14** y cuyos coeficientes de determinación (R^2), se muestran en el mismo cuadro. Haciendo la inversa, $\frac{1}{\bar{D}_{KL}(n)} \approx \frac{1}{K}n$ **(Figura 12b)**, y ajustando por nínimos cuadrados se observó que K es el valor de la entropía de Shannon de P_S.

Se estudió la media de D_{KL} de dos distribuciones muy distintas que tengan una H similar, ya que deberían comportarse del mismo modo. Se realizaron 15 mil simulaciones para cada tamaño de las proporciones $P_1 = [0{,}5, 0{,}25, 0{,}25, 0]$ y $P_2 = [0{,}64, 0{,}12, 0{,}12, 0{,}12]$, ambas con una H de aproximadamente 1.5 **(Cuadro A.15)**. Se observa en la **Figura 13** como la media de D_{KL} para ambas simulaciones se comporta de manera similar. Por lo tanto:

$$\bar{D}_{KL}(n) \approx \frac{H}{n} .$$

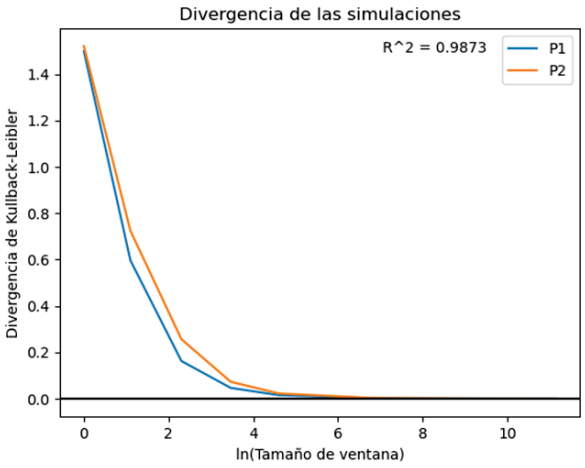
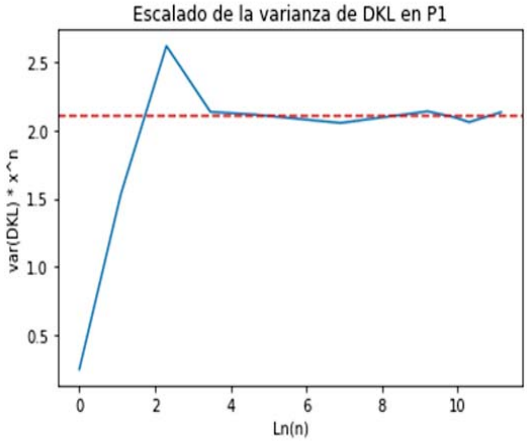
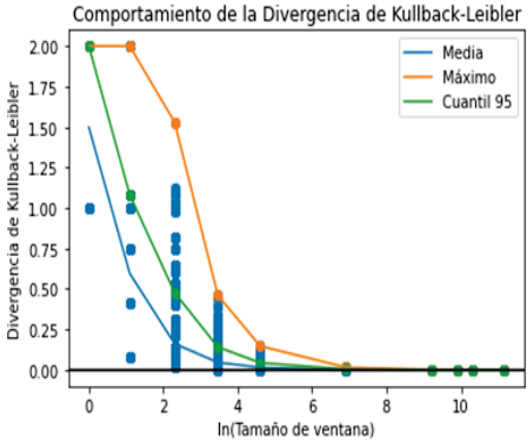


Figura 13: Ajuste entre las medias de $D_{KL}(P_{obs}||P_S)$ frente al logaritmo natural del tamaño de la secuencia para para P_1 y P_2

El **Máximo** y **cuantil 95** de la distribución son parámetros importantes a la hora de estudiar una distribución de probabilidad, puesto que el cuantil 95 se suele emplear como límite para determinar si algo es significativo o no. Como se observa en la **Figura 14b**, ambos parámetros se comportan como una constante dividida entre el tamaño de las secuencias (n), al igual que la media.



(a)



(b)

Figura 14: (a) Varianza de $D_{KL}(P_{obs}||P_1)$, multiplicada por el tamaño de la secuencia al cuadrado frente al logaritmo natural del tamaño de ventana. (b) Comportamiento de $D_{KL}(P_{obs}||P_{S2})$. Modificación de la **Figura 12.2**, que incluye el máximo valor obtenido (verde) y el valor del cuantil 95 (naranja)

En particular:

$$Max\,(n) \approx \frac{k_1}{n} \qquad y \qquad C95(n) \approx \frac{k_2}{n}\,.$$

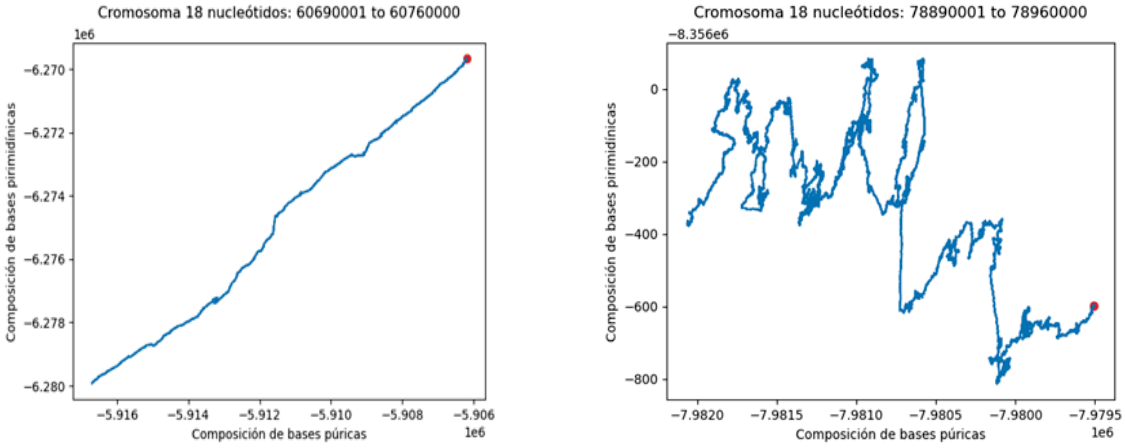
Donde $Max(n)$ denota el máximo valor de la D_{KL} y $C95(n)$, el valor del cuantil 95 obtenido en las simulaciones para el tamaño n. k_1 y k_2 denotan constantes desconocidas que no se han conseguido relacionar con parámetros conocidos.

La **varianza** también es un parámetro importante para caracterizar una determinada función de distribución. De nuevo, no se ha conseguido relacionarla con parámetros conocidos, pero como se observa en la **Figura 14a**, la varianza se comporta como:

$$V\,(n) \approx \frac{k_3}{n^2}\,,$$

donde $V(n)$ denota la varianza de D_{KL} y k_3 es una constante desconocida.

Se aplicó esta suposición sobre la distribución de nucleótidos en zonas no funcionales sobre un caso concreto. Se calcula la D_{KL} de la distribución de nucleótidos de la cadena tomada (P_{obs}) respecto a la distribución de nucleótidos del *isochore* donde se encuentra (P_{iso}). Si este valor es superior al cuantil 95 de la distribución de D_{KL} obtenida en las simulaciones, se rechazará la hipótesis nula, que dice que nos encontramos ante una zona no funcional.



(a) (b)

Figura 15: (a) Representación de Nandy de un segmento de 70 mil bases sin ningún elemento codifiante ni regulador detectado hasta el momento. (b) Representación de Nandy de un segmento de 70 mil bases con elementos reguladores detectados.

Se tomó la secuencia de referencia de NCBI (NC_000018.10) del cromosoma 18 de *Homo sapiens*,

ya que se trata de uno de los cromosomas más pequeños y por tanto es más fácil de tratar, cuya P_{chr} es $[0{,}3003, 0{,}30195, 0{,}19720, 0{,}20055]$. A través del visor Web de NCBI, se localizó una zona de un tamaño de 7×10^4 bases que no tuviese genes ni zonas reguladoras **(Figuras 15a y B.5)**.

Para esta secuencia, P_{obs}= [0.32564, 0.32255, 0.17617, 0.17564], por lo que la $D_{KL}(P_{obs}||P_{chr})$ = 0.0065. El cuantil 95 para esta longitud de cadena es varios órdenes menor, 8.17E-5. Por lo tanto, la premisa de que P_{chr} = P_{iso} es errónea y hay que reformular el planteamiento. Los P_{iso} tienen una distribución de nucleótidos similar a P_{chr}, con una desviación respecto a esta que no puede observarse a través de la representación de Nandy de la **Figura 6**. Dicho de otro modo:

$$P_{chr} \; = \; P_{iso\ (i)} \pm \delta_{(i)} \; \forall \, i \; ,$$

donde $\delta_{(i)}$ es término de error para el $P_{iso(i)}$, con $i = 1 \ldots N$, siendo N el número total de *isochores* en el cromosoma. Averiguar este $\delta_{(i)}$ puede ser muy complejo, así que se siguió el trabajo de **Ray et al., 2019**[70], que empleó un umbral de 0.1 para determinar si dos distribuciones son similares. Como las distribución de dicho estudio y las de este son distintas, se ha decidido ajustar dicho valor. Se ha empleado un umbral 10 veces más pequeño (0.01) equivalente a medir D_{KL} entre una P=[0.27, 0.27, 0.23, 0.23] y la P_{chr}=[0.3003, 0.30195, 0.19720, 0.20055]. Así, el máximo valor de $D_{KL}(P_{iso}||P_{chr})$ (Máximo(D_{KL})) será 0.01, puesto que las consideramos similares. Se rechazará la hipótesis de que una cadena del cromosoma de tamaño n no tenga función cuando $D_{KL}(P_{obs}||P_{chr}) > C95(n)$+Máximo$(D_{KL}) = C95(n) + 0{,}01$. En el caso de estudio, $C95(n)$ es del orden de 10E-5, despreciable frente a 0.01. Por lo tanto, se rechazará la hipótesis nula de que la secuencia es no funcional si $D_{KL}(P_{obs}||P_{chr}) > 0{,}01$. No se rechaza la hipótesis nula para el caso de la **Figura 15a**, ya que $D_{KL}(P_{obs}||P_{chr}) = 0{,}00650 < 0{,}01$.

Se prueba a buscar un caso contrario, un fragmento de 7×10^4 bases de ADN intergénico en el que sí se hallen secuencias funcionales y se observe una distribución de nucleótidos distinta a lo esperado si no tuviesen función **(Figuras 15b y B.6)**. Para este fragmento, se tiene una P_{obs} = [0.2712, 0.2455, 0.24866, 0.23464], y por lo tanto, una $D_{KL}(P_{obs}||P_{chr})$ = 0.02314, que es mayor que el umbral 0.01, por lo que rechazamos que se trate de una zona sin función.

Por lo tanto, la herramienta indica la posible existencia de elementos reguladores o con función cuando se observa una distribución de nucleótidos distinta a la esperada en un segmento de ADN. Es necesario realizar estudios sobre la utilidad y limitaciones de esta herramienta.

Capítulo 4. Conclusiones

1. Las métricas de similitud entre los conjuntos de puntos de las representaciones de Nandy propuestas pueden ser de ayuda a métodos tradicionales de estudios filogenéticos. Sin embargo, cometen errores en la colocación de algunos clados. Esto es debido a la sensibilidad de estas métricas. Es posible que existan funciones de esta clase que ofrezcan mejores resultados, por lo que se debe seguir explorando este campo.

2. Las representaciones de Nandy pueden ser caracterizadas mediante un análisis multifractal si se convierten en una distribución de masas en el plano. Esto es una nueva herramienta de estudio de secuencias de ADN, con un potencial aún por explorar.

3. Las características geométricas y multifractales de las representaciones de Nandy pueden servir para distinguir entre reinos y dominios, y entre secuencias homólogas de un mismo gen. El rango del espectro multifractal $D_{-80,80}$ puede servir para diferenciar entre secuencias genómicas y mRNA de un mismo reino.

4. La herramienta propuesta de localización de secuencias funcionales podría ser de utilidad como complemento a otros procedimientos con el mismo fin. Su utilidad y limitaciones aún no han sido determinadas.

Referencias

[1] DNA Data Bank of Japan (DDBJ). [Inernet]. National Institute of Genetics, 1986. URL: https://www.ddbj.nig.ac.jp/.

[2] European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL). [Inernet]. European Molecular Biology Laboratory, 1980. URL: https://www.embl.org/.

[3] National Center for Biotechnology Information (NCBI). [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, 1988. URL: https://www.ncbi.nlm.nih.gov.

[4] GenBank and WGS Statistics. [Internet]. Bethesda (MD): National Library of Medicine (US), 2023. URL: https://www.ncbi.nlm.nih.gov/genbank/statistics/ (visitado 17-06-2023).

[5] Nandy A. A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Application to Globin Genes. En: *Current science* Current Sc 309 (1994) 309-314 (feb. de 1994).

[6] Sayem AS. A quantitative method for measuring and visualizing species' relatedness in a two-dimensional Euclidean space. En: *Electronic Thesis and Dissertation Repository* (abr. de 2013).

[7] Qi ZH, Li L y Qi XQ. Using Huffman coding method to visualize and analyze DNA sequences. En: *Journal of Computational Chemistry* 32 (15 nov. de 2011), págs. 3233-3240. ISSN: 01928651. DOI: 10.1002/jcc.21906.

[8] Yau SS, Wang J, Niknejad A, Lu C, Jin N y Ho YK. DNA sequence representation without degeneracy. En: *Nucleic acids research* 31 (12 jun. de 2003), págs. 3078-3080. ISSN: 1362-4962. DOI: 10.1093/NAR/GKG432.

[9] Falconer K. Front Matter. En: *Fractal Geometry* (ene. de 2003), págs. i-xxvii. DOI: 10.1002/0470013850.FMATTER.

[10] Mandelbrot B. How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. En: *Science* 156 (mayo de 1967), págs. 636-638. ISSN: 00368075. DOI: 10.1126/SCIENCE.156.3775.636.

[11] Harte D. Multifractals : Theory and Applications. En: *Multifractals* (jun. de 2001). DOI: 10.1201/9781420036008.

[12] Baranowski P, Krzyszczak J, Slawinski C, Hoffmann H, Kozyra J, Nieróbca A et al. Multifractal analysis of meteorological time series to assess climate impacts. En: *Climate Research* 65 (sep. de 2015), págs. 39-52. ISSN: 0936-577X. DOI: 10.3354/CR01321.

[13] Krupovic M, Dolja VV y Koonin EV. The LUCA and its complex virome. En: *Nature reviews. Microbiology* 18 (11 nov. de 2020), págs. 661-670. ISSN: 1740-1534. DOI: 10.1038/S41579-020-0408-X.

[14] Moreira D y Philippe H. Molecular phylogeny: pitfalls and progress. En: *International microbiology : the official journal of the Spanish Society for Microbiology* (2000). DOI: 10.2436/IM.V3I1.9236.

[15] Bacon DJ y Anderson WF. Multiple sequence alignment. En: *Journal of molecular biology* 191 (2 sep. de 1986), págs. 153-161. ISSN: 0022-2836. DOI: 10.1016/0022-2836(86)90252-4.

[16] Palazzo AF y Lee ES. Non-coding RNA: what is functional and what is junk? En: *Frontiers in genetics* 6 (JAN 2015). ISSN: 1664-8021. DOI: 10.3389/FGENE.2015.00002.

[17] Kim TH y Dekker J. ChIP-seq. En: *Cold Spring Harbor protocols* 2018 (5 mayo de 2018), págs. 363-367. ISSN: 1559-6095. DOI: 10.1101/PDB.PROT082644.

[18] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. The sequence of the human genome. En: *Science* 291 (5507 feb. de 2001), págs. 1304-1351. ISSN: 00368075. DOI: 10.1126/SCIENCE.1058040/SUPPL_FILE/C16_SCIENCE.PDF.

[19] Rearick D, Prakash A, McSweeny A, Shepard SS, Fedorova L y Fedorov A. Critical association of ncRNA with introns. En: *Nucleic acids research* 39 (6 mar. de 2011), págs. 2357-2366. ISSN: 1362-4962. DOI: 10.1093/NAR/GKQ1080.

[20] Chorev M y Carmel L. The function of introns. En: *Frontiers in genetics* 3 (APR 2012). ISSN: 1664-8021. DOI: 10.3389/FGENE.2012.00055.

[21] Oliver JL, Carpena P, Román-Roldán R, Mata-Balaguer T, Mejías-Romero A, Hackenberg M et al. Isochore chromosome maps of the human genome. En: *Gene* 300 (1-2 oct. de 2002), págs. 117-127. ISSN: 03781119. DOI: 10.1016/S0378-1119(02)01034-X.

[22] Cohen N, Dagan T, Stone L y Graur D. GC Composition of the Human Genome: In Search of Isochores. En: *Molecular Biology and Evolution* 22 (5 mayo de 2005), págs. 1260-1272. ISSN: 0737-4038. DOI: 10.1093/MOLBEV/MSI115.

[23] Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. En: *Experientia* 6 (6 jun. de 1950), págs. 201-209. ISSN: 00144754. DOI: 10.1007/BF02173653/METRICS.

[24] Hershberg R y Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. En: *PLoS Genet* 6 (9 2010), pág. 1001115. DOI: 10.1371/journal.pgen.1001115.

[25] Galtier N y Duret L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. En: *Trends in genetics : TIG* 23 (6 jun. de 2007), págs. 273-277. ISSN: 0168-9525. DOI: 10.1016/J.TIG.2007.03.011.

[26] Salwan R y Sharma V. Genomics of prokaryotic extremophiles to unfold the mystery of survival in extreme environments. En: *Microbiological research* 264 (nov. de 2022). ISSN: 1618-0623. DOI: 10.1016/J.MICRES.2022.127156.

[27] Ganyecz A, Kaílay M y Csontos JJ. Thermochemistry of Uracil, Thymine, Cytosine, and Adenine. En: (2019). DOI: 10.1021/acs.jpca.9b02061.

[28] Duret L. Neutral Theory: The Null Hypothesis of Molecular Evolution. En: *Nature Education* 1 (1 2008), págs. 803-806.

[29] Kartavtsev YP, Batischeva NM, Bogutskaya NG, Katugina AO y Hanzawa N. Molecular systematics and DNA barcoding of Altai osmans, oreoleuciscus (pisces, cyprinidae, and leuciscinae), and their nearest relatives, inferred from sequences of cytochrome b (Cyt-b), cytochrome oxidase c (Co-1), and complete mitochondrial genome. En: *Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis* 28 (4 jul. de 2017), págs. 502-517. ISSN: 2470-1408. DOI: 10.3109/24701394.2016.1149822.

[30] Kanapathipillai M. Treating p53 Mutant Aggregation-Associated Cancer. En: *Cancers* 10 (6 jun. de 2018). ISSN: 2072-6694. DOI: 10.3390/CANCERS10060154.

[31] Esposti MD. On the evolution of cytochrome oxidases consuming oxygen. En: *Biochimica et biophysica acta. Bioenergetics* 1861 (12 dic. de 2020). ISSN: 1879-2650. DOI: 10.1016/J.BBABIO.2020.148304.

[32] Madani M, Tenuta M y Handoo A. Molecular Characterization and Phylogeny of Ditylenchus weischeri from Cirsium arvense in the Prairie Provinces of Canada. En: *Journal of nematology* 2 (2018). DOI: 10.21307/jofnem-2018-011.

[33] Hernández M, Martín MV, Herrador-Gómez PM, Jiménez S, Hernández-González C, Barreiro S et al. Mitocondrial COI and 16S rDNA sequences support morphological identification and biogeography of deep-sea red crabs of the genus Chaceon (Crustacea, Decapoda, Geryonidae) in the Eastern Central and South Atlantic Ocean. En: *PloS one* 14 (2 feb. de 2019). ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0211717.

[34] Guerrero R, Margulis L y Berlanga M. Symbiogenesis: the holobiont as a unit of evolution. En: *International microbiology : the official journal of the Spanish Society for Microbiology* 16 (3 2013), págs. 133-143. ISSN: 1139-6709. DOI: 10.2436/20.1501.01.188.

[35] Gregg JL, Hershberger PK, Neat AS, Jayasekera HT, Ferguson JA, Powers RL et al. A phylogeny based on cytochrome-c oxidase gene sequences identifies sympatric Ichthyophonus genotypes in the NE Pacific Ocean. En: *Diseases of aquatic organisms* 150 (2022), págs. 61-67. ISSN: 0177-5103. DOI: 10.3354/DAO03677.

[36] Martin WF y Cerff R. Physiology, phylogeny, early evolution, and GAPDH. En: *Protoplasma* 254 (5 sep. de 2017), págs. 1823-1834. ISSN: 0033183X. DOI: 10.1007/S00709-017-1095-Y.

[37] Liu J, Liu H y Zhang H. Phylogeny and evolutionary radiation of the marine mussels (Bivalvia: Mytilidae) based on mitochondrial and nuclear genes. En: *Molecular phylogenetics and evolution* 126 (sep. de 2018), págs. 233-240. ISSN: 1095-9513. DOI: 10.1016/J.YMPEV.2018.04.019.

[38] Coatham SJ, Sellers WI y Püschel TA. Convex hull estimation of mammalian body segment parameters. En: *Royal Society open science* 8 (6 jun. de 2021). ISSN: 2054-5703. DOI: 10.1098/RSOS.210836.

[39] Dou Z, Xin S, Xu R, Xu J, Zhou Y, Chen S et al. Top-Down Shape Abstraction Based on Greedy Pole Selection. En: *IEEE transactions on visualization and computer graphics* 27 (10 oct. de 2021), págs. 3982-3993. ISSN: 1941-0506. DOI: 10.1109/TVCG.2020.2995495.

[40] Zwierzyński M. The improved isoperimetric inequality and the Wigner caustic of planar ovals. En: *Journal of Mathematical Analysis and Applications* 442 (2 oct. de 2016), págs. 726-739. ISSN: 0022-247X. DOI: 10.1016/J.JMAA.2016.05.016.

[41] Foroutan-pour K, Dutilleul P y Smith DL. Advances in the implementation of the box-counting method of fractal dimension estimation. En: *Applied Mathematics and Computation* 105 (2-3 nov. de 1999), págs. 195-210. ISSN: 0096-3003. DOI: 10.1016/S0096-3003(98)10096-6.

[42] Grassberger P. Generalized dimensions of strange attractors. En: *Physics Letters A* 97 (6 sep. de 1983), págs. 227-230. ISSN: 0375-9601. DOI: 10.1016/0375-9601(83)90753-3.

[43] Posadas AND, Giménez D, Bittelli M, Vaz CMP y Flury M. Multifractal characterization of soil particle-size distributions. En: *Soil Science Society of America Journal* 65 (5 sep. de 2001), págs. 1361-1367. ISSN: 0361-5995. DOI: 10.2136/SSSAJ2001.6551361X.

[44] Tamura K, Stecher G y Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. En: *Molecular biology and evolution* 38 (7 jul. de 2021), págs. 3022-3027. ISSN: 1537-1719. DOI: 10.1093/MOLBEV/MSAB120.

[45] Thompson JD, Higgins DG y Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. En: *Nucleic acids research* 22 (22 nov. de 1994), págs. 4673-4680. ISSN: 0305-1048. DOI: 10.1093/NAR/22.22.4673.

[46] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. En: *Nucleic acids research* 32 (5 2004), págs. 1792-1797. ISSN: 1362-4962. DOI: 10.1093/NAR/GKH340.

[47] Alix B, Boubacar DA y Vladimir M. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. En: *Nucleic acids research* 40 (Web Server issue jul. de 2012). ISSN: 1362-4962. DOI: 10.1093/NAR/GKS485.

[48] Gdansk University of Technology. Visual TreeCmp. Programa de comparación de árboles filogenéticos en tiempo polinómico. Software. 2019.

[49] Bogdanowicz D, Giaro K y Wróbel B. TreeCmp: Comparison of trees in polynomial time. En: *Evolutionary Bioinformatics* 2012 (8 ago. de 2012), págs. 475-487. ISSN: 11769343. DOI: 10.4137/EBO.S9657/ASSET/IMAGES/LARGE/10.4137_EBO.S9657-FIG4.JPEG.

[50] Tang Y, Liu G, Zhao S, Li K, Zhang D, Liu S et al. Major Histocompatibility Complex (MHC) Diversity of the Reintroduction Populations of Endangered Przewalski's Horse. En: *Genes* 13 (5 mayo de 2022). ISSN: 2073-4425. DOI: 10.3390/GENES13050928.

[51] Young DS, Chen X, Hewage DC y Nilo-Poyanco R. Finite mixture-of-gamma distributions: estimation, inference, and model-based clustering. En: *Advances in Data Analysis and Classification* 13 (4 dic. de 2019), págs. 1053-1082. ISSN: 18625355. DOI: 10.1007/S11634-019-00361-Y/TABLES/9.

[52] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. En: *Journal of molecular evolution* 16 (2 jun. de 1980), págs. 111-120. ISSN: 0022-2844. DOI: 10.1007/BF01731581.

[53] Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. En: *Molecular Biology and Evolution* 9 (4 jul. de 1992), págs. 678-687. ISSN: 0737-4038. DOI: 10.1093/OXFORDJOURNALS.MOLBEV.A040752.

[54] Tamura K y Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. En: *Molecular Biology and Evolution* 10 (3 mayo de 1993), págs. 512-526. ISSN: 0737-4038. DOI: 10.1093/OXFORDJOURNALS.MOLBEV.A040023.

[55] Mizuta S. Graphical Representation of Biological Sequences. En: *Bioinformatics in the Era of Post Genomics and Big Data* (jun. de 2018). DOI: 10.5772/INTECHOPEN.74795.

[56] Huttenlocher DP, Klanderman GA y Rucklidge WJ. Comparing Images Using the Hausdorff Distance. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15 (9 1993), págs. 850-863. ISSN: 01628828. DOI: 10.1109/34.232073.

[57] Saitou N y Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. En: *Molecular biology and evolution* 4 (4 1987), págs. 406-425. ISSN: 0737-4038. DOI: 10.1093/OXFORDJOURNALS.MOLBEV.A040454.

[58] Kim J, Rohlf FJ y Sokal RR. The Accuracy of Phylogenetic Estimation Using the Neighbor-Joining Method. En: *Evolution* 47 (2 abr. de 1993), pág. 471. ISSN: 00143820. DOI: 10.2307/2410065.

[59] Robinson DF y Foulds LR. Comparison of phylogenetic trees. En: *Mathematical Biosciences* 53 (1-2 feb. de 1981), págs. 131-147. ISSN: 0025-5564. DOI: 10.1016/0025-5564(81)90043-2.

[60] Böcker S, Canzar S y Klau GW. The generalized Robinson-Foulds metric. En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8126 LNBI (2013), págs. 156-169. ISSN: 03029743. DOI: 10.1007/978-3-642-40453-5_13/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-642-40453-5_13.

[61] Shannon CE. A Mathematical Theory of Communication. En: *The Bell System Technical Journal* 27 (1948), págs. 623-656.

[62] Sherwin WB. entropy Entropy, or Information, Unifies Ecology and Evolution and Beyond. En: *Entropy* 20 (2018), pág. 727. DOI: 10.3390/e20100727.

[63] Kullback S y Leibler RA. On Information and Sufficiency. En: *The Annals of Mathematical Statistics* 22 (1 mar. de 1951), págs. 79-86. ISSN: 0003-4851. DOI: 10.1214/AOMS/1177729694.

[64] Bonnici V. Kullback-Leibler divergence between quantum distributions, and its upper-bound. En: (2020).

[65] Keselman HJ y Rogan JC. The Tukey multiple comparison test: 1953-1976. En: *Psychological Bulletin* 84 (5 sep. de 1977), págs. 1050-1056. ISSN: 00332909. DOI: 10.1037/0033-2909.84.5.1050.

[66] Sampson JN, Boca SM, Moore SC y Heller R. FWER and FDR control when testing multiple mediators. En: *Bioinformatics (Oxford, England)* 34 (14 jul. de 2018), págs. 2418-2424. ISSN: 1367-4811. DOI: 10.1093/BIOINFORMATICS/BTY064.

[67] Hanusz Z, Tarasinska J y Zielinski W. Shapiro–Wilk Test with Known Mean. En: *Revstat-Statistical Journal* 14 (1 feb. de 2016), 89–100-89–100. ISSN: 2183-0371. DOI: 10.57805/REVSTAT.V14I1.180. URL: https://revstat.ine.pt/index.php/REVSTAT/article/view/180.

[68] Bawono P, Dijkstra M, Pirovano W, Feenstra A, Abeln S y Heringa J. Multiple Sequence Alignment. En: *Methods in molecular biology (Clifton, N.J.)* 1525 (2017), págs. 167-189. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-6622-6_8.

[69] Seo TK. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. En: *Molecular biology and evolution* 25 (5 mayo de 2008), págs. 960-971. ISSN: 1537-1719. DOI: 10.1093/MOLBEV/MSN043.

[70] Ray GL, Christensen MH y Pinson P. Detection and characterization of domestic heat pumps. En: *2019 IEEE Milan PowerTech, PowerTech 2019* (jun. de 2019), pág. 8810930. DOI: 10.1109/PTC.2019.8810930.

Capítulo A. Anexo de cuadros

Cuadro A.1: Tabla de tamaños de secuencias de ARNr 16S de la base de datos. La longitud de la secuencia se denota con n.

ARNr 16S	**Rango** (n)	**Media** (n)	**Varianza** (n)	**Número de secuencias**
Animales	1016-1219	1112.14	2993.18	1615
Arqueas	1017-1219	1116.95	4937.11	81
Bacterias	1066-1168	1116.78	871.12	1354
Plantas	990-1319	1145.69	10418.21	48
Protistas	971-1320	1164.86	9156.93	81

Cuadro A.2: Tabla de tamaños de secuencias de *cox1* y *cytb* de la base de datos. La longitud de la secuencia se denota con n

Genes	$\mathbf{D_{-80,80}}$ (n)	**Media** (n)	**Varianza** (n)	**Número de secuencias**
cox1	684-1319	1586.9	21574.51	542
cytb	469-1991	1171.54	6801.18	2754

Cuadro A.3: Tabla de tamaños de secuencias genes de animales y ARNm de la base de datos. La longitud de la secuencia se denota con n

ARNm	$\mathbf{D_{-80,80}}$ (n)	**Media** (n)	**Varianza** (n)	**Número de secuencias**
Animales	813-1320	1075.02	21041.13	1999

Cuadro A.4: Tabla de tamaños de conglomerado de secuencias de animales, bacterias y plantas de la base de datos. La longitud de la secuencia se denota con n

Conglomerado	$\mathbf{D_{-80,80}}$ (n)	**Media** (n)	**Varianza** (n)	**Número de secuencias**
Animales	813-1319	1024.13	19487.42	836
Bacterias	813-1320	1079.77	23369.86	1132
Plantas	814-1320	1069.24	18929.73	454

Cuadro A.5: Tabla que recoge el género y especie de las secuencias homólogas de genes empleados en el **objetivo 2**.

citocromo C	***gapdh***	***h3***	***hsp90*α**	**ARNr 18S**
Bos taurus	Apus apus	Canis lupus familiaris	Alligator mississippiensis	Bos taurus
Chaetura pelagica	Canis lupus familiaris	Corvus brachyrhynchos	Bubalus bubalis	Drosophila melanogaster
Equus caballus	Danio rerio	Equus caballus	Canis lupus familiaris	Equus caballus
Gallus gallus	Gallus gallus	Erinaceus europaeus	Equus caballus	Gallus gallus
Homo sapiens	Homo sapiens	Ficedula albicollis	Gallus gallus	Homo sapiens
Meleagris gallopavo	Mus musculus	Gallus gallus	Homo sapiens	Mus musculus
Mus musculus	Oryctolagus cuniculus	Homo sapiens	Mus musculus	Prosopis cineraria
Pan troglodytes	Rattus norvegicus	Mus musculus	Ovis aries	Rattus norvegicus
Rattus norvegicus	Sus scrofa	Pan troglodytes	Rattus norvegicus	Sus scrofa
Sus scrofa	Taeniopygia guttata	Pygoscelis adeliae	Sus scrofa	Zea mays

Cuadro A.6: Tabla que recoge los resultados de las pruebas ANOVA del **Experimento 1** y **Experimento 2**. ANOVA realizada con paquete `Scipy`

Característica	**Experimento 1**	**Experimento 2**	**ANOVA**
$\mathbf{D_0}$	194.4 1.03E-148	33.93 2.97E-15	E_{obs} Valor p
$\mathbf{D_1}$	149.27 8.78E-117	22.09 3.14E-10	E_{obs} Valor p
$\mathbf{D_{-80,80}}$	131.03 1.04E-103	26.39 4.59E-12	E_{obs} Valor p
$\mathbf{C_{iso}}$	1115.89 0	96.28 5.85E-41	E_{obs} Valor p

Cuadro A.7: Tabla que recoge los resultados de prueba Tukey del **Experimento 1**. Caract. denota la característica estudiada, IC el intervalo de confianza, y μ_1 y μ_2 las medias de las clases. Tukey realiazado con el paquete `statsmodels`. La clases se encuentran abreviadas: Ani-Animales, Arq-Arqueas, Bac-Bacterias, Pla-Plantas y Pro-Protistas.

Caract.	$\mathbf{D_0}$			$\mathbf{D_1}$			$\mathbf{D_{-80,80}}$			$\mathbf{C_{iso}}$		
Tukey	Dif.	IC	$\mathbf{v_P}$	Dif.	IC	$\mathbf{v_P}$	Dif.	IC	$\mathbf{v_P}$	Dif.	IC	$\mathbf{v_P}$
Ani/Arq	0.069	[0.028-0.109]	0	0.047	[0.015-0.077]	0.0004	0.295	[0.073-0.515]	0.0025	0.014	[0.008-0.019]	0
Ani/Bac	0.118	[0.104-0.131]	0	0.081	[0.071-0.091]	0	0.534	[0.463-0.605]	0	0.04	[0.038-0.041]	0
Ani/Pla	0.178	[0.126-0.230]	0	0.122	[0.082-0.162]	0	0.696	[0.413-0.980]	0	0.045	[0.038-0.052]	0
Ani/Pro	0.238	[0.197-0.279]	0	0.148	[0.117-0.179]	0	1.002	[0.781-1.223]	0	0.071	[0.065-0.076]	0
Arq/Bac	0.049	[0.008-0.090]	0.009	0.034	[0.003-0.065]	0.023	0.239	[0.018-0.461]	0.0267	0.026	[0.020-0.031]	0
Arq/Pla	0.11	[0.045-0.174]	0	0.075	[0.026-0.125]	0.0003	0.402	[0.049-0.755]	0.0163	0.032	[0.023-0.040]	0
Arq/Pro	0.169	[0.113-0.225]	0	0.102	[0.059-0.144]	0	0.708	[0.403-1.012]	0	0.057	[0.050-0.065]	0
Bac/Pla	0.06	[0.008-0.113]	0.014	0.041	[0.001-0.081]	0.0398	0.163	[-0.122-0.447]	0.5248	0.006	[-0.002-0.013]	0.19
Bac/Pro	0.12	[0.079-0.16]	0	0.067	[0.036-0.099]	0	0.468	[0.246-0.690]	0	0.031	[0.026-0.037]	0
Pla/Pro	0.06	[-0.005-0.126]	0.087	0.026	[-0.023-0.076]	0.5968	0.306	[-0.047-0.659]	0.1259	0.026	[0.017-0.035]	0

Cuadro A.8: Tabla que recoge los resultados de prueba Tukey del **Experimento 2**. Caract. denota la característica estudiada, IC el intervalo de confianza, y dif. la diferencia de medias de las clases y v_p el valor p. Tukey realiazado con el paquete `statsmodels`. La clases se encuentran abreviadas: Ani-Animales, Bac-Bacterias y Pla-Plantas.

Caract.	$\mathbf{D_0}$			$\mathbf{D_1}$			$\mathbf{D_{-80,80}}$			$\mathbf{C_{iso}}$		
Tukey	Dif.	IC	v_p	Dif.	IC	v_p	Dif.	IC	v_p	Dif.	IC	v_p
Ani/Bac	-0.052	[-0.067-(-0.04)]	0	-0.033	[-0.045-(-0.02)]	0	-0.23	[-0.31-(-0.15)]	0	-0.014	[-0.017-(-0.01)]	0
Ani/Pla	-0.046	[-0.07-(-0.03)]	0	-0.027	[-0.04-(-0.012)]	0.0001	-0.17	[-0.27-(-0.08)]	0.0001	-0.018	[-0.02-(-0.014)]	0
Bac/Pla	0.005	[-0.013-0.024]	0.78	0.006	[-0.01-0.02]	0.63	0.062	[-0.031-0.15]	0.266	-0.004	[-0.007-0]	0.047

Cuadro A.9: Tabla de resultados de prueba T de Student (T-test) y la prueba de Levene para los contrastes entre los grupos de secuencias de ADN genómico (Geno) y ARNm de animales. Las columnas debajo de cada contraste se refiere a la prueba T de Student. E_{obs} de refiere al estadístico observado para cada prueba. D_{-80} y D_{80} se refiere al rango del espectro multifractal es calculado como la diferencia entre el D_{-80} y D_{80} de cada representación de Nandy.

Caract.	$\mathbf{D_0}$		$\mathbf{D_1}$		$\mathbf{D_{-80,80}}$		$\mathbf{C_{iso}}$	
T-test	$\mathbf{E_{obs}}$	valor p	$\mathbf{E_{obs}}$	valor p	$\mathbf{E_{obs}}$	valor p	$\mathbf{E_{obs}}$	valor p
Geno/ARNm	-2.5565	0.0106	-0.1385	0.8899	-4.1017	4.21665E-05	0.3547	0.7228

Cuadro A.10: Tabla de resultados de prueba T de Student (T-test) y la prueba de Levene para los contrastes entre los grupos de secuencias de ADN de p53 y cytb. Las columnas debajo de cada contraste se refiere a la prueba T de Student. E_{obs} de refiere al estadístico observado para cada prueba. D_{-80} y D_{80} se refiere al rango del espectro multifractal calculado como la diferencia entre el D_{-80} y D_{80} de cada representación de Nandy.

Caract.	$\mathbf{D_0}$		$\mathbf{D_1}$		$\mathbf{D_{-80,80}}$		$\mathbf{C_{iso}}$	
T-test	$\mathbf{E_{obs}}$	valor p	$\mathbf{E_{obs}}$	valor p	$\mathbf{E_{obs}}$	valor p	$\mathbf{E_{obs}}$	valor p
cytb/p53	-10.8778	5.17E-27	-11.6312	1.46E-30	-5.16	2.64E-07	-7.5675	5.10E-14

Cuadro A.11: Tabla que recoge los resultados de la prueba Levene de homogeneidad de varianzas para los **Experimentos 1-4**. E_{obs} de refiere al estadístico observado para cada prueba. D_{-80} y D_{80} se refiere al rango del espectro multifractal calculado como la diferencia entre el D_{-80} y D_{80} de cada representación de Nandy.

Caract.	**Experimento 1**	**Experimento 2**	**Experimento 3**	**Experimento 4**	**Levene**
$\mathbf{D_0}$	71.2762	33.9377	4.3487	88.0263	$\mathrm{E_{obs}}$
	8.36E-58	6.60405E-12	0.0371	1.31E-20	Valor p
$\mathbf{D_1}$	66.7358	22.0902	0.4277	100.4688	$\mathrm{E_{obs}}$
	3.27E-54	1.25133E-10	0.5132	2.99E-23	Valor p
$\mathbf{D_{-80,80}}$	90.568	26.392	0.4442	69.5726	$\mathrm{E_{obs}}$
	5.27E-73	9.21562E-05	0.5051	1.13E-16	Valor p
$\mathbf{C_{iso}}$	96.3046	96.2804	7.8306	83.3219	$\mathrm{E_{obs}}$
	1.94E-77	1.28543E-10	0.0052	1.29E-19	Valor p

Cuadro A.12: Tabla que recoge el estudio de la normalidad de las clases comparadas en los experimentos de caracterización y descritos en el **Apartado 2.1**. Se empleó el test Shapiro del paquete `Scipy` para ello. La clases se encuentran abreviadas: Ani-Animales, Arq-Arqueas, Bac-Bacterias, Pla-Plantas y Pro-Protistas.

Grupo	**Clase**	$\mathbf{D_0}$		$\mathbf{D_1}$		$\mathbf{D_{-80,80}}$		$\mathbf{C_{iso}}$	
Shapiro		E_{obs}	Valor p	E_{obs}	Valor p	E_{obs}	Valor p	E_{obs}	Valor p
ARNr 16S	Ani	0.909	3.30E-29	0.922	2.48E-27	0.978	4.81E-15	0.956	1.08E-21
	Arq	0.956	0.0072	0.951	0.0036	0.985	4.71E-01	0.916	5.74E-05
	Bac	0.925	3.03E-25	0.930	1.58E-24	0.964	7.70E-18	0.981	3.34E-12
	Pla	0.927	0.0051	0.899	0.0006	0.952	0.0476	0.964	0.1444
	Pro	0.846	1.16E-07	0.848	1.41E-07	0.980	0.202	0.881	1.82E-06
Cong.	Ani	0.957	7.06E-15	0.956	3.69E-15	0.934	8.60E-19	0.938	3.75E-18
	Bac	0.873	2.72E-28	0.847	1.30E-30	0.922	9.40E-24	0.844	5.68E-32
	Pla	0.869	6.03E-19	0.878	3.02E-18	0.951	4.56E-11	0.833	2.11E-21
mRNA	Ani	0.952	1.01E-24	0.962	3.78E-22	0.936	1.66E-28	0.930	1.33E-29
Genes	p53	0.901	2.09E-06	0.889	6.41E-07	0.937	1.97E-04	0.656	7.93E-14
	cytb	0.871	2.07E-42	0.897	5.63E-39	0.959	4.42E-27	0.818	0

Cuadro A.13: Se muestran los algoritmos de alineamientos empleados sobre las secuencias, así como el mejor modelo de sustitución para ese conjunto de datos. El parámetro de la función Gamma, denotado con α **(Sección 2.3)**. G implica que el modelo de sustitución emplea la función Gamma. Cálculos realizados a través de MEGA11

Gen	**Algoritmo de alineamiento**	**Modelo de sutitución**	**Parámetro Gamma** (α)
citocromo C	MUSCLE	K2+G	0.25
gapdg	MUSCLE	T92+G	0.29
h3	MUSCLE	T92+G	0.24
hsp90	MUSCLE	TN93+G	2.22
ARNr18S	ClustalW	K2+G	0.39

Cuadro A.14: Tabla que recoge las distintas tandas de simulación realizadas (S1-S6) para distribuciones P_S con una entropía de Shannon $H(P_S)$ descendiente. R^2 denota el coeficiente de determinación entre la media de $D_{KL}(P_{obs}||P_S)$, siendo P_{obs} la distribución de nucleótidos de cada cadena simulada, y el modelo propuesto en la **Ecuación 3.3**.

Simulación	$\mathbf{P_S}$**=[**$\mathbf{p_A}$**,** $\mathbf{p_T}$**,** $\mathbf{p_C}$**,** $\mathbf{p_G}$**]**	**H(**$\mathbf{P_S}$**)**	$\mathbf{R_2}$
S1	[0.2500, 0.2500, 0.2500, 0.2500]	2	0.987
S2	[0.5000, 0.2500, 0.1250, 0.1250]	1.75	0.984
S3	[0.5000, 0.2500, 0.2500, 0.0000]	1.5	0.996
S4	[0.7035, 0.1965, 0.0500, 0.0500]	1.25	0.970
S5	[0.5000, 0.5000, 0.0000, 0.0000]	1	0.999
S6	[0.7853, 0.2147, 0.0000, 0.0000]	0.75	0.998

Cuadro A.15: Tabla que recoge las distintas tandas de simulación realizadas ($P1$ y $P2$) para distribuciones P_S con una entropía de Shannon $H(P_S)$ descendiente.

Simulación	$\mathbf{P_S}$=[$\mathbf{p_A}$, $\mathbf{p_T}$, $\mathbf{p_C}$, $\mathbf{p_G}$]	**H($\mathbf{P_S}$)**
P1	[0.5000, 0.2500, 0.2500, 0.0000]	1.500
P2	[0.6400, 0.1200, 0.1200, 0.1200]	1.513

Capítulo B. Anexo de figuras

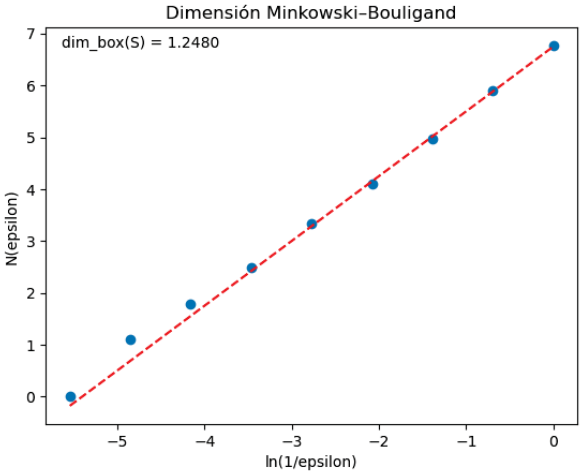


Figura B.1: Estimación de la dimensión Minkowski–Bouligand para la representación de Nandy de la secuencia de DNA de la subunidad delta de la hemoglobina humana (NCBI *Reference Sequence*: NG_063112.2)

Figura B.2: Árboles filogenéticos creados con 10 genes homólogos de ARNr18s. La distancia entre cada pareja de clados ha sido calculada sobre conjunto de puntos obtenidos a partir de representación de Nandy. A la izquierda la función empleada es d_{SM} y a la derecha d_{Sm}. Los recuadros negros señala los taxones *Drosophila melanogaster*, *Prosopis cineraria* y *Zea mays*

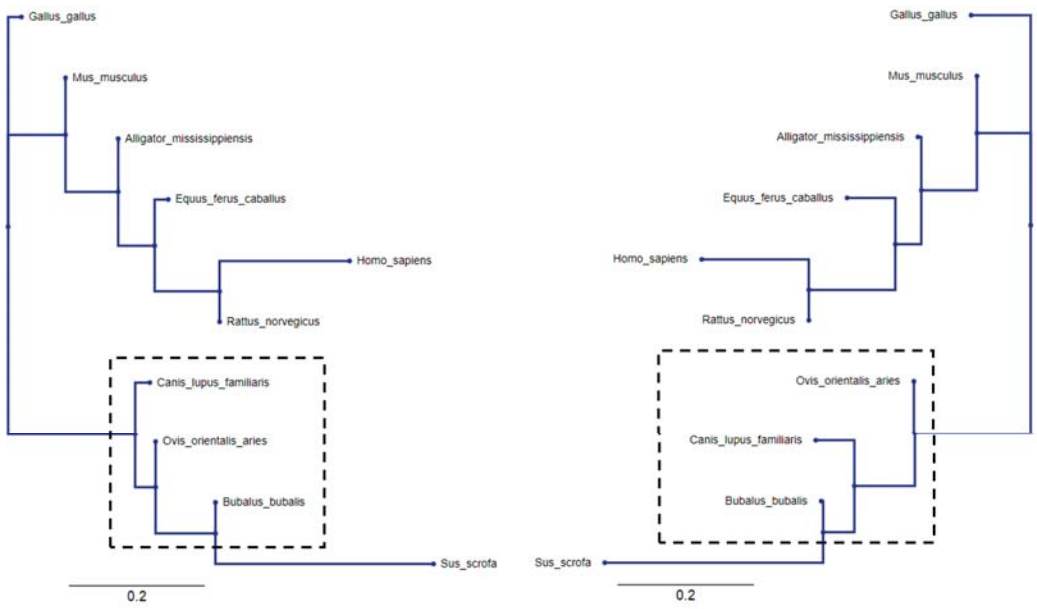


Figura B.3: Árboles filogenéticos creados con 10 genes homólogos de hsp90. La distancia entre cada pareja de clados ha sido calculada sobre conjunto de puntos obtenidos a partir de representación de Nandy. A la izquierda la función empleada es d_{SM} y a la derecha d_{Sm}. Los recuadros negros señala los taxones con una posición relativa distinta.

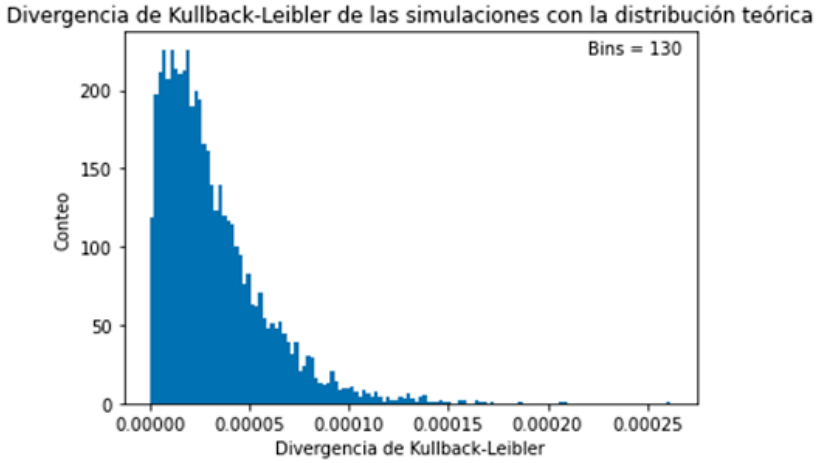


Figura B.4: Histograma de la D_{KL} entre las simulaciones obtenidas a partir de P_S y $P_S = [0{,}5, 0{,}25, 0{,}125, 0{,}125]$. La longitud de las secuencias es de 70 mil nucleótidos. El gráfico emplea 130 intervalos (*bins*). Histograma a partir del paquete `Matplotlib`

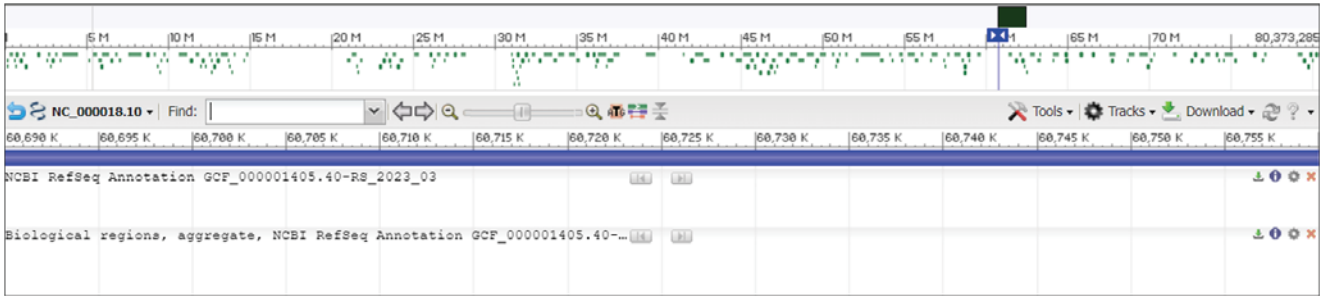


Figura B.5: Visor de NCBI Graphics del segmento correspondiente de la **Figura 15a**.

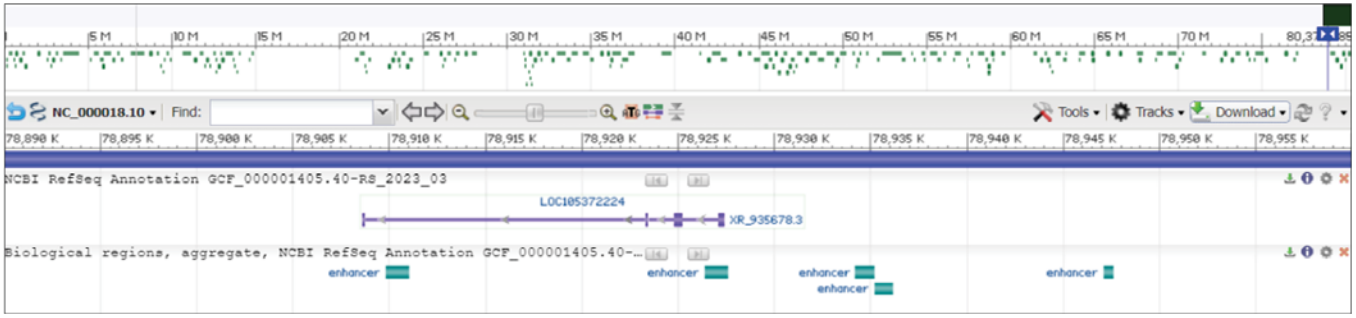


Figura B.6: Visor de NCBI Graphics del segmento correspondiente de la **Figura 15b**.