Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales
Departamento de Ciencias de la Computación

Arquetipos en secuencias biológicas: Estudio del uso de codones y aminoácidos en seres vivos

---

Tesis presentada para optar al título de Magíster en Explotación de Datos y Descubrimiento de Conocimiento

Lucio Aliperti Car

Director:
Dr. Ignacio Enrique Sánchez Miguel

Codirector:
Dr. Pablo Turjanski

Lugar y año:
Buenos Aires, 2021

Índice

Índice de Figuras

Índice de Tablas

1 Agradecimientos

A la Universidad de Buenos Aires.

Al Dr. Ignacio Enrique Sánchez y al Dr. Pablo Turjanski, por aceptar dirigir este trabajo y orientarme en la búsqueda de nuevos interrogantes.

A los docentes y autoridades de la Maestría, por su dedicación y experiencia.

Al Dr. Marcelo Soria, por su orientación y constante predisposición.

A mis compañeros, Florencia y Sergio, por compartir horas de trabajo y buena compañía.

A los integrantes de Laboratorio de Fisiología de Proteínas, por discutir incansablemente perspectivas y resultados.

A mi madre y mi padre, por su apoyo incondicional.

A la familia y amigos, por estar siempre.

A Victoria, por su amor infinito.

2 Resumen / Abstract

2.1 Resumen

En este trabajo se realiza un estudio sobre los sesgos en los usos de codones y aminoácidos en los dominios de la vida. Cada organismo presenta un perfil de abundancias en el uso de codones o aminoácidos propios. Se plantea, entonces, la posibilidad de hallar patrones comunes entre los organismos haciendo foco en organismos extremos o arquetípicos que ayuden a describir las distribuciones de todos ellos. Se aplican distintos algoritmos de búsqueda sobre un gran conjunto de organismos de forma de localizar arquetipos, que representan organismos ideales, sobre los cuales se pueden identificar características propias.

Utilizando bases de datos secundarias que contienen información de secuencias biológicas, se construyen datasets con las abundancias de uso de codones y aminoácidos. Se aplican criterios de selección, de forma de garantizar la representatividad de los datos, y se construyen variables adicionales que puedan caracterizar a los arquetipos hallados desde distintas perspectivas. Se aplican diferentes métodos de búsqueda de arquetipos sobre los mismos datasets de abundancias de manera de seleccionar el modelo más consistente. Se logran hallar 4 arquetipos característicos del conjunto de organismos, a los cuales se le asignaron diferentes propiedades características.

La aplicación de métodos de búsquedas de arquetipos sobre las distribuciones de abundancias relativas de codones y aminoácidos aporta un nuevo enfoque al análisis de los componentes de las secuencias biológicas. La localización de organismos arquetipos favorece la comprensión de los fenómenos en estudio y ayuda a plantear nuevos interrogantes acerca de los sesgos en los usos de codones y aminoácidos.

2.2 Abstract

In this work, an analysis is carried out on codon and aminoacid usage biases across all domains of life. Each organism has an unique profile of codon and aminoacid abundances. Therefore, it is possible to find common patterns among organisms, focusing on extreme or archetypal organisms that can explain their overall distribution. A selection of algorithms was applied on a large set of species in order to find archetypes as ideal organisms, that can be associated with specific characteristics.

Codons and aminoacid abundances datasets were build using secondary databases that contain data from biological sequences. Several selection criteria were applied in order to ensure data representativeness and additional variables were included to characterize archetypes from different approaches. A variety of archetypes searching methods was tested on the same abundances datasets in order to select the more suitable for the analysis. Four archetypes were found and associated with several properties and characteristics.

Applying archetypes searching methods on codon and aminoacid abundances brings a new perspective to biological sequences components analysis. Archetypal organisms identification ease the phenomena understanding and helps to raise new questions about codon and aminoacid usage bias.

3 Introducción

3.1 Complejidad biológica

3.1.1 Biodiversidad

La biodiversidad se define como la variedad que presenta el conjunto de seres vivos en nuestro planeta y evidencia la plasticidad que poseen las estructuras biológicas en diferentes entornos [1]. Actualmente se identificaron más de 1 millón de especies distintas, que si bien poseen rasgos muy diferentes y variados, mantienen principios fundamentales comunes. El 'árbol de la vida' es uno de los principios regentes más importante en el estudio de las ciencias biológicas en donde se condensan los principios que rigen la teoría de la evolución y la existencia de las relaciones de parentesco entre todos los organismos vivos [2] (Fig. 1). Los seres vivos se organizan dentro del árbol agrupados según diferentes criterios que, lejos de permanecer constantes, sufren actualizaciones continuamente en un intento por representar las relaciones de parentesco lo mejor posible. Criterios basados en similitud de rasgos observables de los diferentes organismos fueron dando paso a la construcción de agrupamientos basados en la información genética. Los organismos se agrupan dentro de grupos, denominados taxones, según criterios de similitud. Los diferentes criterios de agrupamiento permiten establecer grupos de organismos que van formando las distintas ramas del árbol dando lugar a relaciones filogenéticas entre ellos [3]. Las relaciones filogenéticas son las relaciones de parentesco que poseen los diferentes taxones o grupos de taxones dentro del árbol. Sobre el modelo teórico del árbol se establecen estas relaciones filogenéticas de los diversos organismos o grupos de organismos partiendo de la base de la existencia de un ancestro común.

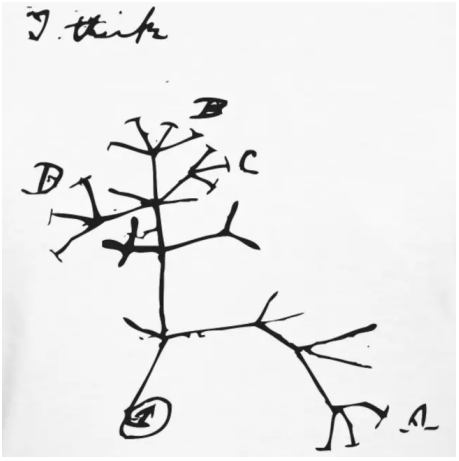

Figura 1: El árbol de la vida según Darwin. Los organismos se agrupan jerárquicamente según criterios variables basándose siempre en la existencia de ancestros comunes para cada uno de los subgrupos. Cada subgrupo forma ramas que a su vez sufre subdivisiones dando origen a la diversidad biológica [4].

La clasificación de los organismos vivos vigente, realizada principalmente a partir de sus moléculas y secuencias biológicas, arroja una relación primaria donde se diferencian 3 grandes dominios capaces de incluir todos los organismo vivos conocidos. Esta categorización logra definir los 3 taxones de mayor jerarquía: Bacteria, Archaea y Eukaryota (también llamado Eukarya) (Fig. 2). En Bacteria se incluyen organismos procariotas (sin núcleo celular), unicelulares y con presencia de diacilglicerol en su membrana; Archaea incluye organismos también procariotas y unicelulares, pero con tetraésteres de diacilglicerol en la membrana celular; y Eukaryota, que presenta organismos con núcleo celular y ésteres de ácidos grasos en su membrana, pudiendo ser uni o pluricelulares [5]. Más allá de estos dominios, diferentes criterios permiten una clasificación en rangos de menor jerarquía de manera de poder llegar a agrupar a los organismos en especies individuales. Por ejemplo, la bacteria Escherichia coli se constituye dentro del siguiente linaje: Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacterales, Enterobacteriaceae, Escherichia, Escherichia coli. Los niveles inferiores a Dominio pueden tener nombres propios según el caso. Típicamente el nivel inmediatamente inferior toma el nombre de Phylum: el Dominio Bacteria, por ejemplo, contiene los Phylum de Acidobacteria, Cianobacteria, Actinobacteria, entre otros.

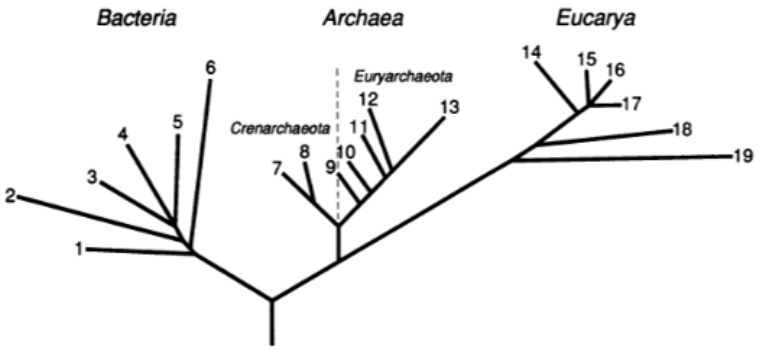

Figura 2: Dominios principales. El consenso actual fija tres dominios principales entre los organismos: Bacteria, Archaea y Eukaryota (también llamado Eukarya). Ésta es la subdivisión de mayor jerarquía dentro de los organismos vivos. Subdivisiones sucesivas de cada subgrupo permiten llegar a cada una de las especies y establecer relaciones de parentesco entre los grupos o especies [5].

3.1.2 La biología y las células

Las células son la unidad básica estructural y funcional de todos los organismos conocidos. Es el lugar donde ocurren las reacciones químicas encargadas del desarrollo y el crecimiento de los seres vivos. Los organismos pueden ser unicelulares, cuando los organismos se componen de una única célula como las bacterias y levaduras, o pluricelulares, cuando se componen de muchas células, como en el caso del ser Humano [1, 6]. Si bien los diferentes organismos poseen células de muy variada forma y función, las funcionalidades básicas son comunes a todos ellos [7]. Todas las células deben poder emplear energía del medio, formar sus estructuras funcionales y transmitir su información genética a la siguiente generación. Estas funciones básicas emplean mecanismos compartidos por todos los organismos que involucran reacciones semejantes. El árbol filogenético refleja estas similitudes en las células de los distintos organismos a la vez que incluye la existencia de una funcionalidad básica común a todos ellos.

3.1.3 La biología y sus moléculas

Las funciones de las células están mediadas por reacciones químicas donde se involucran diversos tipos de moléculas biológicas. Si bien existe una gran diversidad en las moléculas que componen los seres vivos, típicamente se suelen agrupar en 4 grupos generales de macromoléculas (Fig. 3) [7]. Las proteínas son largos polímeros de aminoácidos que pueden o no poseer actividad catalítica y funcionan como enzimas, elementos estructurales, receptores de señal o transportando sustancias

específicas dentro o fuera de las células. Los ácidos nucleicos, ADN y ARN, son polímeros de nucleótidos que almacenan y transmiten información genética y se involucran en la reproducción celular y en la síntesis de las proteínas, entre otras funciones. Los polisacáridos, polímeros de azúcares simples, tienen las funciones de almacenar energía química, formar las estructuras rígidas componentes de las paredes celulares en plantas y en bacterias, y como elementos de reconocimiento en la porción exterior celular. Finalmente, los lípidos sirven como componentes estructurales de las membranas y como combustible rico en energía.

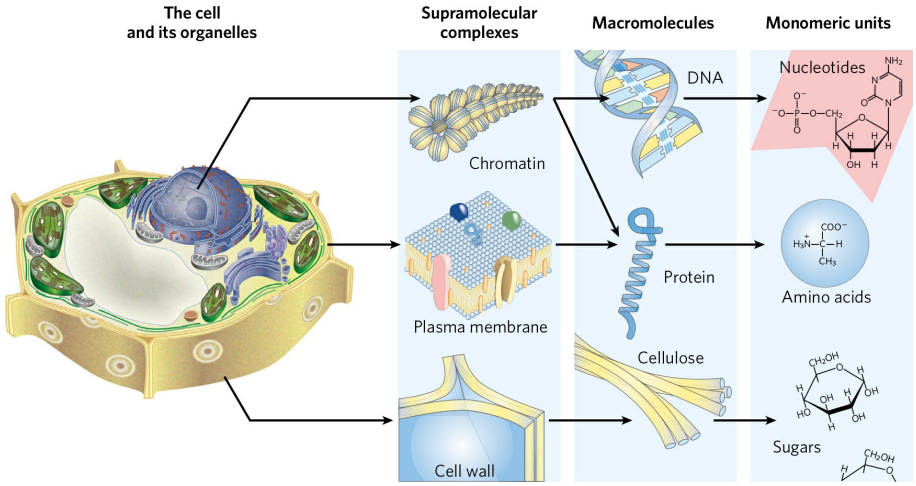

Figura 3: Moléculas biológicas. Las diferentes porciones y estructuras celulares se componen de moléculas diferenciadas. Se muestra una célula vegetal donde se percibe su núcleo celular, su retículo endoplasmático, una vacuola, sus cloroplastos y sus mitocondrias. Se muestran los grupos principales de moléculas biológicas: azúcares, lípidos, aminoácidos y ácidos nucleicos (nucleótidos). Cada uno forma estructuras con complejidad creciente según la cantidad y diversidad de sus componentes. En los casos de azúcares, aminoácidos y ácidos nucleicos, estos diferentes monómeros se concatenan formando moléculas de gran tamaño, que a su vez forman complejos supramoleculares donde se agrupan muchas macromoléculas que interaccionan entre si. Cada macromolécula, a su vez, se compone de uno o más monómeros según el tipo de molécula (polímeros de nucleótidos forman los ácidos nucleicos como el ADN, polímeros de aminoácidos forman las proteínas y polímeros de azúcares forman los polisacáridos como la celulosa). Los lípidos, por otra parte, no se constituyen como polímeros sino que se organizan como grandes complejos supramoleculares. Son los constituyentes principales de las membranas biológicas y componen tanto la membrana celular como las de las diferentes estructuras internas (membrana nuclear, membranas de mitocondrias, membranas de cloroplastos y membranas de retículos endoplasmáticos) [7].

3.2 Secuencias biológicas

3.2.1 El código genético

Los genomas de los diferentes organismos se componen por largas secuencias de nucleótidos y tienen como principal función conservar la información necesaria para el desarrollo de cada uno de ellos. La información se encuentra almacenada como sucesiones de 4 nucleótidos, cada uno conteniendo una base nitrogenada

característica (Adenina, Guanina, Citosina y Timina) [7]. Los genomas pueden contener entre 10^6 (pequeñas bacterias) y 10^{11} (Eukaryota superiores) pares de bases de longitud. Dentro de los genomas existen porciones que son traducidas a proteínas comúnmente denominadas Secuencias Codificantes o CDS ('Coding Sequences') [8]. Además, existen secuencias de nucleótidos que no son traducidas a proteínas. Esta proporción suele ser menor en organismos más simples (Bacterias, Eukaryota unicelulares) pero puede alcanzar grandes proporciones en algunos Eukaryota superiores donde constituyen la mayoría del genoma [9]. Además, existen fenómenos de procesamiento del mRNA denominados 'postranscripcionales' que pueden afectar las porciones de las cadenas de nucleótidos que son efectivamente traducidas a proteínas (típicamente, fenómenos de 'splicing'), por lo que, en la práctica, la relación entre genoma y proteínas está muy lejos de correlacionar directamente [10]. Las proteínas se componen, en una primera aproximación, de cadenas de 20 aminoácidos diferentes con un largo medio de 300 aminoácidos, que pueden combinarse y plegarse para formar sus estructuras funcionales [7, 8].

El 'dogma central' de la biología molecular establece que la información de las secuencias genómicas, compuestas por genes y éstas por codones (tripletes de nucleótidos adyacentes), son capaces de traducirse en proteínas, compuestas por aminoácidos, según las reglas de un código genético prácticamente universal (Fig. 4) [8]. El código genético se presenta como una relación directa entre los grupos de tres nucleótidos y los aminoácidos empleados para las síntesis proteicas [7]. Esta transmisión de información se realiza de manera semejante en todos los organismos vivos dando cuenta de una funcionalidad común, a la vez que se realiza con sutiles variaciones, evidenciando cierta flexibilidad, y capacidad de adaptación [2]. Dentro de cada CDS, podemos suponer una relación entre la presencia de un determinado codón y su traducción a un aminoácido según lo que determina el código genético. Si bien se registran diversas variaciones del código genético (en NCBI se registran hasta 33 códigos diferentes), su funcionamiento es esencialmente el mismo. Más del 99 % de los organismos utiliza un código estándar para la traducción de RNA genómico en su citoplasma, pudiendo haber variaciones en los codones de inicio [11]. Por otro lado, es importante mencionar que existen diferencias en los códigos utilizados en el genoma nuclear y en las mitocondrias. Además, algunos organismos poseen ligeras singularidades en sus códigos genéticos genómicos. En *Candida ssp.*, por ejemplo,

el codón CUG se traduce en Ser en lugar de Leu, a diferencia de como ocurre en la mayoria de los organismos. De cualquier forma, la información necesaria para especificar cada proteína individual, se almacena, en última instancia, en ácidos nucleicos.

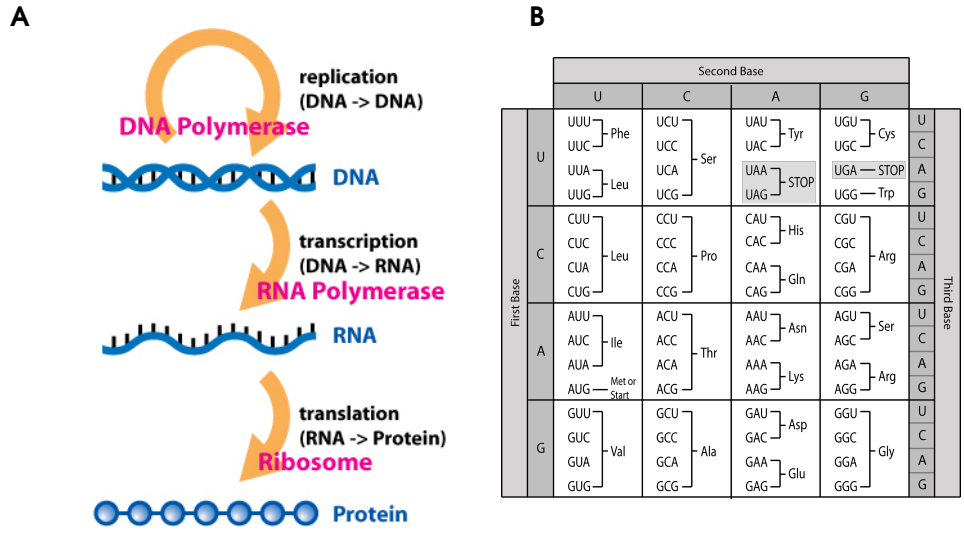

Figura 4: 'Dogma central' de la biología molecular. A. La información fluye desde las secuencias de nucleótidos que conforman el ADN hacia las secuencias de los ARN mensajeros, que transportan la información hasta los ribosomas, estructuras especializadas encargadas de realizar el proceso de traducción, donde se interpretan las secuencias de nucleótidos construyéndose las secuencias de aminoácidos que forman las proteínas [12]. B. Durante el proceso de traducción, cada triplete de nucleótidos se decodifica en un aminoácidos particular según un código genético compartido por la mayoría de los organismos vivos. En este caso, se muestra el código genético comúnmente denominado 'estándar', considerado como código de referencia y sobre el cual pueden existir pequeñas variantes en grupos particulares de organismos. Existiendo 20 aminoácidos (más una señal de 'stop') y 64 codones diferentes, cada aminoácido tiene entre 1 y 6 codones que lo codifican.

3.2.2 Uso de codones

Los genomas de los distintos organismos presentan sesgos en el uso de codones [13]. Dado que, en el código genético, un mismo aminoácido puede ser codificado por más de un codón, los organismos podrían utilizar cualquiera de los codones posibles para un aminoácido dado (Fig. 5A) [14, 15]. Si un organismo determinado necesitase, por ejemplo, incluir una Prolina (Pro) en una secuencia de aminoácidos dada, podría utilizar cualquiera de los codones correspondientes: CCA, CCU, CCG o CCC (Fig. 4B). Sin embargo, se observa que la elección de los codones podría tener funciones asociadas como la regulación de la expresión génica o de la eficiencia de los procesos de traducción. El uso de determinados codones ó aminoácidos por diferentes especies de organismos no sucede de manera azarosa, sino que presenta patrones en el espacio de configuraciones posibles. Los patrones evidencian los sesgos

existentes en el uso de los diferentes codones (Fig. 5B) [16].

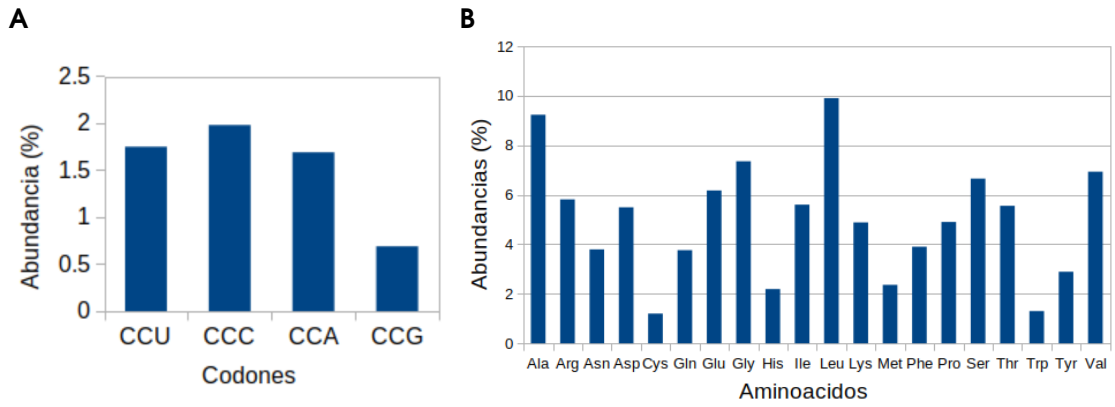

Figura 5: Sesgos de abundancias de codones y aminoácidos. A. Abundancias porcentuales de los codones de la Prolina en el genoma humano. La abundancia de cada codón es diferente para el grupo de codones sinónimos. Esto se repite en las abundancias de los codones de varios aminoácidos. B. Abundancias porcentuales de los aminoácidos en UniProtKB/TrEMBL. Tomando todas las proteínas de la base de datos se calculan las abundancias de aparición de cada aminoácido. Nuevamente, se observa una distribución no uniforme. Tomado de UniProtKB/TrEMBL (`https://www.ebi.ac.uk/uniprot/TrEMBLstats`).

Dado que todos los organismos comparten los mecanismos de transcripción y traducción, es posible postular la existencia de líneas rectoras en las distribuciones de estos componentes que pueden aportar indicios acerca de las razones subyacentes de dichas distribuciones [17, 18]. Como mencionamos previamente, los organismos utilizan algunos codones sinónimos por sobre otros y el uso de los codones no parece ser al azar. El uso de los codones alternativos no se observa de manera uniforme, sino que en la mayoría de los organismos aparece un desbalance [19].

El sesgo en el uso de codones se asocia tradicionalmente con el contenido GC de cada organismo [2, 20, 21, 22]. El contenido GC se calcula para los diferentes genomas, genes, CDS o secuencias en general y mide la proporción de Guanina + Citosina sobre el total de bases. El contenido GC genómico es ampliamente utilizado para explicar el uso de los diferentes codones en los grupos de organismos.

Sin embargo, no existe un consenso general acerca de las líneas rectoras de los sesgos en el uso de codones entre especies ni de las fuerzas evolutivas involucradas en las composiciones de los genomas. Dentro de organismos particulares se observaron correlaciones entre el uso de codones específicos y las abundancia de sus tRNA citosólicos correspondientes [19, 23, 24]. El tRNA (Acido Ribonucleico de transferencia) esta involucrado en el proceso de traducción y tiene unido un codón y su aminoácido correspondiente. En líneas generales, existe un tRNA por cada uno de los codones codificantes. Las abundancias de tRNA son capaces de predecir el uso de

codones en las porciones codificantes de genes de alta expresión [25, 26]. Se observa un fenómeno de coevolución entre el uso de codones en los genes y las abundancias de los diferentes tRNA, dando cuenta de una posible fuerza de selección natural tanto en genes de alta [15] como de baja [27] expresión. Sin embargo, no se alcanza a discernir una relación causal clara entre el uso de codones y las abundancias de tRNA. Las relaciones filogenéticas también han mostrado una relación con el uso de codones. Los grupos de organismos con linajes similares han mostrado una correlación en el uso de codones, si bien esto no permite explicar lo sesgos de codones en taxones filogenéticamente más lejanos [28]. Es posible relacionar las abundancias de codones con los linajes solo en casos de grupos específicos. De hecho, se han logrado realizar clasificaciones de secuencias en grupos taxonómicos a partir de los sesgos en los usos de codones, pero solamente aplicado a grupos de bacterias [29, 30]. Finalmente, se han observado cómo los sesgos en el uso de codones aparecen más marcadamente en las secuencias con alta tasa de expresión. Secuencias más traducidas muestran una selección más fuerte a la hora de emplear los codones, mientras que segmentos que no sufren traducción no tienen un sesgo tan marcado [18]. Nuevamente, se postula una fuerza de selección natural asociada al sesgo de codones en secuencias que sufren el proceso de traducción más intensamente y su relación con la evolución de genomas completos [31, 32, 33]. Se ha demostrado una correlación entre los usos de codones y los distintos fenómenos mencionados, pero esta correlación no implica necesariamente una relación causal.

3.2.3 Secuenciación masiva de genomas

En los últimos 20 años las nuevas tecnologías de secuenciación masiva han logrado generar grandes cantidades de información relativa a secuencias nucleotídicas presentes en una gran cantidad de organismos diferentes. A partir de la secuenciación de segunda generación se llevaron a cabo la identificación de millones de genes de miles de especies, al mismo tiempo que se pusieron a disposición de la comunidad científica mediantes la construcción de bases de datos públicas. Diferentes consorcios mantienen actualmente bases de datos donde se almacena masivamente y se hace pública la información fruto de secuenciaciones. Por ejemplo, Genbank maneja más de 3×10^{12} secuencias biológicas [34, 35] mientras que Refseq mantiene datos de más de 9×10^{8} pares de bases [36].

Las bases de datos previamente mencionadas contienen datos de diversas fuentes, ya sean genes individuales, ensambles de secuencias, genomas completos, genes transcriptos, etc. y pueden tener distinto grado de curado o calidad en sus anotaciones. Además, son de acceso público y pueden utilizarse como fuente de bases de datos secundarias o como insumos de herramientas bioinformáticas de búsqueda o análisis. Si bien son las bases de datos primarias las que almacenan información tomada directamente de observaciones o mediciones, son las bases de datos secundarias las que recopilan información de diferentes bases de datos y las relacionan entre si. Por ejemplo, bases de datos de secuencias nucleotídicas o proteicas suelen ser diferentes pero se integran de manera de poder relacionar genes y proteínas y construir una base de datos secundaria. Así es posible construir herramientas capaces de integrar diferentes fuentes de información con objetivos más o menos específicos, según el caso, y facilitar la labor de análisis o búsqueda de datos por parte del usuario. En este trabajo se utilizaron datos provenientes de CoCoPUTs[1] [37], una base de datos secundaria pública que relaciona y procesa las secuencias existentes en Genbank y Refseq, dándoles una estructura diferente.

3.3 Arquetipos y Fenotipos

3.3.1 Optimización de Fenotipos

Cada organismo contiene en su genomas información que, en última instancia, influirá en la definición de su estructura y sus características. En líneas generales, esta información propia toma el nombre de 'genotipo', mientras que lo que se manifiesta como morfología o función se denomina 'fenotipo' [38]. Los diferentes sistemas biológicos, a lo largo de su recorrido evolutivo, incrementan su éxito reproductivo tanto como sea posible. El éxito reproductivo mide la capacidad de los organismos de producir descendencia. El éxito reproductivo depende, en general, de poder cumplir con varios objetivos (o tareas) simultáneamente, por lo que los diferentes fenotipos tienden a enfrentar una situación fundamental: un fenotipo dado suele no ser óptimo en todas las tareas que realiza [39, 40, 41, 42]. Dado un conjunto de características o tareas a optimizar, un frente de Pareto se forma cuando no es posible mejorar una de ellas sin ocasionar una merma en alguna de las otras. La presión de selección tiende a deshacerse de fenotipos no óptimos seleccionando mayoritariamente fenotipos que

[1] CoCoPUTs: `https://hive.biochemistry.gwu.edu/review/codon2`

logren un conjunto de características compatibles con el frente de Pareto [40, 43] (Fig. 6). Es por esto que los diferentes fenotipos muestran diferentes estrategias surgidas de combinaciones de rasgos o características que les permiten desarrollarse y reproducirse en un entorno dado con un elevado éxito reproductivo.

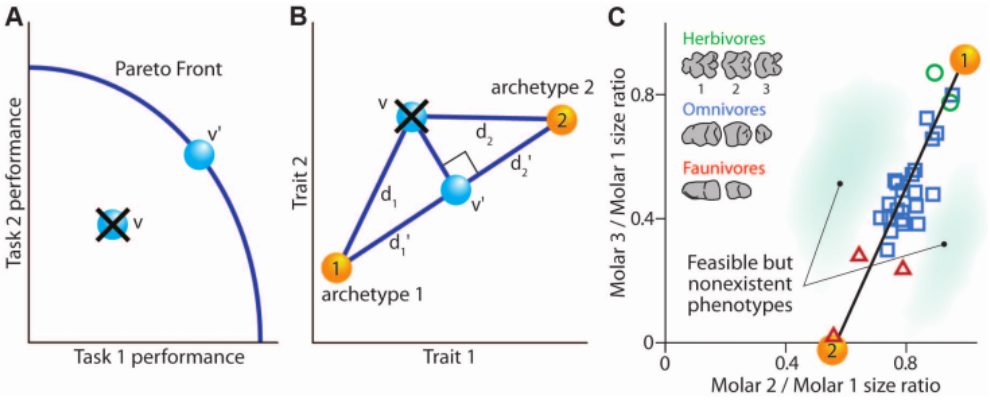

Figura 6: Arquetipos como organismos especializados. A. Dadas dos tareas diferentes, los organismos intentan alcanzar la máxima performance en ambas al mismo tiempo. El frente de Pareto aparece cuando la mejora en alguna de las tareas trae aparejada una disminución en la performance de la otra. Los arquetipos se conciben como organismos especializados idealmente en alguna tarea en particular. B. Dado un conjunto de rasgos propios de los fenotipos, se pueden definir combinaciones de rasgos con éxito reproductivo óptimo y combinaciones de rasgos con éxito reproductivo no óptimo. Cada una de estas combinaciones definen diferentes zonas dentro de los múltiples valores de cada rasgo. Las zonas de éxito reproductivo óptimo tienen su correlación con las eficiencia de las tareas que se ubican dentro del frente de Pareto. La forma del conjunto de individuos observables, considerados como con éxito reproductivo óptimo, determina los vértices que pueden considerarse como organismos especializados para tareas particulares (arquetipos). C. El conjunto de organismos observados puede considerarse como una suma ponderada de estos arquetipos. Los organismos observables pueden pensarse como una combinación lineal de los organismos especializados, donde conforme se acerquen a un arquetipo, su performance en su tarea asociada aumentara a costa de alejarse de otro [41].

Los fenotipos mejor compensados pueden pensarse como sumas ponderadas de fenotipos especializados para objetivos únicos [39]. Estos fenotipos especializados se conforman como entes potencialmente ideales con características particulares asociadas, directa o indirectamente, a las tareas que otorgan el grado de especialización [41] (Fig. 6). Los arquetipos funcionan como organismos con propiedades especificas. Sus propiedades deberán colocarlos en los extremos del grupo de organismos observados considerados como fenotipos optimizados en cuanto a su éxito reproductivo. Organismos fuera de este perímetro no estarán optimizados y por lo tanto no serán observables (Fig. 6B). Luego, los arquetipos con características específicas podrán asociarse a tareas específicas a partir de las propiedades adicionales de organismos cercanos (Fig. 6C).

3.3.2 Búsqueda de Arquetipos

La búsqueda de arquetipos en disciplinas biológicas se realiza mediante un enfoque general común. Se utiliza un conjunto de instancias o elementos observados sobre los que se pretende hallar arquetipos. En el caso puntual de este trabajo, estos elementos están constituidos por organismos tomados de todas las ramas de la vida. Cada uno de estos elementos posee diferentes características que deben tomar forma de variables numéricas descriptoras. El conjunto de estas variables numéricas continuas conforma un espacio de variables donde es posible ubicar a los distintos elementos (Fig. 6C). En este trabajo, las variables están constituidas por las características de los organismos y las frecuencias de uso de codones y aminoácidos, entre otras. Sobre este mismo espacio de variables es posible ubicar a los arquetipos tras la aplicación de los algoritmos de búsqueda de arquetipos. Otro claro ejemplo lo constituye el análisis de arquetipos sobre perfiles de expresión celular (Fig. 7A), donde las instancias o elementos representan a diferentes células y las variables de características describen perfiles de expresión de varios genes.

El análisis de arquetipos utiliza un conjunto de instancias o elementos situados en un espacio de variables e intenta localizar un número reducido de elementos extremos en el mismo espacio que puedan construir, mediante una combinación lineal, cada uno de los elementos originales [44]. El análisis requiere poder discernir el número de arquetipos para el conjunto de datos y luego ubicarlos en el espacio de forma que sean representativos [41] (Fig. 6). Como entrada se toma un conjunto de instancias o elementos con sus variables asociadas. El conjunto de elementos toma una forma en el espacio de sus variables que se intentará reproducir a partir de un número reducido de arquetipos [45, 46]. Idealmente, cada elemento deberá poder estimarse a partir de la suma ponderada de los arquetipos hallados. Los conjuntos de individuos pueden así considerarse como sumatorias convexas de los arquetipos, tomando distintas proporciones de cada uno de ellos. Una vez localizados los arquetipos, es posible analizar su posición absoluta en el espacio de variables y su ubicación relativa con respecto a los elementos tomados como entrada. Por ejemplo, para buscar arquetipos tomando perfiles de expresión celular (Fig. 7a), cada elemento se constituye por una célula a la cual se le determinaron las expresiones de una serie de genes. Estos niveles de expresión constituyen las variables que describen a cada célula. El conjunto de células, entonces, toma una forma en el espacio de estas variables. Los arquetipos

hallados se ubican en el mismo espacio pudiendo estimarse su nivel de expresión. Una estrategia similar puede emplearse para analizar transcriptomas bacterianos [47].

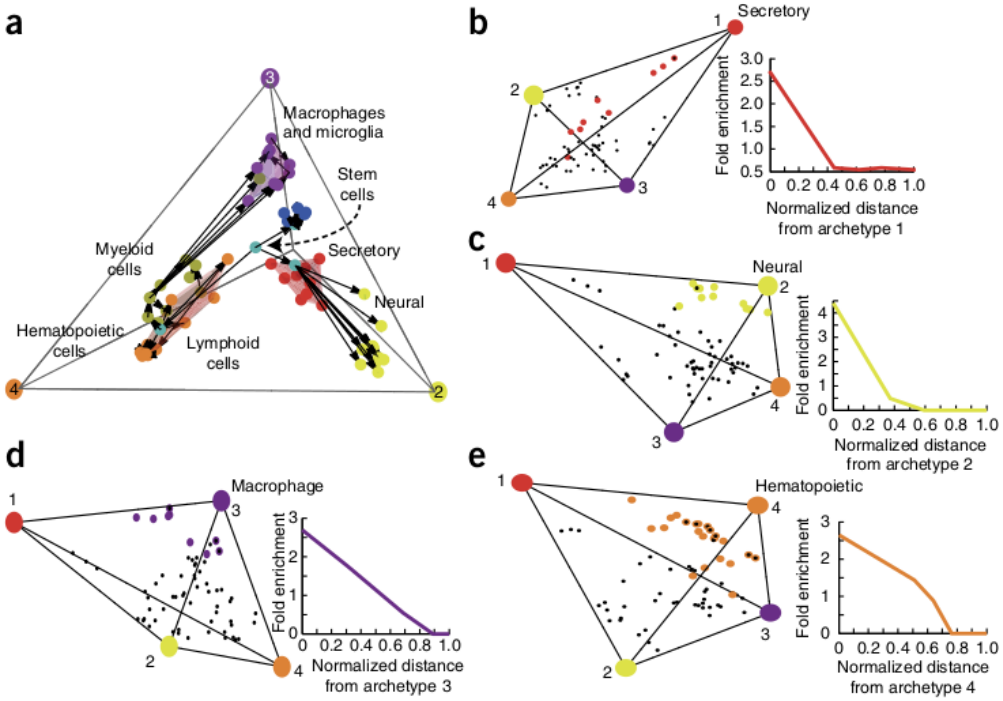

Figura 7: Metodología ParTI. a. Tras realizar una reducción de dimensionalidad, utiliza diferentes algoritmos de búsqueda de arquetipos para comparar un politopo con la superficie convexa del conjunto de elementos. Luego se compara los diferentes elementos con las posiciones de cada arquetipo. En este ejemplo [41], se utiliza la expresión de genes de diferentes células para definir 4 arquetipos de tipos celulares. b, c, d, e. Para cada arquetipo se puede determinar un perfil de distintas características, en este caso, se utiliza la expresión de genes asociados a líneas celulares. Perfiles decrecientes de estas características implican un enriquecimiento en las proximidades de cada arquetipo. Las características enriquecidas permiten asignar propiedades a los arquetipos más próximos.

Diversas metodologías se postularon para hallar arquetipos. PCHA ('Principal Convex Hull Analysis') [48] fue propuesto como un primer algoritmo capaz de ubicar arquetipos que puedan reproducir el resto de los elementos como combinaciones lineales a partir de ajustar una superficie convexa. El método en sí sufrió mejoras y se comprobó su utilidad en diversas áreas como evolución, genética o análisis de textos. Por otro lado, recientes avances han logrado postular enfoques alternativos en donde la búsqueda de arquetipos se realiza a partir de la construcción de redes neuronales y autoencoders con diferentes variantes [49], ampliando la aplicabilidad de los arquetipos a otras áreas como procesamiento de imágenes o volúmenes extensos de datos.

3.3.3 Propiedades de los Arquetipos

Una vez conocidos el número y la localización relativa de los arquetipos al conjunto de organismos, el análisis de sus variables y características pueden ayudar a inferir funciones o tareas. Las propiedades tomadas de las variables sobre las que se realizan las búsquedas de arquetipos son complementadas por propiedades de los organismos que, si bien no se incluyeron en el el análisis inicial, pueden dar cuenta de las características. En este trabajo denominaremos estas propiedades como variables de 'enriquecimiento' en concordancia con los autores del método [41]. Tomando por ejemplo la Fig. 6C, si bien los arquetipos pueden ubicarse a partir de determinaciones morfométricas de seres vivos, sus propiedades más relevantes pueden derivar de sus comportamientos, hábitats o edad: aunque los arquetipos aparecen como fenotipos extremos en un espacio de dos variables de morfología dental, sus características y tareas se denotan al incluir el tipo de dieta de los organismos involucrados. Solo así puede observarse claramente que los arquetipos constituyen organismos especializados en procesar determinados tipos de dieta. En el caso de los perfiles de los niveles de expresión génica (Fig. 7b, c, d y e), las funciones de los arquetipos no surgen directamente de su ubicación en el espacio, sino a partir de los conjuntos de genes con alto nivel de expresión de las células cercanas. Las categorías de los genes con elevado nivel de expresión de las células cercanas a cada arquetipo permiten caracterizar mejor a cada arquetipo y asignarles roles o tareas específicas. Para esto se tiene en cuenta la distancia entre cada célula y cada arquetipo (Fig. 7 b, c, d y e). Un arquetipo rodeado por células con elevada expresión de genes propios de células 'secretoras' puede caracterizarse como un arquetipo con la función 'secretora' (Fig. 7b).

3.4 Objetivos e Hipótesis de trabajo

3.4.1 Objetivo general

La hipótesis de trabajo se basa en que las distribuciones de abundancias de los codones o aminoácidos pueden concebirse como guiadas por objetivos generales. Tomando a los arquetipos como fenotipos extremos, se intentarán describir "objetivos" o "tareas" hacia las cuales se han orientado los diferentes organismos. Esto implicaría que las abundancias de los diferentes elementos no se distribuyen

uniformemente o al azar sino que están, de alguna manera, respetando un óptimo dentro de sus combinaciones posibles. En el presente trabajo se intentarán definir y caracterizar los arquetipos existentes en las distribuciones de codones y aminoácidos de todos los órdenes de la vida, como medio para poder hallar líneas rectoras que ayuden a explicar su variabilidad.

3.4.2 Objetivos específicos

Se plantean los siguientes objetivos específicos:

- **Entendimiento de los Datos**

  Recolección de datos. Obtener una base de datos de abundancias de codones y aminoácidos de una gran cantidad de organismos diferentes con gran cantidad de variables descriptoras.

  Descripción de datos. Analizar y describir las características principales de la base de datos de trabajo.

- **Preparación de datos**

  Procesamiento de bases de datos. Procesar y seleccionar un grupo de los datos de origen de forma tal de mejorar su confianza y, a su vez, intentar alcanzar una buena representatividad.

  Armado de datasets de trabajo. Construir los datasets de trabajo a partir de las bases de datos de abundancias y el dataset de enriquecimiento a partir de variables adicionales.

- **Modelado**

  Elección de modelo. Evaluar métodos de localización de arquetipos y seleccionar el más adecuado para la búsqueda de arquetipos en nuestro dominio de trabajo.

  Búsqueda de arquetipos. Aplicar el modelo elegido sobre los datasets de trabajo de forma de hallar los arquetipos que los caracterizan.

- **Evaluación**

  Análisis de localización de arquetipos. Analizar las ubicaciones de los arquetipos hallados y las relaciones entre ellos.

Caracterización de arquetipos. Caracterizar los arquetipos hallados a partir de las variables de enriquecimiento.

4 Métodos

4.1 Flujo general de trabajo

Este trabajo se realiza dentro de un flujo general de tareas que abarcan desde el conocimiento del tema y la bibliografía relacionada con sus temas específicos hasta la elaboración de la presente tesis. Se recorren diferentes etapas donde se van desarrollando las actividades necesarias para la obtención tanto de resultados y conclusiones consistentes, como de nuevas perspectivas para el enfoque del problema planteado. A efectos de poder analizar mejor las etapas de trabajo, se toma el marco teórico descripto a partir de la metodología CRISP-DM ('Cross Industry Standard Process for Data Mining') [50]. CRISP-DM presenta un método estandarizado utilizado para describir los ciclos de vida de proyectos de 'Data Mining' (Fig. 8). Si bien plantea seis etapas bien delimitadas, es lo suficientemente flexible como para poder adaptarlo a múltiples tareas. Se describen, entonces, seis etapas generales sobre las cuales se enmarcan las diferentes actividades de forma de poder ordenar las tareas llevadas a cabo [51].

La etapa de Entendimiento de Dominio (Etapa 1) involucró la investigación de las temáticas asociadas al tópico de estudio y al problema a resolver. Conocer los aspectos generales del uso de los codones y aminoácidos en los diferentes seres vivos y su dinámica fueron el puntapié inicial para plantear los primeros interrogantes y elegir las herramientas adecuadas. Además, se orientó la búsqueda bibliográfica hacia los métodos de análisis de arquetipos y su marco teórico subyacente. Se plantearon las preguntas iniciales y se establecieron los objetivos generales: hallar y caracterizar los arquetipos que pudieran surgir de los usos relativos de los conjuntos de codones y aminoácidos en todas las ramas de la vida. A partir de la información disponible, se plantearon las etapas subsiguientes. En la sección de Introducción de este trabajo se describen los aspectos más importantes de esta etapa. La aplicabilidad de las búsquedas de arquetipos al problema del uso de codones sirvió como guía en esta primera etapa.

El Entendimiento de los Datos (Etapa 2) incluyó la recolección y el análisis de los datos disponibles de diferentes fuentes. La amplia diversidad de bases de datos provee abundante información que debe pasar por un proceso posterior de modo de mejorar la calidad de los datos para resolver el problema planteado. Se

buscaron bases de datos de abundancias de codones con información confiable y suficiente para el análisis. La búsqueda se extendió a variables de enriquecimiento, es decir, a propiedades adicionales características de organismos vivos que puedan usarse para caracterizar a los arquetipos. La exploración inicial permitió conocer las características generales de los datos y seleccionar aquellas más adecuadas para responder los interrogantes planteados. Se detalla esta etapa en la sección 4.2.

La Preparación de los Datos (Etapa 3) surge como una etapa clave dada la vasta cantidad de secuencias biológicas existentes. Se seleccionaron bases de datos secundarias capaces de simplificar el enfoque y se aplicaron procesamientos sucesivos sobre los datos a fin construir los datasets de trabajo. Se buscó disponer de la mayor cantidad de datos posibles con buena representatividad. Se construyeron los datasets usados tanto como entrada en la búsqueda de arquetipos como los datasets de variables de enriquecimiento. Se detalla esta etapa en las secciones 4.3, 5.1 y 5.2.

La etapa de Modelado (Etapa 4) consistió en evaluar y seleccionar un método idóneo de búsqueda de arquetipos. Una vez seleccionado un método, se realizó la búsqueda de arquetipos sobre los dataset de trabajo ya construidos. Esta etapa se detalla en las secciones 4.4, 5.3 y 5.4.

Durante la Evaluación (Etapa 5) de los resultados obtenidos se caracterizaron los arquetipos hallados mediante diferentes enfoques. Se tomó la localización de cada uno de los arquetipos y se incluyeron las variables adicionales de enriquecimiento surgidas de bases de datos adicionales. Esta etapa se detalla en las secciones 4.5, 5.5 y 5.6.

Finalmente, en la etapa de Implementación (Etapa 6), se discuten los resultados y las conclusiones, se analizan los próximos pasos y se proponen nuevas preguntas. Esta etapa se incluye en la sección de Discusión.

Etapa 1	Etapa 2	Etapa 3	Etapa 4	Etapa 5	Etapa 6
Entendimiento de Dominio	Entendimiento de los Datos	Preparación de Datos	Modelado	Evaluación	Implementación
Objetivo general.	Recolección de datos.	Procesamiento de base de datos.	Elección de modelo.	Análisis de localización de arquetipos.	Evaluación final.
Metas específicas.	Descripción de datos.	Construcción de datasets de trabajo.	Búsqueda de arquetipos.	Caracterización de arquetipos.	Nuevas hipótesis.

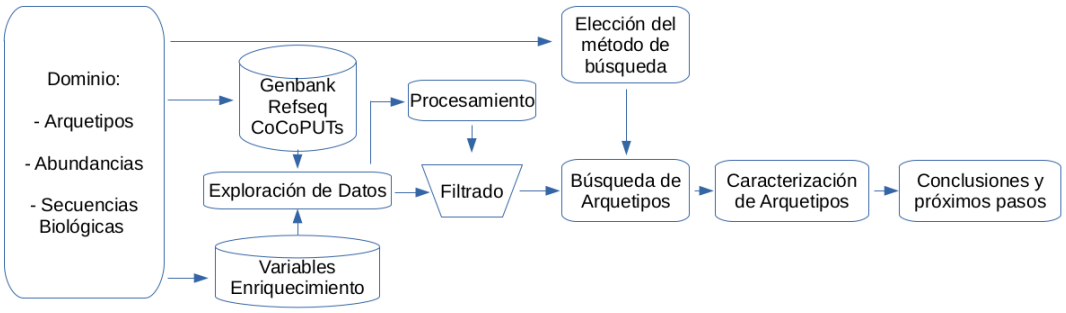

Figura 8: Flujo general de trabajo. A partir de la metodología CRISP-DM, se plantean las etapas de este trabajo. Se encuadran las actividades principales a modo de mapa general.

4.2 Entendimiento de los Datos

4.2.1 Recolección de datos

Los datos en los que se basa el presente trabajo fueron tomados de la base situada en CoCoPUTs[2] [37] y descargada en Abril de 2019 (versión o576245). La base de datos recopila los datos presentes en las bases de datos primarias Refseq y Genbank, que son repositorios donde se almacenan secuencias biológicas. Genbank es un repositorio público originalmente diseñado para almacenar secuencias de nucleótidos fruto de secuenciaciones masivas [35, 34]. Cualquier autor puede cargar sus secuencias en esta base de datos, por lo que esta base de datos no tiene un proceso de curado. Las secuencias almacenadas pueden estar anotadas detalladamente o bien pueden ser fruto de secuenciaciones de muestras complejas, por lo que pueden tener poca o nula anotación. Genbank pertenece al Instituto Nacional de Salud de los Estados Unidos y está mantenido por el 'National Center for Biotechnology Information' (NCBI). Refseq también depende del NCBI pero funciona como un registro no redundante de secuencias biológicas [36]. Tiene un proceso de curado y una anotación detallada en sus registros. Tiene un único registro por cada secuencia conocida y no es posible agregar secuencias sin un proceso de verificación. El objetivo de esta base de datos es almacenar secuencias conocidas con ubicaciones, funcionalidades y referencias claras y verificables. Esta base de datos

[2]CoCoPUTs: `https://hive.biochemistry.gwu.edu/review/codon2`

es significativamente más reducida que Genbank, pero es mucho más confiable.

El servicio web de CoCoPUTs [37] presenta tanto datos a nivel de CDS, como datos agregados por genoma de especie. En todos los casos el registro de CoCoPUTs detalla el origen primario del dato, es decir, si se trata de un registro de Genbank o de Refseq. CoCoPUTs recopila los registros de todos los CDS de Genbank y Refseq, y registra la cuenta de cada uno de los codones registrando las especies y los orígenes de cada CDS. CoCoPUTs trata de igual manera los registros de Genbank y de Refseq y no muestra campos diferenciados para cada base de datos. Si bien las bases de datos Genbank y Refseq son abiertas y su información esta disponible, las bases de datos secundarias como CoCoPUTs facilitan el acceso y presentan los datos de manera estructurada y estandarizada.

Para este trabajo se descargaron de CoCoPUTs los datos agregados por especie y se incluyeron todas la variables disponibles. Las especies se designan con su nombre y se acompañan por su 'TaxID', un número característico de cada taxón en las bases de datos de taxonomía de NCBI. Cada registro tiene anotado si fue obtenido a partir de la base de datos Refseq o Genbank. Se obtuvieron dos datasets de 1.118.703 registros para Genbank y de 152.902 registros para Refseq.

4.2.2 Descripción de los datos

Cada uno de los registros cuenta con las mismas variables. Estas variables constituyen los datos crudos sobre los cuales se construyen variables adicionales para enriquecer el estudio. Los registros anotados como provenientes de Refseq o de Genbank poseen las mismas variables, por lo que pueden compararse. Se detallan las variables originales presentes en la base de datos CoCoPUTs (Tabla 1). Además, se muestra una serie de registros de ejemplo (Fig. 9).

Variable	Descripción
Division	Refseq ó Genbank.
Assembly	Código de genoma asignado por GTDB (Genome Taxonomy DataBase).
TaxID	ID de especie asignado por NCBI.
Species	Nombre de la especie.
Organelle	Ubicación intracelular de las secuencias.
Translation Table	Código genético utilizado para la construcción del genoma. En NCBI se ordenan de 1 al 33.
nCDS	Cantidad de CDS involucrados.
nCodons	Cantidad de codones involucrados.
GC %	Contenido GC del organismo. Es el porcentaje de bases G o C sobre el total de bases codificantes. Se define para cada organismo a partir de los CDS considerados en la base de datos. Se toman en cuenta las abundancias de los 64 codones.
GC1 %	Contenido GC en la primera base de cada codón tenido en cuenta para el cálculo de la variable GC %.
GC2 %	Contenido GC en la segunda base de cada codón tenido en cuenta para el cálculo de la variable GC %.
GC3 %	Contenido GC en la tercera base de cada codón tenido en cuenta para el cálculo de la variable GC %.
Abundancias Codones (x64)	64 variables representando las abundancias relativas de los 64 Codones. Se toman las abundancias de los CDS de cada organismo.

Tabla 1: Variables originales. Se muestran una lista de las variables originales tomadas directamente de CoCoPUTs.

Division	Assembly	Taxid	Species	Organelle	Translation Table	n CDS	n Codons	GC%	GC1%	GC2%	GC3%	TTT	TTC	TTA
refseq	GCF_000001405.38	9606	Homo sapiens	genomic	11	119846	78581299	51.12	55.64	42.38	55.34	1341983	1377246	690119
refseq	GCF_002573245.1	562	Escherichia coli	genomic	11	4242	1310521	51.39	58.86	40.64	54.66	29635	21106	18517
refseq	GCF_000001635.24	10090	Mus musculus	genomic	11	76534	52306284	51.86	55.82	42.69	57.08	835058	981216	383207
refseq	GCF_000001735.3	3702	Arabidopsis thaliana	genomic	11	48147	20853569	43.9	50.45	40.17	41.07	471992	403279	279241

Figura 9: Muestra de datos de CoCOPUTs. Ejemplo de algunos registros de la base de datos utilizada. En la imagen se muestran solo 3 de los 64 codones que figuran efectivamente en la base de datos para cada registro.

4.3 Preparación de los Datos

4.3.1 Construcción de los datasets 'Primarios'

En esta etapa se toman los datos presentes en las bases de datos y se construyen los datasets utilizados para las etapas posteriores. A partir de las cuentas de cada codón se construyen 3 datasets de trabajo, que denominaremos 'Primarios'.

El primer dataset 'Codones' se compone de las abundancias relativas de los 61 codones codificantes para cada organismo (ver ecuación (1)). Se remueven los codones correspondientes a la señal de Stop debido a que no se espera que se comporten de manera semejante al resto de los codones ya que que no tienen un proceso de traducción comparable, sino que funcionan como señal de terminación. q_j representa la abundancia del codón j.

$$q_j = \frac{\#codones_j}{\sum_j^{61} \#codones} \tag{1}$$

El segundo dataset 'Aminoácidos' se compone de las abundancias relativas de los 20 aminoácidos (ver ecuación (2)), tomadas de los codones descritos en los datos crudos y de su relación con el código genético. p_i representa la abundancia relativa del aminoácido i y j_k los codones sinónimos para el aminoácido i.

$$p_i = \sum_{j_k} q_{j_k} \tag{2}$$

El tercer dataset 'Sinónimos' se compone de las proporciones de cada codón sinónimo dentro del aminoácido que codifica, siendo la sumatoria de cada conjunto de codones sinónimos igual a 1 (ver ecuación (3)). Se remueven los codones UGG (Trp) y AUG (Met), visto que sus valores son iguales a 1 en todos los casos, totalizando de esta manera 59 variables. r_j representa la proporción de cada codón j sobre el aminoácido i que codifica.

$$r_j = \frac{q_j}{\sum_{j_k} q_{j_k}} = \frac{q_j}{p_i} \tag{3}$$

Estos 3 datasets constituyen los datasets 'Primarios' sobre los cuales se aplicará la búsqueda de arquetipos (ver secciones posteriores).

4.3.1.1 Análisis de los 3 datasets 'Primarios'

Para este primer análisis se utilizan los datos de los datasets 'Codones', 'Aminoácidos' y 'Sinónimos', y se aplica un Análisis de Componentes Principales (PCA) [52, 53] para conocer la distribución general y los subconjuntos más relevantes de sus variables. Al aplicar el PCA sobre estos 3 datasets de manera individual se logran evaluar las distribuciones de sus variables de forma independiente y compararlas. Si bien las variables se relacionan entre sí y existe cierto grado de redundancia, se espera poder observar los grupos de variables de cada dataset de forma tal de poder mejorar la interpretación.

4.3.2 Construcción del dataset de 'Enriquecimiento'

Además de los 3 datasets primarios, se construye un cuarto dataset de 'Enriquecimiento' a partir de procesar variables adicionales. Para construir el dataset

'Enriquecimiento' se tomaron, en un principio, las variables presentes en la base de datos CoCoPUTs [37]. Luego, se realizó una búsqueda bibliográfica orientada, por un lado, a variables adicionales, y por otro, a parámetros relacionados con el funcionamiento celular que permitiesen construir nuevas variables. A partir de la información de las bases de datos de CoCoPUTs y de bibliografía, se generó el dataset 'Enriquecimiento' y sus respectivas variables continuas y discretas (Tabla 2) [41]. Estas variables no forman parte de la entrada del algoritmo de localización de arquetipos, sin embargo permiten asociar características adicionales a cada organismo.

Algunas variables toman valores directamente de los datasets 'Primarios' y otras utilizan los valores de abundancia combinándolos con información externa hallada en la bibliografía. En el caso de las variables de crecimiento bacteriano, estos valores fueron tomados directamente de la bibliografía sólo para un grupo reducido de organismos.

Grupo	Variable	Descripción
Abundancias	Codones (x64)	Se utilizan las abundancias de los 61 codones codificantes (q_j) y los 3 codones 'stop' de los organismos con código genético 'estándar'
	Aminoácidos (x20)	Abundancias de los 20 aminoácidos (p_i)
	Adenina	Abundancias del nucleótido Adeninda
	Guanina	Abundancias del nucleótido Guanina
	Citosina	Abundancias del nucleótido Citosina
	Timina	Abundancias del nucleótido Timina
	AG-Purines (Adenina + Guanina)	Abundancias de la suma de los nucleótidos Adenina + Guanina
	GT-Keto (Guanina + Timina)	Abundancias de la suma de los nucleótidos Guanina + Timina
Originales	nCDS	Cantidad de CDS
	nCodons	Cantidad de Codones
	GC	Proporción de G o C en los CDS
Distribución	h_C	Entropía de abundancias de codones
	h_A	Entropía de abundancias de aminoácidos
	h_R	Entropía de sinónimos
	w_j	Adaptabilidad relativa de cada codón
	w_mean	Adaptabilidad relativa promedio
	CAI	Índice de adaptabilidad de los codones
	fop	Frecuencia de codones óptimos
	Nc_sum	Suma de cantidad efectiva de codones
Fisiológicas	Nucleot_metab	Costos metabólicos de nucleótidos
	AA_cost_abs	Costos metabólicos de aminoácidos
	AA_cost	Costos metabólicos de aminoácidos ponderados por tiempo
	Robustness	Robustez de cada codón
Dinucleótidos	Rise	Parámetro de apareamiento entre bases Grado de separación entre bases (Angstrom)
	Roll	Parámetro de apareamiento entre bases Grado de inclinación (grados)
	Shift	Parámetro de apareamiento entre bases Grado de corrimiento (Angstrom)
	Slide	Parámetro de apareamiento entre bases Grado de corrimiento en eje alternativo (Angstrom)
	Tilt	Parámetro de apareamiento entre bases Grado de rotación (grados)
	Twist	Parámetro de apareamiento entre bases Grado de torsión (grados)
	Major Groove Depth	Profundidad del surco mayor del ADN (Angstrom)
	Major Groove Width	Ancho del surco mayor del ADN (Angstrom)
	Minor Groove Depth	Profundidad del surco menor del ADN (Angstrom)
	Minor Groove Width	Ancho del surco menor del ADN (Angstrom)
	Persistance length	Rigidez del ADN (nanómetros)
	Entropy	Entropía acoplada a la formación de un segmento de la doble hélice (cal/mol/K)
	Enthalpy	Entalpía acoplada a la formación de un segmento de la doble hélice (kcal/mol)
	Free Energy	Energía Libre acoplada a la formación de un segmento de la doble hélice (kcal/mol)
	Mobility	Constante de movilidad del ADN
	Melting Temperature	Temperatura a la cual se desnaturalizan 50 % de las bases acopladas (°C)
	Stacking Energy	Energía de Apilamiento acoplada a la formación de un segmento de la doble hélice (kcal/mol)
Crecimiento bacteriano	OGT	Temperatura óptima de crecimiento (°C)
	d_h	Tiempo de duplicación (horas)
Taxonomía	Dominio	Clasificación taxonómica con jerarquía de Dominio
	Phylum	Clasificación taxonómica con jerarquía de Phylum

Tabla 2: Variables del dataset 'Enriquecimiento'. Se agrupan las variables según su significado biológico.

4.3.2.1 Variables Continuas

4.3.2.1.1 Abundancias

Estas variables toman los valores de las abundancias de los elementos que constituyen los datasets 'Primarios'. Se agregan como variables de enriquecimiento a efectos de poder compararse con el resto. Como se mencionó anteriormente, estas variables son abundancias relativas por lo que toman valores entre 0 y 1. Se incluyen las abundancias de los 3 codones 'stop' de modo de observar su comportamiento.

- Abundancias de codones: Abundancias tomadas del dataset primario 'Codones'. Se incluyen las abundancias de los 64 codones por separado. Estas variables servirán como control y para poder orientarse entre los arquetipos obtenidos. Surgen de las cuentas de codones de cada organismo y toman valores entre 0 y 1.

- Abundancias de Aminoácidos: Al igual que con las abundancias de codones, se toman las abundancias del dataset primario 'Aminoácidos'. Cada una de estas variables surgen de la cuenta de los 20 aminoácidos partiendo de los codones correspondientes. Cada una de las 20 variables toma valores entre 0 y 1.

4.3.2.1.2 Originales

Estas variables se extraen directamente de la base de datos CoCoPUTs [37]. No sufren transformación alguna. Constituyen características generales del conjunto de secuencias utilizadas para las determinaciones de abundancias de codones.

- nCDS: cantidad de CDS involucrados en la especie. Esta variable resulta de contar la cantidad de CDS asociados a cada organismo que se utilizaron para el cálculo de codones. Al ser una cantidad, toma la forma de números enteros positivos.

- nCodons: cantidad de codones involucrados en la especie. De forma similar a 'nCDS', esta variable contiene la cantidad total de codones asociada a cada organismo. Nuevamente, da una idea de la cantidad de datos sobre los cuales se realizaron las determinaciones de abundancias. Toma valores enteros positivos.

- GC: Contenido GC del organismo, calculado a partir de los CDS. Esta variable registra la proporción de Guanina + Citosina con respecto al total de bases

consideradas en las secuencias. Es un parámetro tradicionalmente utilizado para caracterizar genomas o genes [2, 20, 21]. Toma valores entre 0 y 1.

4.3.2.1.3 Distribución

Estas variables dan cuenta, mediante diferentes métricas, de la distribución de codones o aminoácidos. Funcionan como diferentes medidas que evalúan el sesgo presente en los usos de codones o aminoácidos, según el caso.

- h_C: Entropía del uso de codones. A partir del concepto de entropía de Shannon [54, 55], se calcula esta métrica asociada a las abundancias de codones. Toma valores positivos, donde valores cercanos a cero implican una distribución de abundancias heterogénea y un valor más elevado representa una distribución más homogénea. Utilizando el logaritmo natural, se expresa en *nats* y se define como:

$$h_C = -\sum_{j=1}^{61} q_j.ln(q_j) \tag{4}$$

- h_A: Entropía de uso de aminoácidos. Al igual que en el caso de h_C, esta métrica se calcula en base a la Entropía de Shannon [54], solo que utilizando las abundancias de aminoácidos. Toma valores positivos, donde valores cercanos a cero implican una distribución de abundancias más heterogénea y un número más elevado da cuenta de una distribución más homogénea. Utilizando el logaritmo natural, se expresa en *nats* y se define como:

$$h_C = -\sum_{i=1}^{20} p_i.ln(p_i) \tag{5}$$

- h_R: Entropía asociada al uso de codones sinónimos. Al igual que en el caso de h_C y h_A, esta métrica se calcula en base a la Entropía de Shannon [54], solo que utilizando los valores calculados de r_j (fracción de codones sinónimos dentro de un mismo aminoácido). Toma valores positivos, donde valores cercanos a cero implican una distribución de abundancias más heterogénea y un número más elevado da cuenta de una distribución más homogénea. Utilizando el logaritmo natural, se expresa en *nats* y se define como:

$$h_R = -\sum_{i=1}^{20}\sum_{j=1}^{61} p_i.r_j.ln(r_j) \tag{6}$$

Teniendo en cuenta que $q_j = p_i.r_j$, a partir de la entropía de uso de codones se puede establecer una relación entre las entropías descriptas.

$$h_C = -\sum_{j=1}^{61}\sum_{i=1}^{20} p_i.r_j.ln(p_i.r_j) = -\sum_{j=1}^{61}\sum_{i=1}^{20} p_i.r_j.ln(p_i) - \sum_{j=1}^{61}\sum_{i=1}^{20} p_i.r_j.ln(r_j) \quad (7)$$

Dado que la suma de los r_j dentro de cada aminoácido i es igual a 1, el primer término puede reducirse dejando de lado r_j, quedando definido h_C como la suma de h_A y h_R:

$$h_C = -\sum_{i=1}^{20} p_i.ln(p_i) - \sum_{j=1}^{61}\sum_{i=1}^{20} p_i.r_j.ln(r_j) = h_A + h_R \quad (8)$$

Esta entropía de codones sinónimos h_R da cuenta de la homogeneidad en la distribución de codones dentro de cada aminoácido. Nuevamente toma valores positivos y un valor cercano a cero implica que en cada aminoácido solo se utiliza un solo codón. Valores elevados dan cuenta de que, dentro de cada aminoácido, el uso de codones es más homogéneo.

- w_j: Adaptabilidad relativa ('Relative Adaptiveness') [56]. Se calcula para cada uno de los 59 codones j donde haya más de un codón que corresponda a un mismo aminoácido. Es el cociente entre la abundancia relativa de cada codón con la abundancia relativa del codón sinónimo más abundante. Toma valores entre 0 y 1, donde los codones sinónimos más abundantes tienen valores de 1. Esta métrica se utiliza como variable intermedia para el cálculo de CAI y w_mean.

$$w_j = \frac{q_j}{max_i(q_j)} \quad (9)$$

- w_mean: Media aritmética de los valores de adaptabilidad relativa de cada codón. Se toman los valores de w_j de los 59 codones y se calcula su promedio. Esta variable toma valores positivos y da cuenta de la preferencia de los organismos por el uso del codón sinónimo más utilizado. Valores más bajos implican el uso de un único codón para cada aminoácido, mientras que valores

más elevados implican el uso de codones de forma más homogénea.

$$w_{mean} = \frac{\sum_{j=1}^{59} w_j}{59} \tag{10}$$

- CAI: 'Codon Adaptation Index'. Media geométrica de todos los índices de adaptabilidad relativa. Al igual que en el caso anterior, esta variable resume el conjunto de Adaptabilidad Relativa de todos los codones [56, 57, 58]. Esta métrica es tradicionalmente utilizada para comparar sesgos en el uso de codones. Toma valores positivos, donde valores cercanos a 0 implica que cada aminoácido utiliza un solo codón codificante y valores elevados implica que cada aminoácido utiliza sus respectivos codones codificantes de forma homogénea.

$$CAI = \sqrt[59]{\prod_{i=1}^{59} w_i} \tag{11}$$

- fop: Frecuencia de codón óptimo ('Frequency Optimal Codon') [59]. Cociente entre la suma de las frecuencias de los codones más usados dentro de cada aminoácido y la suma total de codones. Representa la proporción de uso de los codones mayoritarios de cada aminoácidos, considerados como codones 'óptimos'. Toma valores entre 0 y 1 donde $FOP = 1$ implica que se utiliza solo un codón por cada aminoácido y donde valores más cercanos a 0 implica que cada aminoácido utiliza sus respectivos codones codificantes en igual proporción, es decir, de forma más homogénea.

- Nc_sum: Número efectivo de codones (Ncf) [60]. Este parámetro da idea de la diversidad de los codones utilizados. Inicialmente, se calcula para cada aminoácido por separado. Un número efectivo de codones mayor implica que las abundancias relativas de los codones es más homogénea. Se incluye la suma de codones totales como variable a ser utilizada. Toma valores entre 20 y 61. Si todos los aminoácidos se codifican con un solo codón, Nc_sum toma un valor de 20, si los aminoácidos utilizan sus codones homogéneamente, el valor de Nc_sum se acercará a 61. Se calcula según:

$$N\_sum = \sum_{i}^{20} \frac{1}{\sum_{j_k} r_j^2} \tag{12}$$

Donde i representa los aminoácidos, j cada codón, j_k los codones sinónimos dentro de cada aminoácido, r_j, la proporción de abundancias de cada codón dentro de su codón sinónimo, como se describió anteriormente.

4.3.2.1.4 Fisiológicas

- Nucleot_metab: Costo metabólico de los nucleótidos tomados de manera individual [61] (Tabla 3). Se pondera por las abundancias relativas de cada nucleótido en los genomas. Toma valores positivos, siendo que valores elevados indican el uso de nucleótidos de mayor costo de síntesis.

Nucleótido	**Costo metabólico (ATP)**
Adenina	21.13
Timina	13.42
Guanina	20.37
Citosina	15.77

Tabla 3: Costos metabólicos de nucleótidos. Se muestran los costos metabólicos de los nucleótidos utilizados [61].

- AA_cost_abs: Costo metabólico absoluto de aminoácido. Se toman de bibliografía [62] los valores asociados al costo metabólico de cada aminoácido (Tabla 4). Estos valores estimados se ponderan por las abundancias relativas de cada aminoácido, obteniendo un valor promedio de costo metabólico en aminoácidos. Toma valores positivos. Valores elevados de esta variable implican que existe una preponderancia de aminoácidos de mayor costo de síntesis.

Aminoácido	Costo metabólico absoluto(kcal/mol)	Costo metabólico (kcal/(mol.t))
A	11.7	12
C	24.7	741
D	12.7	114
E	15.3	77
F	52	208
G	11.7	12
H	38.3	536
I	32.3	65
K	30.3	242
L	27.3	55
M	34.3	446
N	14.7	147
P	20.3	61
Q	16.3	130
R	27.3	109
S	11.7	70
T	18.7	112
V	23.3	47
W	74.3	892
Y	50	350

Tabla 4: Costos metabólicos de aminoácidos. Se muestran los costos metabólicos de aminoácidos, utilizados para el cálculo de las variables AA_cost_abs y AA_cost [20, 62]

- AA_cost: Costo metabólico de aminoácido. Este costo es comparable a 'AA_cost_abs' en cuanto que toma sus valores de costos metabólicos de síntesis de aminoácidos. Sin embargo, en este caso se incorpora un factor de degradación de cada aminoácido, tomado de bibliografía [20]. Este costo involucra parámetros de estabilidad de cada aminoácido, reflejando la durabilidad de cada uno de ellos (Tabla 4). Este factor de degradación multiplica al costo metabólico absoluto y es mayor cuanto menos estable sea cada aminoácido. Así, valores elevados de 'AA_cost' implican un alto costo por unidad de tiempo en la síntesis de aminoácidos. El uso de aminoácidos más sensibles y con mayor decaimiento implica que los sistemas biológicos involucrados en su síntesis deberán realizar un gasto mayor de recursos por unidad de tiempo. Esta variable toma valores positivos, donde valores elevados dan cuenta del uso de aminoácidos con un mayor costo relativo a su durabilidad.

- Robustness: Medida de robustez del patrón del uso de codones. Dentro del código genético, cada codón puede convertirse en cualquiera de 9 codones diferentes debido a mutaciones donde una base cambia por otra. Por cada codón existen 9 posibles codones a los cuales mutar con una única mutación. Estas mutaciones puntuales pueden involucrar o no un cambio en el aminoácido

correspondiente. Hay codones que, ante un cambio de una sola base, tienen una mayor probabilidad de cambiar el aminoácido codificado. Además, los cambios de un aminoácido pueden ser por uno fisicoquímicamente similar o por uno con una estructura muy diferente. La matriz de Grantham [63] relaciona todos los aminoácidos y les asigna distancias entre cada uno de ellos basado en su naturaleza fisicoquímica. Aminoácidos más semejantes fisicoquímicamente (Leu y Val, por ejemplo) tendrán una distancia menor a otros más diferentes entre si (Trp y Pro, por ejemplo). Teniendo en cuenta las distancias entre aminoácidos de Grantham, se calcula la distancia de cada codón a sus codones adyacentes (con una única base de diferencia). Para esto se tiene en cuenta el aminoácido codificado por cada codón. Luego se pondera por la probabilidad de ocurrencia de ese codón dado el contenido GC del organismo. La robustez por organismo se calcula teniendo en cuenta las abundancias relativas de cada codón:

$$R = -\sum_{j=1}^{61}\sum_{k=1}^{9} q_j.c_{jk} \tag{13}$$

Donde R representa el grado de robustez por especie y c_{jk} el costo de cada codón j con cada codón adyacente k. Un codón es adyacente a otro si sólo tienen una base de diferencia, por lo que cada codón tiene 9 codones adyacentes. El costo se pondera por la abundancia relativa q_j de cada uno de los 61 codones. Dentro del cálculo del costo c_{jk}, G_{aa} representa el costo del cambio de un aminoácido por otro, teniendo en cuenta la matriz de Grantham. Este costo se calcula teniendo en cuenta la probabilidad s_{jk} de transición al codón adyacente, basado en el contenido GC y en si se cambia hacia GC o hacia AU [64].

$$c_{jk} = s_{jk}.G_{aa} \tag{14}$$

Si el cambio en cuestión deriva en un codón con una ganancia de G o C, se utiliza la ecuación 15. Si por el contrario, el cambio analizado deriva en un codón con ganancia de A o U, se utiliza la ecuación 16. t representa la cantidad de codones adyacentes donde hay ganancia de G o C.

$$s_{jk} = \frac{GC}{t.GC + (9-t)(1-GC)} \tag{15}$$

$$s_{jk} = \frac{1 - GC}{t.GC + (9 - t)(1 - GC)} \tag{16}$$

Por ejemplo, en el caso del codón AGU (Ser), el valor de s_{jk} con respecto al codón AGG (Arg) dado un GC de 0.6 es 0.13. Este valor aumenta a 0.15 si el contenido GC es de 0.7. G_{aa} para este cambio de codones toma el valor de 110 dado que es la distancia de Grantham entre Ser y Arg. En el caso de que el cambio de una base derive en un codón de 'stop', se toma el valor 215 como G_{aa} ya que es el valor máximo de la matriz. Codones sinónimos tienen un $G_{aa} = 0$ por lo que no suman para el valor de R. Organismos con abundancia en codones cuyos codones adyacentes impliquen cambios mayores en los aminoácidos, tendrán un costo mayor y, consecuentemente, un valor más chico en esta variable. Organismos con codones más robustos implica que utilizan codones cuya sustitución no implica un cambio mayúsculo en el aminoácido a ser traducido, por lo que tendrán en esta variable un valor más elevado.

4.3.2.1.5 Dinucleótidos

- Propiedad_#(1 a 145): Se toman 145 propiedades de los 16 dinucleótidos posibles a partir de DiProDB[3] [65]. Estas propiedades se relacionan con la estructura que toman los nucleótidos en el espacio, sus ángulos, distancias, dimensiones, energía libre, energía de torsión, etc. Ponderando estos valores por cada codón, se pudo establecer un promedio general para cada organismo. Por ejemplo, el valor de entalpía de AC es de -9.4 kcal/mol y de -6.6 kcal/mol para el caso del dinucleótido CU, por lo que el codón ACU tomará el valor del promedio de los dos valores (-8 kcal/mol). Este valor se pondera según la abundancia relativa del codón ACU. Por consiguiente, cada una de las propiedades se calcula según:

$$Propiedad\_\# = \sum_{j}^{61} q_j \frac{dinucl_j^1 + dinucl_j^2}{2} \tag{17}$$

Donde qj representa la abundancia relativa de cada codón, $dinucl^1$ y $dinucl^2$ son las propiedades de lo dos dinucleótidos presentes en cada codón j. Se generan, así, 145 variables continuas para cada organismo.

[3]DiProDB: `http://diprodb.fli-leibniz.de/`

Se seleccionaron las propiedades de interés y se agruparon según su significado para poder facilitar el análisis:

- Propiedades geométricas: se incluyen todas las propiedades asociadas a determinaciones de Slide, Shift, Rise, Tilt, Roll y Twist de las hebras de ADN y ARN, ya sean unidas o no a proteínas (Fig. 10). Estas variables surgen de diferentes fuentes y describen las posiciones en el apareamiento de bases adyacentes en las hebras de ADN. Tilt, Roll y Twist toman valores de ángulos, mientras que Slide, Shift y Rise toman valores de distancia. En algunos casos se hallaron parámetros medidos con ADN unido a proteínas. En este grupo se agregan las variables referidas a la geometría de los surcos del ADN ('Major Groove Width','Minor Groove Width','Major Groove Depth' y 'Minor Groove Depth'). Al igual que con el apareamiento de las bases, es posible medir la geometría de los surcos del ADN.

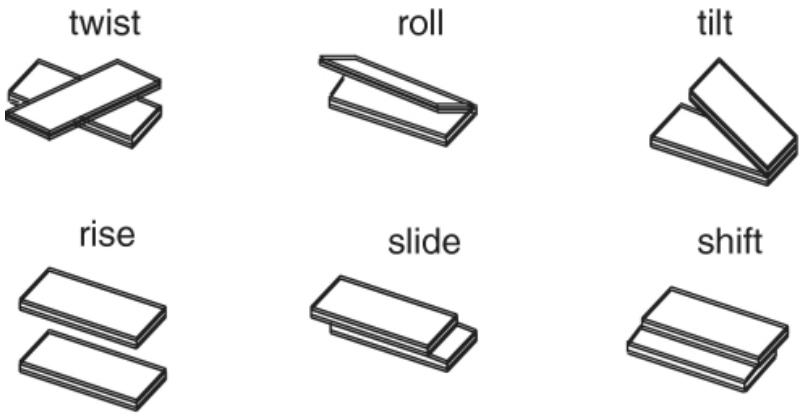

Figura 10: Parámetros de apareamiento de bases. Se muestran los 6 parámetros estructurales característicos de la geometría del los pares de bases. Dentro de la estructura del ADN, las bases adyacentes se aparean y es posible medir los parámetros resultantes de la geometría de los pares de bases contiguos. Los pares de bases contiguos toman una posición en el espacio influenciada por la naturaleza de las bases involucradas. Los seis parámetros toman valores distintos según los pares de bases involucrados y la interacción del ADN, con factores externos. La base de datos consultada contiene determinaciones de estos parámetros surgidas de diferentes fuentes [66].

- Propiedades asociadas a energías: se incluyen variables asociadas a determinaciones de energía libre, entalpía y entropía del proceso de acoplamiento de bases en las hebras de ADN. Estas variables son calculadas a partir de la ecuación de Gibbs:

$$\Delta G = \Delta H - T\Delta S \tag{18}$$

Donde G representa la energía libre, H la entalpía del proceso, T la temperatura y S la entropía. Las variables de energía libre toman valores negativos donde valores cercanos a cero implican una baja energía de interacción y valores más negativos implican una interacción más fuerte.

La entropía mencionada (ecuación 18) es diferente de la descripta anteriormente, ya que en este caso hace referencia al proceso de acoplamiento de las bases del ADN. También se incluye en este grupo a la propiedad 'melting temperature'. Esta temperatura es la necesaria para mantener separadas a la mitad de las moléculas ADN en solución. En este caso, los valores de esta variable tienen en cuenta las contribuciones que aporta cada dinucleótido a esta temperatura. Esta temperatura se relaciona íntimamente con la energía libre, según:

$$T_m = -\frac{\Delta G}{R.ln(\frac{C_T}{2})} \tag{19}$$

Donde T_m representa a la 'melting temperature', ΔG a la energía libre, R a la constante de los gases ideales y C_T a la concentración total de ADN. Esta temperatura toma valores positivos, expresados en grados y es inversamente proporcional a la energía libre. A efectos de poder comparar, esta variable de 'melting temperature' se construye cambiando su signo de forma de mantener la proporcionalidad de esta variable con respecto a las variables de energía libre y facilitar así su interpretación. También se incluyeron en este grupo las energías de 'stacking' entre pares de bases adyacentes y movilidad, dado que existe una fuerte asociación entre éstas energías y las de interacción.

4.3.2.1.6 Crecimiento bacteriano

- OGT: tras una búsqueda bibliográfica, se hallaron 139 valores de Temperatura Óptima de Crecimiento ('Optimal Growth Temperature', OGT) correspondientes a diferentes especies de bacterias presentes en los datasets primarios [67]. La tasa de crecimiento de los microorganismos está influenciada por la temperatura. Hay organismos que tienen una tasa de crecimiento mayor a elevadas temperatura y otros que crecen más rápido a bajas temperaturas. La OGT se calcula para cada microorganismo y se define como la temperatura

donde la tasa de crecimiento es máxima. Se utilizan estas temperaturas como variable dada su relación con el metabolismo celular general.

- d_h: Tiempo de duplicación. También llamado tiempo de generación, es el tiempo que demora una población de microorganismos en duplicar su cantidad. Toma valores expresados en tiempo donde mayor tiempo implica una velocidad de crecimiento más lenta. Tiempos más chicos implican que los organismos se duplican más rápidamente. Se toman de bibliografía valores para las misma especies que en el caso de las OGT [67].

4.3.2.2 Variables Discretas

4.3.2.2.1 Taxonomía

A partir de los 'TaxID' de los organismos tomados de CoCoPUTs, se realiza la búsqueda de todos los linajes en los servicios de NCBI. Los linajes dan cuenta de los taxones a los cuales pertenece cada organismo. Esto permite agrupar los diferentes organismos según el caso y la profundidad del análisis, prefiriendo, en este caso puntual, trabajar con los taxones de mayor nivel dado que poseen un número mayor de organismos y facilitan la observación. En el caso de Phylum, se eligen los taxones más representados dentro de cada dominio para facilitar la interpretación.

4.3.3 Procesamiento de las bases de datos

Con el objetivo de obtener un conjunto de especies con abundancias representativas, tanto a nivel codones como de aminoácidos, y que a su vez logren la mayor diversidad biológica posible, aplicamos una sucesión de filtros al dataset generado a partir de la base de datos CoCoPUTs [37] (Fig. 11). Cada decisión tomada en cada paso del proceso de filtrado se realizó en función del dataset obtenido en la etapa previa. A continuación se detallan cada uno de los filtros realizados.

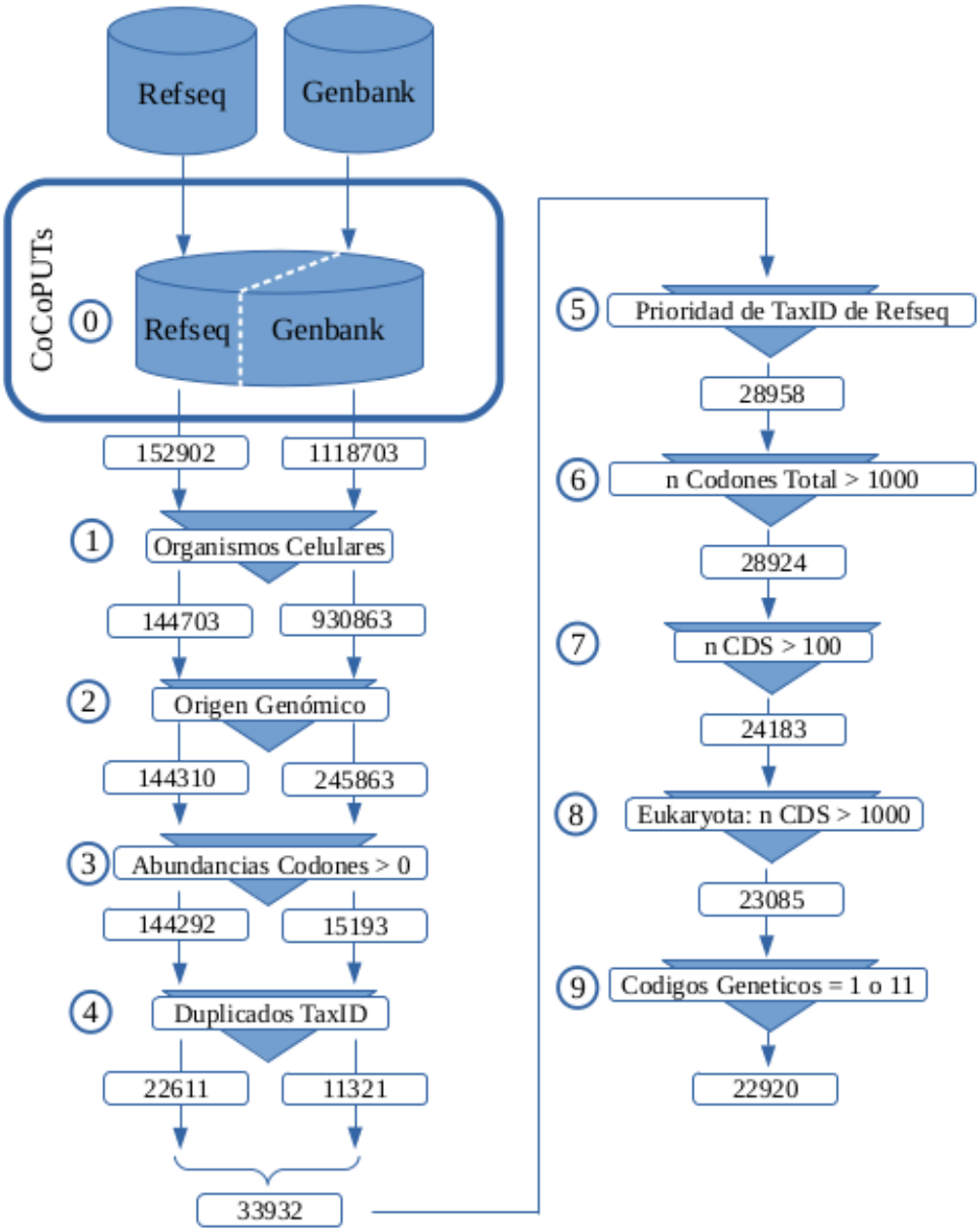

Figura 11: Flujo de selección de registros de trabajo. La base de datos secundaria CoCoPUTs toma los registros de las dos bases de datos primarias (Refseq y Genbank) para mantener datos estructurados y fácilmente disponibles. Desde allí se aplican filtros sucesivos sobre los registros de manera de poder construir los datasets de trabajo. Se numeran los pasos para facilitar su referencia. Tras cada criterio de filtrado, se muestra la cantidad de registros remanentes.

-0- Base de datos CoCoPUTs

La base de datos CoCoPUTs es nuestro punto de partida. Incluye todos los registros etiquetados como de Refseq y Genbank. Contiene 1.271.605 registros, 152.902 de Refseq y 1.118.703 de Genbank

-1- Filtro por Organismos Celulares

Como primer filtro, se retuvieron solamente aquellos registros que poseían CDS de organismos celulares, descartando secuencias de origen viral o plasmídico. Dada la naturaleza diferente de los virus o plásmidos que no son organismos independientes ni se replican de manera independiente, no se espera que los mecanismos subyacentes en los componentes de las secuencias sean necesariamente los mismos. Este primer filtro elimina el 15 % de las secuencias provenientes de las dos bases de datos, dejando 1.075.566 registros.

-2- Filtro por Origen Genómico

A continuación aplicamos un filtro, de forma de trabajar con secuencias de origen nuclear o cromosómico, descartando porciones mitocondriales, de plásmidos, de cloroplastos o de origen incierto. Nuevamente, estas porciones de secuencias descartadas pueden tener comportamientos diferenciados. En este paso se elimina el 63,7 % de los registros del paso anterior, la mayoría de ellos de Genbank, quedándonos finalmente con 390.173 registros.

-3- Filtro por Abundancias

A efectos de garantizar la representatividad, se separan todos los organismos que carezcan de algún codón en sus secuencias. Nuevamente, se intenta garantizar la representatividad del dataset de trabajo. En este caso se deja afuera el 59,1 % de los registros del paso anterior, la mayoría correspondientes a Genbank, quedándonos finalmente con 159.485 registros.

-4- Filtro de Duplicados

A partir de los nombres de cada organismo y de sus TaxID presentes en las bases de datos, se buscaron en NCBI sus TaxID correspondientes de forma de evitar errores de notación. Si bien la base de datos de NCBI mantiene un nombre primario para cada TaxID, existen varios casos en donde nombres parecidos pueden corresponder al mismo organismo y, por consiguiente, al mismo TaxID. Por ejemplo, los registros de 'Escherichia coli 08BKT77219' y 'Escherichia coli strain 08BKT77219' tienen el mismo TaxID 1081895. Se realiza un paso inicial de curado de los TaxID de manera

de identificar las especies de trabajo. Ante los duplicados de TaxID con igual base de datos de origen (Genbank o Refseq), se tomaron los de mayor número de nCDS, considerándolos más representativos de la especie. Así, se deja afuera el 78,7 % de los registros debido a TaxID duplicados, dejando 33.932 registros.

-5- Filtro de Prioridad de TaxID de Refseq

Existen varias especies duplicadas en ambas bases de datos (Refseq / Genbank) con diferentes número de secuencias en cada una. Dado que Genbank funciona como un repositorio de secuencias no necesariamente curadas, se priorizan los registros obtenidos a partir de Refseq ya que a los de este último repositorio se los considera de mejor calidad. Este criterio es compartido por la base de datos CoCoPUTs ante las especies que existen en ambas bases de datos primarias. En caso de que las especies figuren únicamente en Genbank, se mantiene el registro intacto. Así se descarta el 14,6 % de los registros del paso anterior, dejando 28.958 registros.

-6- Filtro de número total de codones

A pesar de tener al menos un codón de cada tipo en todos los organismos, existían registros con cantidades muy bajas de codones totales. Dada la extensión del dataset, se toma una cota inferior de 1000 codones para el número de codones totales de cada organismo. Así, en este paso, se descartan 34 organismos adicionales, dejando 28.924 registros.

-7- Filtro de cantidad total de CDS

Se detectaron registros con gran cantidad de codones que provenían de una cantidad muy reducida de nCDS. Esto implica que la muestra de codones obtenido representa, no el comportamiento general del genoma, sino de unos pocos genes (en algunos casos genes homólogos) aislados. A partir de las distribuciones de los tamaños de los genomas (ver sección resultados, Fig. 14), se plantea una cota inferior para la cantidad de secuencias codificantes (CDS) exigiendo al menos 100 CDS diferentes. De esta manera se dejan de lado organismos con muy pocos CDS a la vez que no se afecta mayormente el tamaño de la muestra. En este paso, se remueven 4.741 organismos (16,4 % de los presentes), dejando 24.183 registros.

-8- Filtro de cantidad total de CDS en Eukaryota

Con un criterio similar al anterior, se restringe aún más la representatividad para el caso de organismos Eukaryotas, dado que suelen tener un genoma más grande. Se amplía la cota inferior para este subgrupo a 1000 CDS dejando de lado organismos con cantidades bajas de CDS sin afectar el número total de organismos de Bacteria o Archaea. Se remueven 1098 organismos adicionales (4,5 % de los presentes en el paso anterior), dejando 23.085 registros.

-9- Filtro por código genético

Finalmente, y a efectos de remover posibles sesgos, se dejan de lado los organismos asociados a códigos genéticos distintos al generalizado tomado como estándar (Fig. 4B). Actualmente NCBI mantiene 33 códigos genéticos diferentes, y para este trabajo se mantienen únicamente los códigos 1 y 11. El código 1 es el código genético considerado estándar y el código 11 es una variación observada en algunas bacterias donde los procesos de traducción puede iniciarse también en los codones GUG y UUG, no habiendo diferencias en las correspondencias entre los codones y los aminoácidos que cada uno de ellos codifica. Se intenta despejar posibles comportamientos anómalos en las abundancias de codones debido a este factor. Se dejan afuera 165 organismos adicionales.

El dataset obtenido cuenta finalmente con 22.920 TaxID diferentes.

4.4 Modelado: Búsqueda de arquetipos y Elección de modelo

4.4.1 Búsqueda de Arquetipos mediante AAnet

Para este análisis se utilizó la herramienta denominada AAnet por los autores del método [49]. Este método toma un dataset y realiza una búsqueda de arquetipos en el mismo espacio de variables. Como resultado se obtiene la localización de los arquetipos hallados. La metodología se basa en la construcción de una red neuronal en forma de autoencoder con un número de capas que puede variar según el dataset involucrado. La bibliografía muestra un número de capas relativamente reducido para problemas con datasets de tamaños comparables a los que involucra nuestro estudio. La construcción del autoencoder involucra además, una modificación en la ecuación objetivo empleada para su entrenamiento (Fig. 12). Este autoencoder

incorpora dos restricciones adicionales que convierten un autoencoder tradicional en otro capaz de definir un espacio latente donde las diferentes instancias usadas para el entrenamiento se organicen alrededor de arquetipos. Esto se logra, por un lado, reduciendo el espacio latente del autoencoder a partir del número de arquetipos deseado, y por otro, aplicando un regularizador que limita a que los coeficientes de la capa latente tengan suma = 1, de modo que el espacio definido constituya una suma convexa de estos coeficientes [49]. La función de pérdida definida reduce el error de estimación a la vez que construye el espacio latente con un número definido de arquetipos y coeficientes de suma = 1. A modo exploratorio, se realiza una optimización de los hiperparámetros en la búsqueda de 4 arquetipos a través de un optimizador de tipo bayesiano utilizando el dataset 'Codones'. Se alcanza un set de hiperparámetros del autoencoder tras 250 iteraciones con 4 arquetipos (Tabla 5).

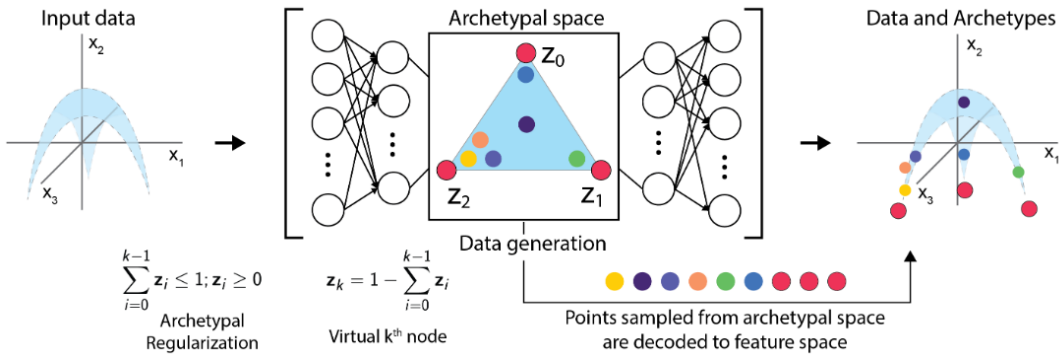

Figura 12: Método general de AAnet. El autoencoder permite llegar a un espacio de arquetipos con dos regularizadores que establecen funciones de activación de sumatoria menor o igual a 1 con sumandos positivos, y fija el número de nodos internos como $n-1$, donde n representa el número de arquetipos. Se puede reconstruir el espacio de variables originales junto con las posiciones de los arquetipos hallados [49].

Parámetro	Valor
noise_z_std	0.08
función de activación	Linear
gamma_mse	1,0
gamma_nn	1,0
gamma_convex	1,0
learning_rate	0.00088533
batch_size	3850
num_batches	2000
Dimensiones de autoencoder	7 capas internas
Número de neuronas	[105, 204, 99, 3, 99, 204, 105]

Tabla 5: Hiperparámetros de AAnet. Se detallan los hiperperámetros utilizados en la construcción del autoencoder para 4 arquetipos utilizando como entrada el dataset 'Codones' y luego de 250 iteraciones.

Esta estructura es comparable en dimensiones a los datasets utilizados por los

autores de la metodología [49]. Hallados los hiperparámetros para el caso particular mencionado (Tabla 5), se ensayan las búsquedas de 1 a 7 arquetipos y se mide la función de pérdida ('Loss'). Se grafica esta función en relación a la cantidad de arquetipos, esperando encontrar un codo o punto de quiebre, considerado como un indicador del número de arquetipos razonablemente asignado al dataset. Los arquetipos encontrados se muestran, posteriormente, en gráficas de componentes principales realizadas sobre el conjunto de datos originales.

4.4.2 Búsqueda de Arquetipos mediante ParTI

Este método, al igual que el método de AAnet, toma un dataset y realiza una búsqueda de arquetipos en el mismo espacio de variables. Como resultado se obtiene la localización de los arquetipos hallados. Esta metodología se basa en la búsqueda de un una politopo (generalización de un polígono en N dimensiones) en un espacio multidimensional [41]. Se inicia con un paso de PCHA (Principal Convex Hull Análisis), de manera de utilizar la superficie convexa ('convex hull') de los elementos en estudio, para definir un politopo que se ajuste lo mejor posible [48]. Este paso involucra la reducción de dimensionalidad mediante un Análisis de Componentes Principales y permite la estimación de un número óptimo de arquetipos (Fig. 7).

Para definir el número óptimo de arquetipos, se ensaya el ajuste del politopo con un número creciente de arquetipos y se define una curva de Varianza Explicada (EV) para diferentes cantidades de arquetipos. La EV se calcula teniendo en cuenta las diferencias entre cada punto observado (p_i) y su estimado (s_i).

$$EV(n) = (1/N) \sum_{i=1}^{N} (1 - \|p_i - s_i\| \,/\, \|p_i\|) \tag{20}$$

Se utilizan todos los puntos (N) y se calcula para cada arquetipo (n). s_i se estima a partir de las combinaciones lineales de los arquetipos hallados. Para puntos dentro del politopo s_i puede estimarse sin error ya que se calcula como una combinación convexa de los arquetipos, de forma que $p_i = s_i$, por lo que no reduce la $EV(n)$. Se construye, entonces, una curva de EV en función de n y se busca el punto de quiebre como punto de balance entre complejidad del modelo y la Varianza Explicada. Metodológicamente, se busca la distancia desde cada punto de esta curva a la recta que une sus extremos. Dada una recta $y = \alpha_1 x + \alpha_0$ que une el punto

mínimo y el máximo de la curva:

$$d_n = \frac{|\alpha_1 x_n - y_n + \alpha_0|}{\sqrt{\alpha_1^2 - 1}} \tag{21}$$

Donde d_n representa la distancia a la recta, α_1 y α_0 representan la pendiente y ordenada al origen de la recta y n la cantidad de arquetipos. El codo o quiebre se considera en el punto de la curva más alejado de esta recta donde d_n se hace máximo. La cantidad óptima de arquetipos será el n que logre el mayor valor de d_n.

Una vez definido el número óptimo de arquetipos, se realiza una reducción de dimensionalidad a partir de un Análisis de Componentes Principales equivalente al utilizado en el PCHA, para luego ajustar sobre el conjunto de datos de un politopo de tantos vértices como arquetipos. Para esto, los autores del método proponen 5 métodos alternativos capaces de realizar el ajuste y determinar las posiciones de los arquetipos. Cada uno de ellos utiliza un algoritmo distinto para ajustar el politopo al conjunto de datos de entrada. A continuación, se describen brevemente las 5 metodologías propuestas.

- **MVSA.** 'Minimum volume simplex analysis' [68] es un algoritmo que intenta ajustar un politopo, también llamado 'simplex', a un conjunto de elementos de un espacio multivariado. El politopo o simplex consiste en la envoltura convexa que surge de $(n + 1)$ puntos en un espacio de n dimensiones. Partiendo de suponer:

$$Y = MS \tag{22}$$

$$S \geq 1 \tag{23}$$

$$1_p^T S = 1_n^T \tag{24}$$

  Donde Y representa la matriz $p \times n$ de n observaciones en un espacio de p variables, M representa la matriz de p vértices del politopo en un espacio de p variables y S representa la matriz de las abundancias $p \times n$, es decir, da cuenta de las proporciones de cada M que permiten calcular cada Y. 1_p^T representa un vector fila de p valores igual a 1 y 1_n^T representa un vector fila de n valores igual a 1. De esta manera, se restringe para que cada elemento de S sea mayor a 0 y su sumatoria a lo largo de p sea igual a 1. El ajuste del politopo se realiza reduciendo el volumen del mismo sin violar las restricciones. Se plantea

un problema de optimización donde se intenta reducir el determinante de la matriz M, sujeto a las misma restricciones.

$$M^* = argmin(|det(M)|) \tag{25}$$

Considerando $Q = M^{-1}$, y por lo tanto, que $det(Q) = 1/det(M)$, se define la función de optimización como:

$$Q^* = argmax(-log\,|det(Q)|) \tag{26}$$

Queda definida S como:

$$S = Q^*Y \tag{27}$$

El algoritmo entonces intenta definir una matriz Q que cumpla con los criterios establecidos, por lo que se define M como la matriz que contiene los vértices del politopo. El algoritmo se inicia con componentes aleatorios de Q, calcula un gradiente según la función de optimización e itera hasta alcanzar un máximo.

- **Sisal.** 'Simplex identification via split augmented lagrangian' [69]. Este método es comparable con el anterior, solo que recibe una modificación en la función de optimización. En este caso, la función se define como:

$$Q^* = argmax(-log\,|det(Q)|) + \lambda\|QY\|_h \tag{28}$$

Donde $\|QY\|_h$ representa $\sum_{ij} h(QY)$ y $h(x) = max\{-x, 0\}$. λ es un factor de regularización del último término. Además, se relaja la primera restricción ($S \geq 1$), dejando únicamente:$1_p^T QY = 1_n^T$. Así, quedan permitidos los valores de S negativos aunque penalizados por el término regularizador. Se logra un ajuste del politopo con un algoritmo similar al caso de MVSA. El método Sisal es el definido por defecto por los autores de ParTI.

- **MVES.** 'Minimum volume enclosing simplex' utiliza una función de optimización similar a MVSA, pero el proceso iterativo de optimización se realiza de a una dimensión por vez [70]. Nuevamente, se intenta minimizar el volumen del simplex, es decir, su determinante. Operativamente, al igual que con los métodos anteriores, se maximiza el determinante de la matriz inversa a M.

- **SDVMM.** 'Successive decoupled volume max-min' utiliza un algoritmo basado en una función objetivo similar a las anteriores, pero incluyendo un concepto adicional [71]. Se plantea que ante la presencia de un error, el volumen del simplex esta sobreestimado, por lo que se optimiza una variable adicional que reduce el espacio dentro del conjunto de puntos observados dentro del cual se maximiza el volumen del politopo.

$$\max_{v_i} \left\{ \min_{u_i<r} |det(\Delta(v_i - u_i, ..., v_N - u_N))| \right\} \tag{29}$$

  Donde v_i constituye los vértices del politopo y deben incluirse dentro de la envolvente convexa de las observaciones ('convex hull'), u_i representa los vectores que reducen el volumen del politopo hasta un r máximo, y Δ construye la matriz M sobre la cual se aplica el determinante. Este método posee un factor que reduce el volumen y acerca los v_i a los puntos experimentales. Este método ubica los arquetipos cercanos a los puntos experimentales, a diferencia de los otros 4 métodos.

- **PCHA.** 'Principal convex hull analysis' plantea un método donde se busca estimar la matriz de datos observados a partir de las sumas ponderadas de los arquetipos, mientras que los arquetipos se estiman a partir de los datos observados [45, 47, 48]. Se plantea así un proceso de optimización de la función:

$$RSS = \sum_{i=1}^{n} \left\| x_i - \sum_{k=1}^{p} \alpha_i \sum_{j=1}^{n} \beta_{kj} x_j \right\| \tag{30}$$

  donde x_i representa la observacion i, α_i representa la proporción de cada arquetipo en cada observación i y β_{kj} representa la proporción de cada observación j en cada arquetipo k. El problema consiste en hallar los α_i y los β_{kj} a partir de iterar sobre cada uno de ellos alternativamente. Se restringen α_i y β_{kj} de forma que su sumatoria sea igual a 1, siendo z_k cada uno de los arquetipos:

$$z_k = \sum_{j=1}^{n} \beta_{kj} x_j \tag{31}$$

Los ajustes intentan reducir el área existente entre el 'convex hull' de los puntos y cada politopo candidato. Se ensayaron los 5 métodos de ajuste con 4 arquetipos sobre

el dataset 'Codones', a fin de comparar las distintas aproximaciones. Además, y a modo de establecer un grado de incertidumbre sobre la localización de los arquetipos, los autores proponen un proceso de 'bootstrapping', donde se toman muestras del dataset hasta conseguir el mismo número de elementos, y se realiza la búsqueda de los arquetipos iterativamente de forma de alcanzar un intervalo de posiciones de cada uno, centrado en la media y usando los desvíos estándar (SD) como error.

En una etapa posterior, se aplica el método ParTI sobre los 3 datasets de estudio. Se realiza el screening inicial con el método PCHA de acuerdo con lo recomendado por los autores del método [41] para definir el número de arquetipos. Luego, tras evidenciar la concordancia entre los diferentes métodos, se buscan los arquetipos en cada dataset utilizando la metodología 'Sisal', obteniendo 4 arquetipos en cada caso. Los arquetipos encontrados son ubicados luego en el espacio de dimensionalidad original a fin de compararlos y evaluar sus propiedades.

4.4.3 Métodos de reducción de dimensionalidad

4.4.3.1 Análisis de Componentes Principales (PCA)

Como se mencionó en la descripción del método ParTI, se realiza una reducción de dimensionalidad previamente a la búsqueda de arquetipos. Si bien es posible emplear otros métodos, los autores emplean PCA de manera de reducir el espacio de búsqueda y aumentar la eficiencia de los algoritmos. El análisis de componentes principales es un método de análisis multivariado ampliamente utilizados en diversas disciplinas científicas [52, 53, 72] . Este análisis parte de una matriz de datos conteniendo las observaciones con cada una de sus variables como características, y permite obtener una serie de autovectores y autovalores capaces de describir los datos en ejes ortogonales. Además, los autovectores logrados se ordenan de forma decreciente en cuanto a su capacidad de explicar los datos originales. El análisis de componentes principales parte de un SVD (Singular Value Decomposition) de la matriz de datos originales:

$$X = P\Delta Q^T \tag{32}$$

Donde X es la matriz de datos, P y Q, los vectores singulares izquierdo y derecho, y Δ representa la matriz de valores singulares ('Singular values'). Desde aquí se

definen las componentes del PCA como :

$$F = P\Delta \tag{33}$$

Donde F representa a las componentes ortogonales. Estas componentes se calculan a partir de la matriz de datos X y Q

$$F = XQ \tag{34}$$

Donde Q representa la matriz de proyección sobre las componentes principales. La matriz Q guarda las combinaciones lineales de las diferentes variables que logran construir las componentes ortogonales F. Q puede interpretarse como una matriz de proyección de los datos originales X en un espacio de componentes ortogonales. Las columnas de Q coinciden con los autovectores de las matriz X. Δ^2 representa las matriz de autovalores de la matriz X. Se obtienen entonces una matriz con los autovectores Q y una lista de autovalores que dan cuenta de la variabilidad total tomada de cada autovector. Cada autovector obtenido da cuenta de la proporción de la variabilidad de la matriz X equivalente a la proporción de su autovalor sobre la suma total de los autovalores. Es posible proyectar solo una porción de los autovectores captando la mayor variabilidad posible si se ordenan los autovectores en orden decreciente de autovalores y proyectando así la matriz X en un nuevo espacio con una dimensionalidad reducida [53].

4.5 Evaluación: Caracterización de arquetipos

4.5.1 Correlación de distancias

Dentro de cada dataset 'Primario' ('Codones', 'Aminoácidos' y 'Sinónimos'), se calcularon las distancias euclídeas entre cada organismo y cada arquetipo obtenido. A fin de caracterizar cada arquetipo, se buscaron las características con una elevada correlación con la distancia calculada, ya sea positiva o negativa. Arquetipos con valores muy altos o muy bajos en estas variables indican que la propiedad en cuestión podría estar ligada a alguna función en particular. A efectos de poder realizar comparaciones entre las variables y todos los arquetipos, se utiliza una métrica de distancia a partir del índice de correlación de Spearman [73, 74].

El coeficiente de correlación de Spearman ρ toma valores entre -1 y 1, y se

determina a partir del orden que toma cada uno de los elementos. Cada elemento toma un número de orden (rango) que se utiliza para realizar el cálculo del coeficiente:

$$\rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{35}$$

Donde d_i representa la diferencia entre los valores de igual rango y n el número de elementos a comparar. A fin de utilizarla como métrica de distancia, se utiliza el mismo principio comúnmente aplicado en la distancia de correlación de Pearson, que toma valores entre 0 y 2 donde variables perfectamente correlacionadas tienen una distancia de 0 [74].

$$dSpear = 1 - \rho \tag{36}$$

De esta manera se forma una métrica de distancia comúnmente utilizada para evaluar diferencias en las correlaciones de manera más robusta. Comparándolo con el coeficiente de correlación de Pearson, el coeficiente de correlación de Spearman es más robusto, es independiente de las distribuciones que pudiesen tener los datos y representa mejor la tendencia entre pares de variables ya que, por ejemplo, dos variables con tendencia creciente entre ellas tendrán un alto coeficiente de correlación de Spearman independientemente si su relación es lineal o no.

Se construye una matriz de distancias a partir del coeficiente de correlación de Spearman entre cada una de las variables del dataset 'Enriquecimiento' a las que se agregan las distancias a cada arquetipo. Así se logran distancias de las variables entre sí, y entre las variables y cada uno de los arquetipos en estudio. Arquetipos altamente correlacionados se ubican a distancias menores entre si con respecto al grupo de organismos.

Para poder asociar las variables con los diferentes arquetipos, se busca que las correlaciones entre las distancias organismo-arquetipo y las propiedades de los organismos sean cercanas a 1 o a -1. Si una variable tiene una correlación cercana a 1 con la distancia a un arquetipo, el arquetipo en cuestión podría asociarse con valores bajos de esa variable, y su distancia *dSpear* será cercana a cero [62, 67]. Por otro lado, si la correlación entre una variable y un arquetipo tienen una correlación cercana a -1, podría asociarse ese arquetipo con valores elevados de esa variable, y la distancia *dSpear* entre el arquetipo y la variable será elevada. Dado que es de interés poder analizar ambos aspectos, se construyen dos medidas para cada par organismo-arquetipo. La distancia euclídea entre cada organismo (O) y cada

arquetipo (A).

$$d_E(O, A) = \sqrt{\sum_{i=1}^{n}(o_i - a_i)^2} \tag{37}$$

donde i representa cada una de las dimenciones en cada dataset y, a_i y o_i, las i-ésimas componentes (i=1...n) de los A y O, respectivamente. Además, se plantea una métrica complementaria:

$$p_E(O, A) = -d_E(O, A) \tag{38}$$

Esta segunda métrica intenta reducir *dSpear* a cero en casos en donde la correlación entre una variable y la distancia sea cercana a -1. Variables con correlaciones de Spearman cercanas a 1 con respecto a p_E alcanzarán *dSpear* cercanas a cero, a la vez que se podrán asociar valores elevados de la variable en cuestión con el arquetipo analizado.

4.5.2 Multidimensional Scaling (MDS)

El método MDS toma como entrada una matriz de disimilaridades o distancias, y permite obtener coordenadas de los elementos involucrados tal que se reproduzcan las distancias originales tanto como sea posible. El método es similar al de PCA, ya que utiliza la factorización de matrices mediante SVD y alcanza un conjunto de autovalores y autovectores, solo que en lugar de obtener una proyección a partir de las variables de un conjunto de observaciones, lo hace a partir de los valores de las distancia entre cada una de las observaciones. El método intenta realizar una proyección en un número reducido de dimensiones del conjunto de datos, de manera de reproducir las distancia de la matriz de entrada lo más fielmente posible. Para esto se utilizan las dos dimensiones principales únicamente, de manera de representarlo gráficamente. Este método permite la realización de los mapas de variables descriptos en la sección de caracterización de los arquetipos [73, 75].

4.5.3 Clustering Jerárquico

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) es un método de agrupamiento ('clustering'), que toma como datos de entrada a una serie de observaciones con sus respectivas variables que los caracterizan en forma de matriz [76]. Es un método de agrupamiento aglomerativo que parte de calcular distancias

entre todos las observaciones, de forma de construir una matriz de distancias inicial [77]. Al inicio, cada punto representa un 'cluster' propio y, de forma iterativa, busca la menor distancia entre los 'clusters' y los combina en un 'cluster' de mayor nivel. En cada paso, las nuevas distancias calculadas se computan promediando las distancias entre los componentes de cada par de clusters. El algoritmo termina al llegar a un solo 'cluster' en el nivel más alto [78]. Así se llega a construir un dendrograma representando la relación entre todos los componentes. En el caso de este trabajo, se emplea para definir los agrupamientos de los arquetipos, y se utiliza la distancia euclídea entre ellos.

5 Resultados

5.1 Procesamiento de las bases de datos

5.1.1 Descripción de base de datos original

5.1.1.1 Asimetría de los dominios de las bases de datos

En una primera etapa, se realizó una exploración de las distribuciones de las principales variables, presentes en la base de datos CoCoPUTs, tal cual fueron descargadas. Es importante hacer notar que los registros etiquetados como provenientes de Genbank y Refseq muestran distribuciones diferentes en varios aspectos. Los registros marcados como Refseq contienen información proveniente mayoritariamente de ensambles ('Assemblies') de secuencias bacterianas, dejando una minoría de registros asignados a Eukaryota o Archaea. Genbank, por el contrario, se nutre de un número más elevado de fuentes por lo que incluye un número mayor de especies, sobre todo asociadas a los dominio Eukaryota y Archaea (Fig. 13). Visto que nuestra hipótesis de trabajo propone una caracterización a nivel especie o conjuntos de especies, la diversidad de Genbank se muestra como un aporte fundamental para llevar a cabo el estudio.

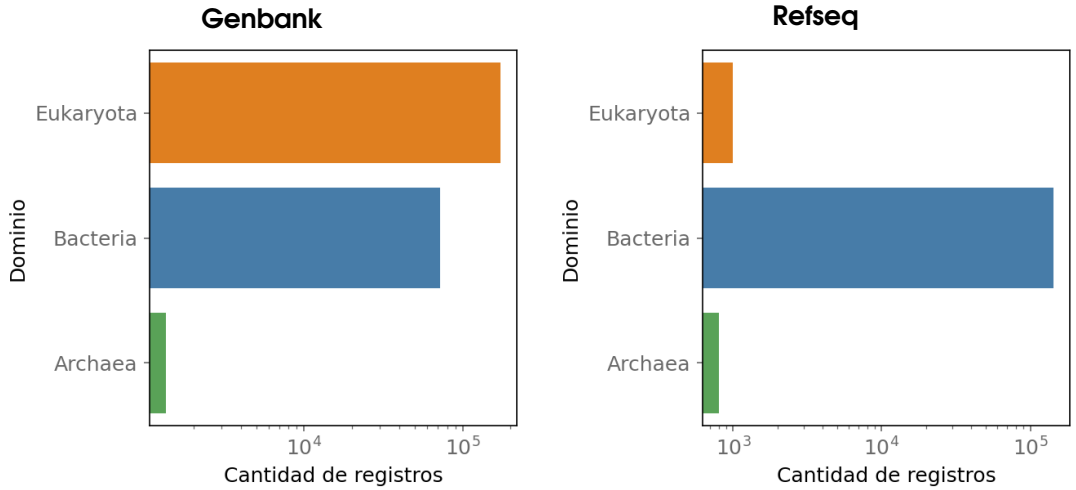

Figura 13: Proporciones de dominios. Se muestran las proporciones en las bases de datos de los dominios. Los valores fueron obtenidos tras aplicar los filtros 1 y 2 (Fig. 11) donde se remueven las secuencias de origen mitocondrial o plasmídico (filtro 1) y donde se remueven secuencias virales o de origen incierto (filtro 2). Genbank posee un mayor número de organismos Eukaryota con componentes de CDS bacterianos. Refseq contiene fundamentalmente secuencias de origen bacteriano.

5.1.1.2 Tamaños de Genomas

En el presente trabajo realizamos un estudio de la distribución de las cantidades de secuencias existentes para cada organismo, en función de la fuente de datos de origen. En el caso de Genbank, la gran mayoría de los organismos de Bacteria y Archaea presentan una cantidad de entre 1 y 10^6, y entre 1 y 10^4 CDS, respectivamente (Fig. 14). En el caso de Refseq, la distribución en estos dos dominios muestra una alta frecuencia, de entre 10^3 y 10^4 CDS (Fig. 14). En el caso de Genbank, aparecen una gran cantidad de organismos con muy bajo número de CDS. Esto podría evidenciar un posible sesgo en los datos que puede comprometer la representatividad de los organismos a utilizar. En el caso de Eukaryota, los perfiles de ambas bases de datos son marcadamente distintos: mientras que Genbank muestra un gran número de organismos totales y una gran proporción de ellos con una cantidad de CDS muy bajos, Refseq mantiene menor cantidad de organismos pero mejor representados (Fig. 14). Una búsqueda más detallada sobre Genbank reveló algunas especies con menos de 10 CDS, que bien podrían contener sesgos en las cuentas de codones para estos organismos. Dada la disponibilidad de diferentes organismos, definimos los filtros 7 y 8 en función de la cantidad de CDS (Fig. 11). En dichos filtros definimos una cota inferior -para las Bacterias y Archaea de 10^2 CDS y de 10^3 CDS para Eukaryota- de manera de remover los organismos que no cuentan con un número de CDS representativo. La cota más alta de Eukaryota responde, también, al hecho de que estos organismos suelen poseer mayor cantidad de genes, por lo que organismos con aproximadamente 100 CDS en este dominio pueden tener un sesgo debido a una baja cantidad de registros de CDS.

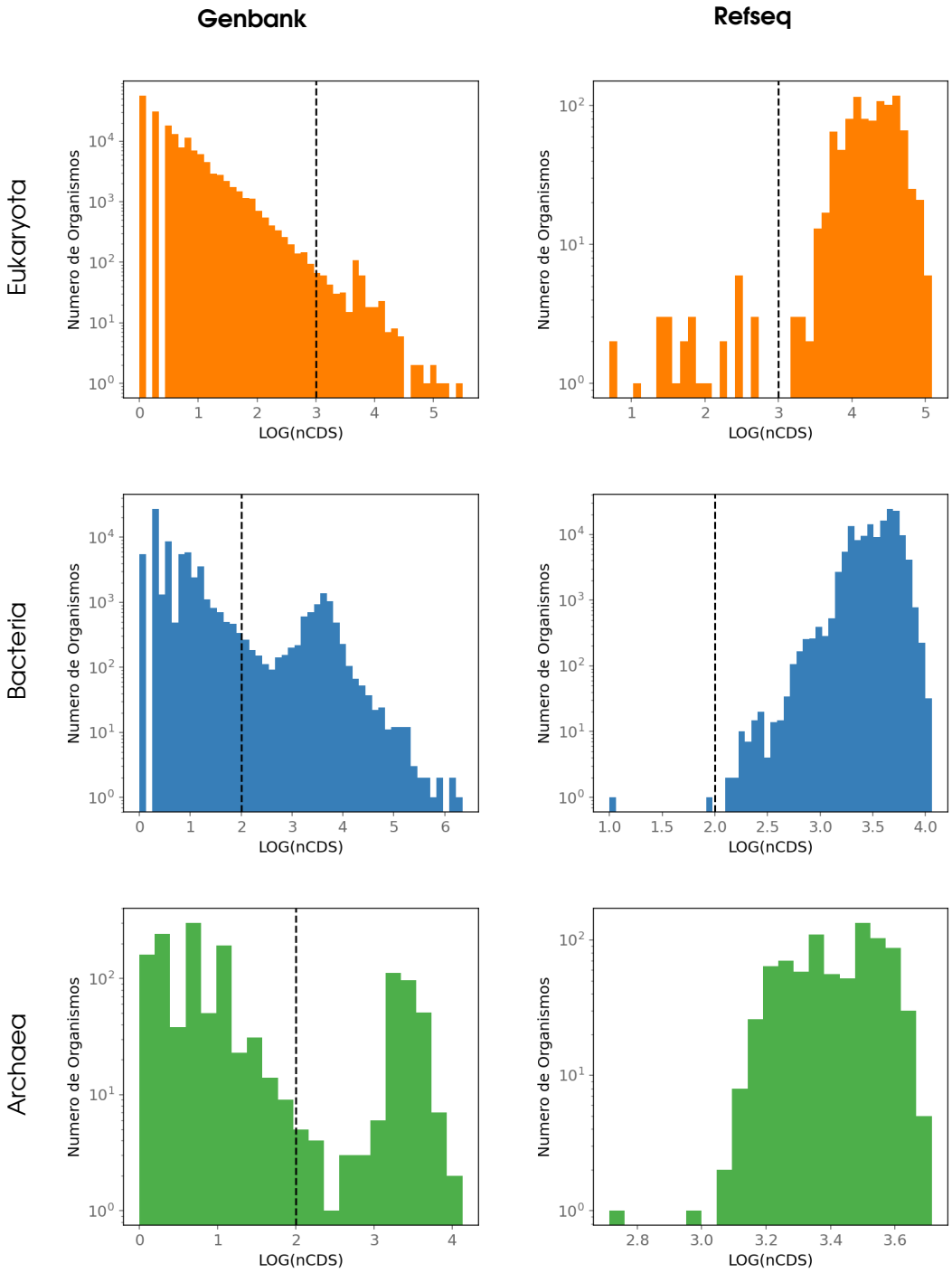

Figura 14: Distribución de tamaños de genomas. Las distribuciones de cantidades de CDS segmentado por dominio para las dos bases de datos utilizadas: Genbank (izq) y Refseq (der). Los valores fueron obtenidos tras aplicar los filtros 1 y 2 (Fig. 11). Refseq muestra genomas Bacterianos y de Archaea entre 10^3 y 10^4 CDS. Posee poca cantidad de organismos Eukayota. Por otro lado, Genbank posee organismos con cantidades de CDS muchos menores. Esto se ve marcadamente en Eukaryota, donde aparece un pico de organismos en cantidades de CDS cercanas a cero. Se muestra como líneas punteadas a los puntos de corte propuestos para cada taxón.

Dada la relevancia para el tópico a estudiar, se realiza un análisis preliminar de la variable "Contenido GC" de todos los organismos involucrados. Se observa una distribución no homogénea en cada una de las bases de datos (Fig. 15). Las

distribuciones de organismos se ven ligeramente asimétricas en ambos casos, en donde organismos de mayor GC parecen tener más cantidad de CDS en los casos donde el número de CDS supera las 10^3 secuencias. En el caso de los organismos provenientes de Genbank, se hallan organismos celulares en todo el rango de cuentas de CDS. Refseq encuentra sus organismos celulares organizados dentro de un rango acotado, con una leve tendencia hacia mayores cantidades de CDS en organismos con mayor GC. Por otro lado, en el caso de Genbank, se observa un grupo de menor cantidad de CDS, en donde hay menos abundancia de organismos con alto GC y un segundo grupo de mayor número de CDS, donde, al igual que en la base de datos anterior, hay un mayor número de secuencias en organismos de alto GC. Entre los registros de ambas fuentes (Genbank o Refseq) se encuentran proporciones minoritarias de secuencias anotadas como 'NN' u 'other sequences' que constituyen grupos de CDS de orígenes diversos o no anotados que serán removidos del análisis posterior.

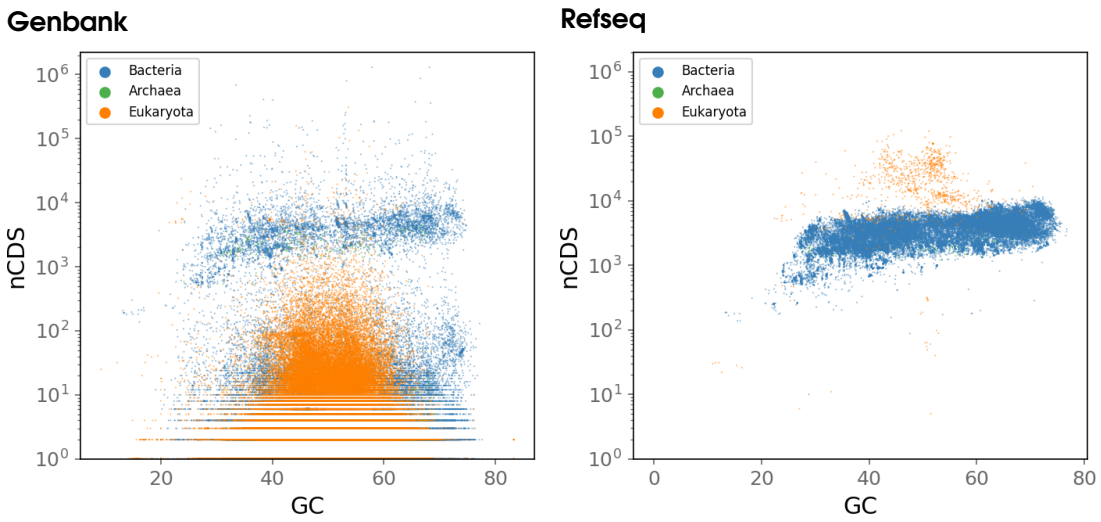

Figura 15: Representatividad de las secuencias a lo largo del espectro de contenido GC. Se muestra el número de CDS para cada Dominio discriminado por valor de contenido GC en cada una de las fuentes (Refseq o Genbank) utilizadas. Se colorean según el Dominio correspondiente.

5.1.2 Descripción de base de datos filtrada

Se conforma un dataset final tras la aplicación de sucesivos filtros (Fig. 11), teniendo en cuenta ambas fuentes (Refseq o Genbank) en donde el dato de Refseq toma precedencia sobre la de Genbank, siguiendo el mismo criterio empleado por CoCoPUTs (ver sección 4.3.3). En la Figura 16 se puede observar una descripción de dicho dataset. Se constituye el dataset definitivo de trabajo compuesto por 22.920

registros, cada uno con un 'TaxID' único.

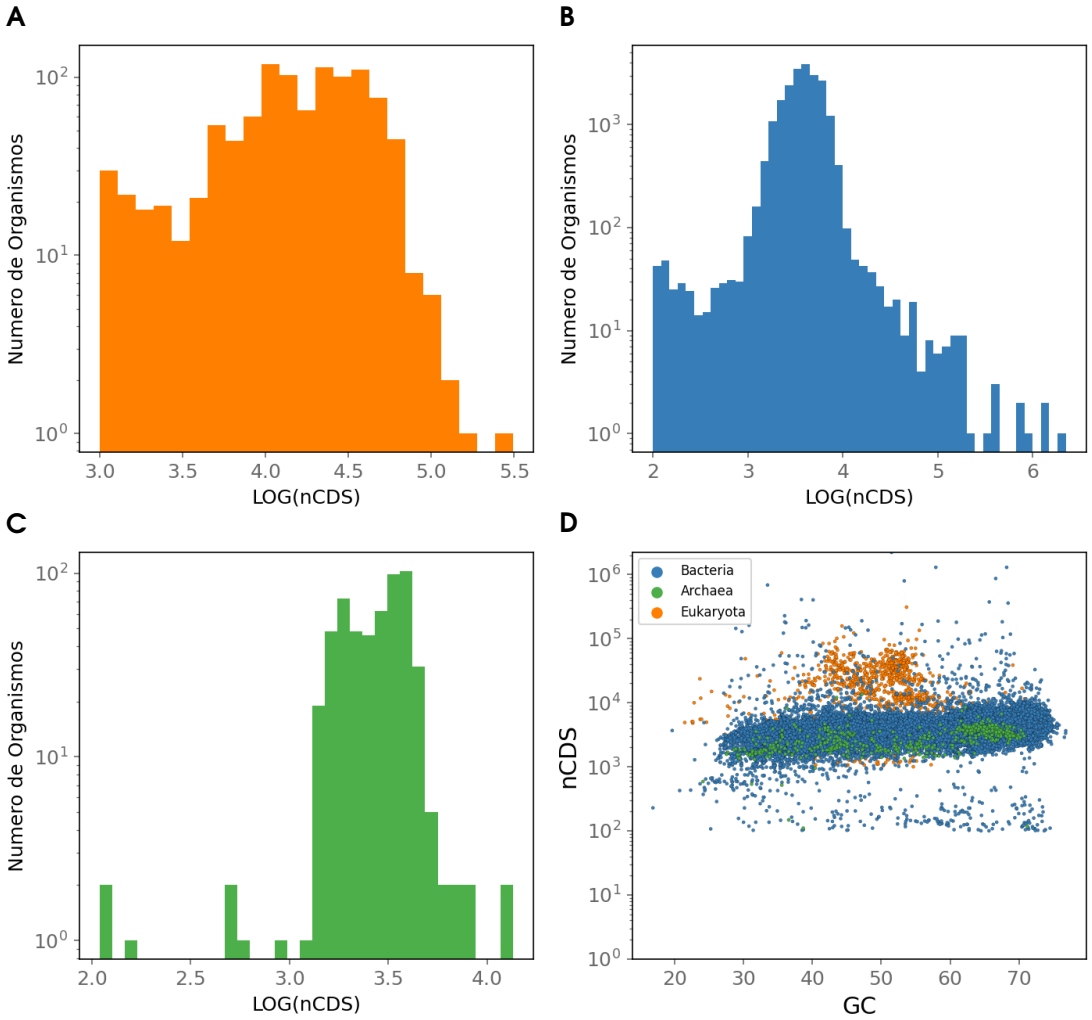

Figura 16: Distribución de nCDS en el dataset definitivo. Tras el filtrado inicial de la base de datos, se constituye un set de organismos de trabajo. A, B y C: Se muestran los perfiles obtenidos para el volumen de genes utilizados (nCDS) para Eukaryota, Bacteria y Archaea, respectivamente. D: se muestra el contenido 'GC' y las abundancias de Dominios en el conjunto definitivo de organismos.

5.2 Construcción de datasets de trabajo

5.2.1 Características de los datasets de trabajo

La construcción de los tres datasets primarios ('Codones', 'Aminoácidos' y 'Sinónimos') se realizó en base a las frecuencias de aparición de los codones / aminoácidos en los organismos. Los resultados de un análisis preliminar de PCA sobre los tres datasets muestran que en el dataset 'Codones', y particularmente en el 'Sinónimos', aparece como primera componente principal aquella que ordena los codones en función del contenido GC (Fig. 17A y 17B). Esto implica que la primera componente da cuenta de las diferencias de comportamiento de los codones

en ambos datasets. En el caso de 'Sinónimos', se observa que las cargas de los codones se alinean con la primera componente, pero mantienen una dispersión más homogénea. El dataset 'Codones' mantiene sus codones más alineados, con solo algunas excepciones que aportan a la segunda componente (PC2).

En el caso del dataset 'Aminoácidos', el PCA marca una primera componente donde se incluye más del 80 % de la variabilidad. Los aminoácidos se alinean con esta componente dejando un número reducido que se dispersan sobre la componente 2 (Fig. 17C). Si se analiza la correlación entre todos los aminoácidos del dataset, se observa más claramente que se forman dos grandes grupos discretos anticorrelacionados entre sí (Fig. 18). Aparecen dos pares de aminoácidos que se separan ligeramente de estos grandes grupos: Asp-Thr y His-Leu que mantienen perfiles particulares. Los dos grandes grupos poseen distintos tipos de aminoácidos dando cuenta de que cualquiera sea el grupo utilizado, existe diversidad en la composición de las proteínas. Estos dos grupos muestran una diferencia en cuanto al contenido GC de sus codones codificantes. El grupo de aminoácidos Val, Trp, Pro, Gly, Ala y Arg tienen codones con valores de mayor contenido GC mientras que el grupo de Phe, Lys, Asn, Ile y Tyr utilizan codones de menor contenido GC.

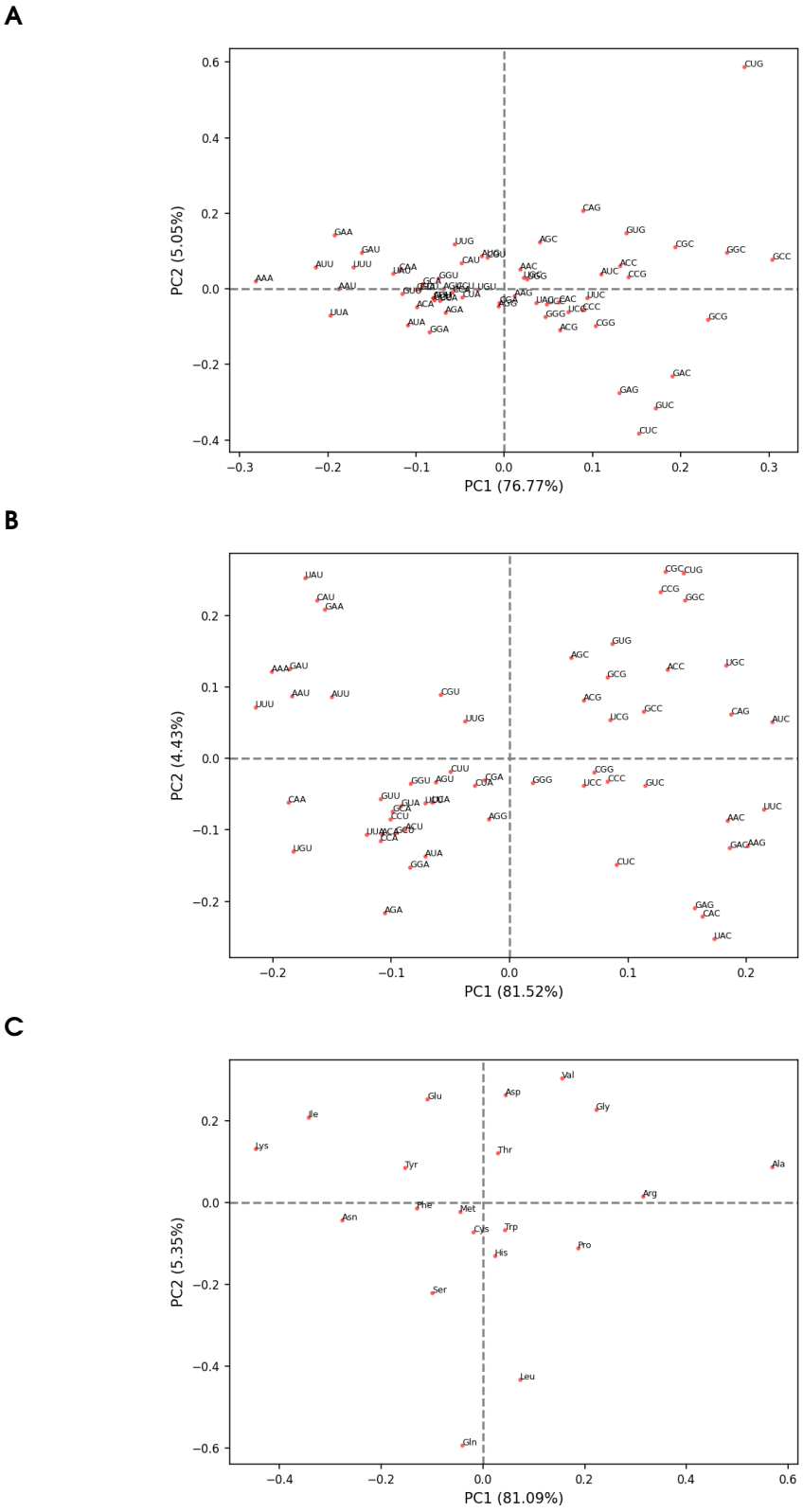

Figura 17: Cargas en las Componentes Principales. A. Primeras dos componentes del dataset 'Codones'. B. Primeras dos componentes del dataset 'Sinónimos'. C. Primeras dos componentes del dataset 'Aminoácidos'. En cada caso, la primera componente posee una gran proporción de la varianza total, dando cuenta de las distribución general de las variables en estos datasets. Puntos más cercanos entre sí dan cuenta de una correlación entre las variables de cada dataset. En todos los casos, se observa una tendencia a alinearse con la primera componente, mostrando que la PC1 es capaz de discriminar el comportamiento de las variables involucradas.

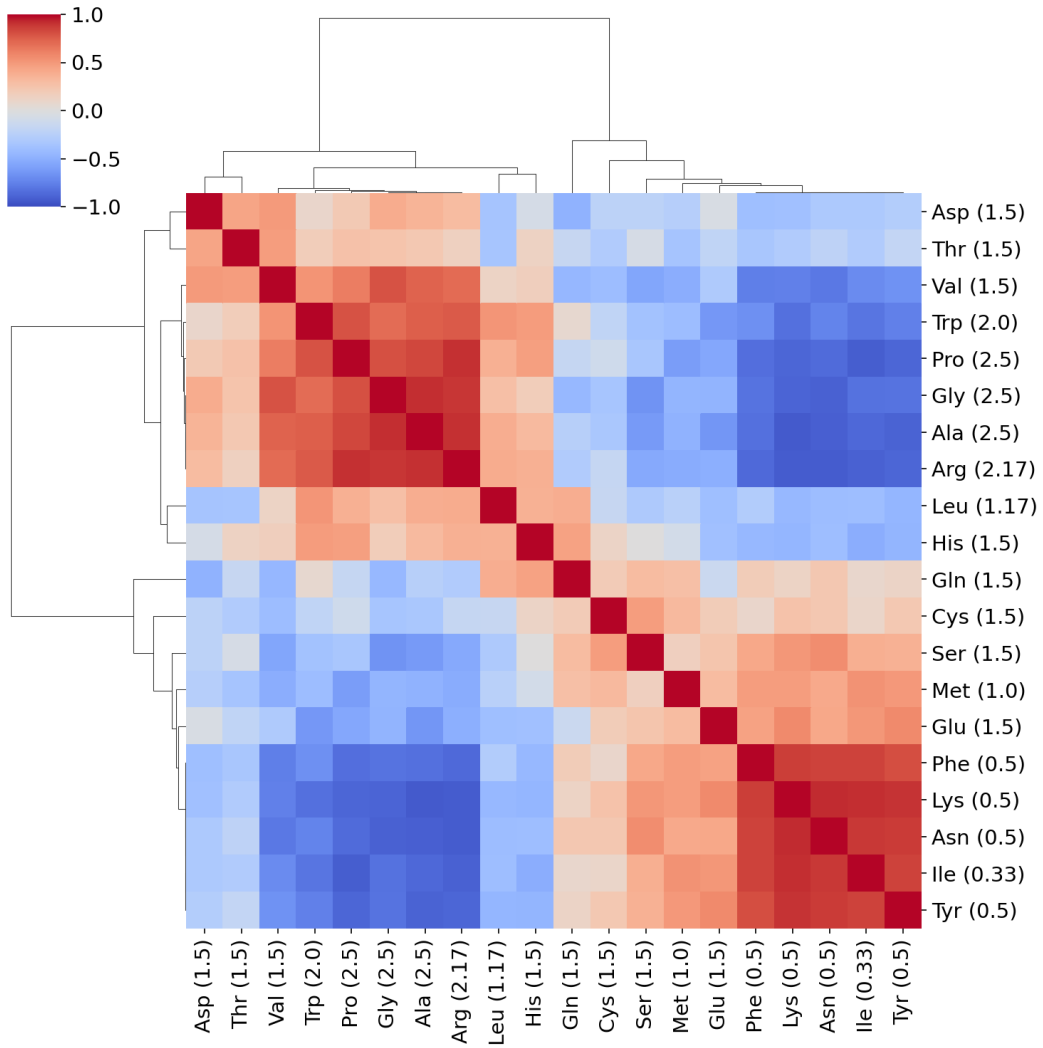

Figura 18: Correlograma de abundancias de aminoácidos. Se construye un correlograma tomando las abundancias de aminoácidos de todo el dataset 'Aminoácidos'. Se muestran los aminoácidos agrupados según UPGMA de manera de identificar agrupamientos generales. Se visualizan dos grandes grupos principales y algunos aminoácidos que se diferencian ligeramente, sobre todo el caso de His y Leu. Junto a cada aminoácido figura entre paréntesis el contenido GC promedio de todos sus codones que lo codifican.

5.3 Elección del modelo de búsqueda

Durante la búsqueda bibliográfica realizada al comienzo de este trabajo, se hallaron unos pocos métodos comúnmente utilizados para las búsquedas de arquetipos, por lo que se realizaron ensayos preliminares sobre dos de ellos, para seleccionar el más adecuado para el problema planteado. El método ParTI fue hallado en varias publicaciones relacionadas con disciplinas biológicas, mientras que AAnet figura como el método más reciente aplicado tanto a problemáticas biológicas como a tratamiento de imágenes.

5.3.1 AAnet

A efectos de evaluar el método AAnet, se realizaron búsquedas de una cantidad variable de arquetipos sobre el dataset 'Codones' (ver sección 4.4.1). La metodología utilizada por AAnet mostró un éxito parcial en la búsqueda de arquetipos sobre los datasets utilizados. Los ensayos preliminares se realizaron sobre el dataset 'Codones' con 2, 3 y 4 arquetipos, permitiendo ubicar arquetipos dada la distribución de los organismos. Se elige este dataset ya que es el que contiene mayor número de variables, y el que reviste mayor interés ya que describe las abundancias de codones de manera directa.

AAnet requiere de un ajuste de varios hiperparámetros para estructurar el autoencoder adecuadamente. Estos hiperparámetros incluyen: los coeficientes que definen las capas y cantidad de neuronas del autoencoder, los parámetros del proceso de aprendizaje, los coeficientes que definen la función de pérdida, y el tipo de función de activación de las neuronas. Los cambios en los parámetros afectan la ubicación de los arquetipos hallados. Se hizo necesario, entonces, hallar un set de hiperparámetros capaces de localizar un número variable de arquetipos de manera consistente.

Durante el proceso de optimización, se logró reducir la función de pérdida del autoencoder. Sin embargo, resultados bajos de la función de pérdida no necesariamente arrojaron una distribución de los arquetipos consistentes con las instancias entregadas como entrenamiento (Fig. 20C). Hiperparámetros capaces de reducir la función de pérdida del autoencoder no siempre generan un espacio latente donde las distribuciones de los organismos concuerden con los arquetipos hallados. No solo eso, sino que la optimización para diferente número de arquetipos arroja distintos hiperparámetros, lo que dificulta la comparación y la elección del número óptimo de arquetipos (Fig. 19). La función de pérdida por sí sola no alcanza para definir el número óptimo de arquetipos, y la regularización impuesta no parece suficiente para localizar arquetipos en el dataset 'Codones'.

Para alcanzar un resultado interpretable y compatible con los conceptos de arquetipos definidos originalmente, fue necesario optimizar el autoencoder hasta lograr una baja función de perdida, y posteriormente redefinir los hiperparámetros de acuerdo a los utilizados por lo autores para datasets de dimensiones similares. De esta manera, se logra encontrar una estructura robusta frente a un número variable de arquetipos que permite obtener resultados reproducibles e interpretables en cada

ejecución.

A fin de evaluar el número óptimo de arquetipos, se utilizó el criterio de la ubicación del codo en el gráfico de error de aproximación en función del número de arquetipos, en donde se intenta buscar un equilibrio entre la complejidad del modelo y el error de estimación alcanzado (Fig. 19) [49]. Las sucesivas búsquedas de arquetipos muestran una gráfica donde el número óptimo de arquetipos podría considerarse entre 2 y 4, lugar donde la curva muestra un codo.

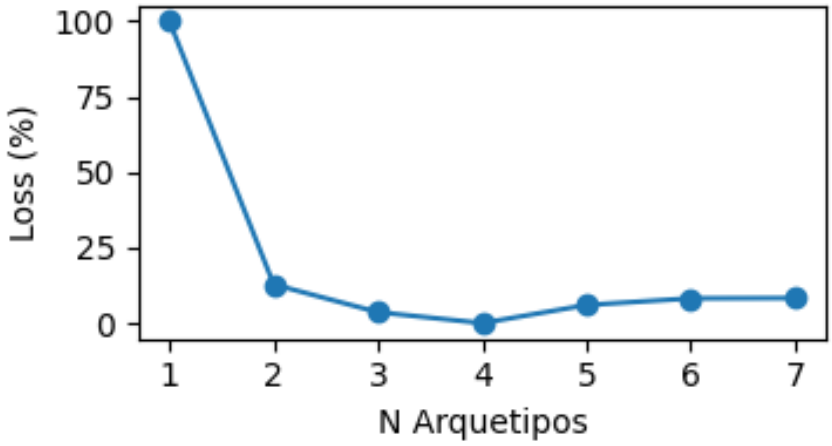

Figura 19: Identificación de número de arquetipos utilizando AAnet. Se muestran los valores de la función de pérdida relativa en el dataset 'Codones' ('Loss'), al finalizar el entrenamiento para el rango de entre 1 y 7 arquetipos. Las función alcanza un codo entre 2 y 4 arquetipos para el set de datos elegido. Ensayos de más de 7 arquetipos arrojaron valores de 'Loss' comparables a los alcanzados con 7 arquetipos [49].

Las aristas de las figuras resultantes dan cuenta de la transformación que sufren los datos para ser incluidos en un politopo de 3 o 4 vértices (Fig. 20B y 20C). Los arquetipos se ubican por fuera de los puntos observados, resultando en un conjunto de instancias que se ubican por dentro del espacio que queda delimitado en los casos de 3 y 4 arquetipos. En el caso de buscar 4 arquetipos, uno de ellos se halla alejado del grupo de las variables, desformando el politopo. En el caso de la búsqueda de 2 únicos arquetipos, los organismos se alinearon alrededor de una curva delimitada por estos arquetipos, mostrando cómo se organizan los organismos en una única dimensión construida desde los dos arquetipos hallados. La búsqueda de 3 arquetipos los localiza adecuadamente en torno a la forma que toma el conjunto de organismos, sugiriendo que este sería la cantidad óptima de arquetipos para este dataset.

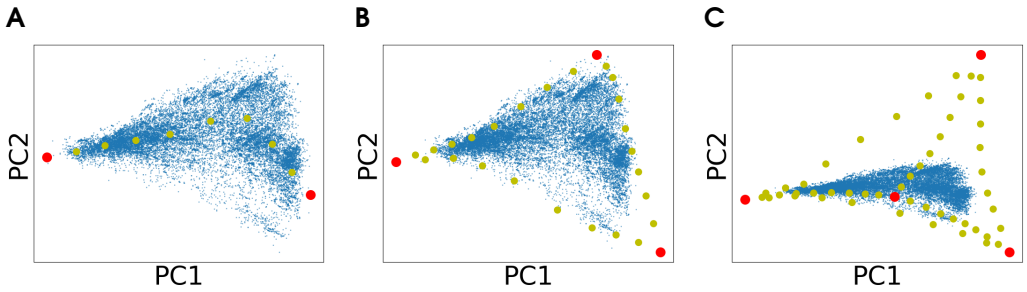

Figura 20: Análisis preliminar de arquetipos utilizando AAnet. Se muestran los organismos proyectados en las dos primeras componentes principales (azul) junto con las posiciones de los arquetipos (rojo) y las aristas de los politopos obtenidos (verde). Se observan los politopos obtenidos con 2 (Fig. A), 3 (Fig. B) y 4 (Fig. C) vértices, que intentan reproducir el conjunto de datos. En el caso de 4 arquetipos, uno de ellos se observa muy alejado del conjunto de datos original.

Este método logra ubicar a los arquetipos en torno al conjunto de organismos y, mediante la observación de la Figs. 19 y 20, determinar un número óptimo de arquetipos. La alta sensibilidad a los cambios de los hiperparámetros y la baja reproducibilidad deja este método como una alternativa válida en datasets de muy alta dimensionalidad. Para este trabajo, sin embargo, un modelo más simple puede lograr resultados consistentes con las dimensiones de nuestros datasets, a la vez que facilitan el análisis.

5.3.2 ParTI

Este método se basa en localizar arquetipos a partir de considerar cada organismo como combinaciones lineales de cada arquetipo. Utiliza distintos algoritmos alternativos capaces de reducir los errores de estimación de las localizaciones de los organismos a partir de las combinaciones de arquetipos.

5.3.2.1 Número óptimo de arquetipos

Como primer paso en este método, se intenta hallar el número óptimo de arquetipos para cada set de datos. La búsqueda, realizada sobre los 3 datasets primarios con un número creciente de arquetipos, mostró curvas de errores estándar similares. Al igual como se realizó con el método AAnet, se localiza el número óptimo de arquetipos en el codo de la curva. Las curvas de ParTI muestran codos aparentes en 4 arquetipos para los 3 datasets primarios (Fig. 21A). Esto se corrobora con un criterio cuantitativo, a partir de calcular las máximas distancias desde cada punto a una recta que una los puntos máximos y mínimos de cada curva (Fig. 21B).

Cada dataset fue analizado de manera independiente, por lo que, si bien se esperan resultados similares, es importante destacar que el número óptimo de arquetipos es el mismo en los tres casos. Se sustenta, entonces, la propuesta de que 4 arquetipos son capaces de explicar adecuadamente los 3 datasets de estudio.

Si bien el criterio para hallar el punto de quiebre arroja resultados consistentes, los codos hallados distan de poder considerarse como codos definidos que inequívocamente describen el número óptimo de arquetipos, dejando lugar a la posibilidad de que el número de arquetipos recomendado no sea el correcto. De todas formas, el poder hallar un número reducido de arquetipos en todos los datasets, sin observar variaciones profundas ni una curva de error sin quiebres, permite suponer la existencia de un número limitado de arquetipos (no más de 5) en los conjuntos de datos, que pueden ayudar a la interpretación de su distribución. Dada la dimensionalidad inicial, y las curvas observadas, se fija en 4 el número de arquetipos de trabajo.

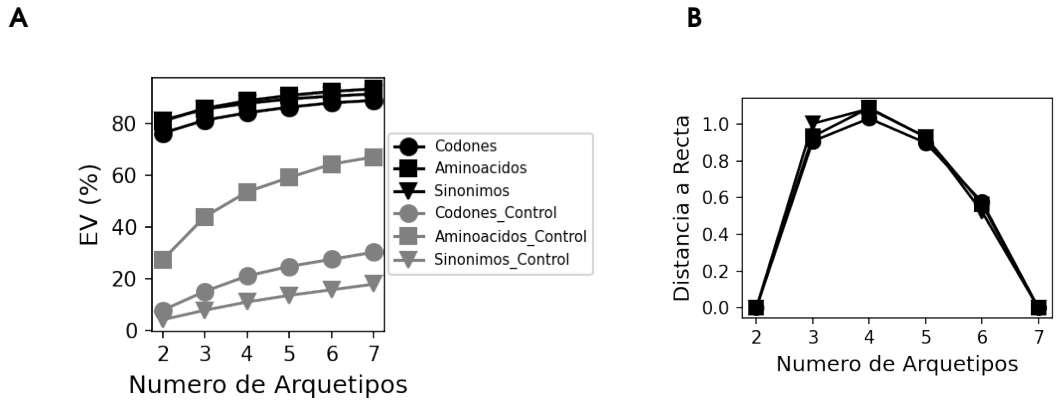

Figura 21: Identificación del número óptimo de arquetipos utilizando ParTI. A. Se muestra la varianza explicada porcentual (EV %) para un creciente número de arquetipos. Se busca el codo de cada curva a fin de establecer el número óptimo de arquetipos. Si bien las curvas no muestran un codo evidente, los autores proponen un criterio cuantitativo para definirlo a partir de maximizar la distancia de cada punto a la recta que une los puntos máximos y mínimos. Este criterio de ParTI para estos datasets señalan 4 arquetipos para todos los casos. Se incluyeron los resultados obtenidos a partir de correr el mismo método en datasets de control. Los datasets control se construyeron a partir de los 3 datasets primarios pero mezclando aleatoriamente los valores de cada una de las variables por separado, de forma de romper la estructura general de los datos mientras se mantienen las distribuciones dentro de cada variable. B. Se muestran las distancia entre cada punto y la recta que une el máximo y mínimo de cada curva para los tres datasets primarios. El valor máximo de distancia señala el número óptimo de arquetipos según ParTI.

Finalmente, se aplicó el mismo método pero en 3 datasets tomados como control. Estos datasets permiten evaluar el comportamiento del método de búsqueda en datasets sin estructura y compararlo con sus resultados en los datasets primarios.

Se espera que estos datasets de control muestren un comportamiento marcadamente diferente de los datasets primarios. Estos datasets control se realizaron tras una mezcla aleatoria de todos los valores dentro de cada variable, de forma de romper las correlaciones en los datasets, manteniendo las distribuciones propias dentro de cada variable. Como es de esperar, estos nuevos datasets de control alcanzaron proporciones de varianza explicada muy bajas en comparación con los originales (Fig. 21A).

5.3.2.2 Ensayos de algoritmos alternativos

Tomando el dataset 'Codones', se utilizaron los 5 algoritmos de búsqueda de ParTI ('Sisal','MVES','MVSA','SDVMM' y 'PCHA') arrojando todos ellos 4 arquetipos en locaciones similares. Se utilizó la metodología de bootstrapping 'leave-one-out' para estimar un intervalo de confianza de cada posición [41]. En todos los casos se hallaron posiciones de los arquetipos comparables con intervalos de confianza adecuados. Además, los intervalos de confianza, con el dataset 'Codones', muestran robustez en el proceso (Fig. 22). Testeos con 3 o 5 arquetipos arrojaron resultados igualmente reproducibles en todos los casos. Este método no requiere de hiperparámetros, sólo de la elección del ajuste del politopo que no conlleva diferencias importantes ni de interpretación para el problema planteado. Además, el método de reducción de dimensionalidad mediante PCA garantiza que las nuevas dimensiones sean combinaciones lineales de las originales, facilitando su interpretación y reduciendo posibles distorsiones en la distribución de los organismos. Por estos motivos se eligió ParTI, en vez de AAnet, para el análisis de arquetipos sobre los 3 datasets primarios definidos ('Codones', 'Aminoácidos' y 'Sinónimos'). Finalmente, dada la concordancia entre las posiciones obtenidas por los diferentes algoritmos, se definió 'Sisal', de manera arbitraria, como la metodología de referencia para el ajuste del politopo.

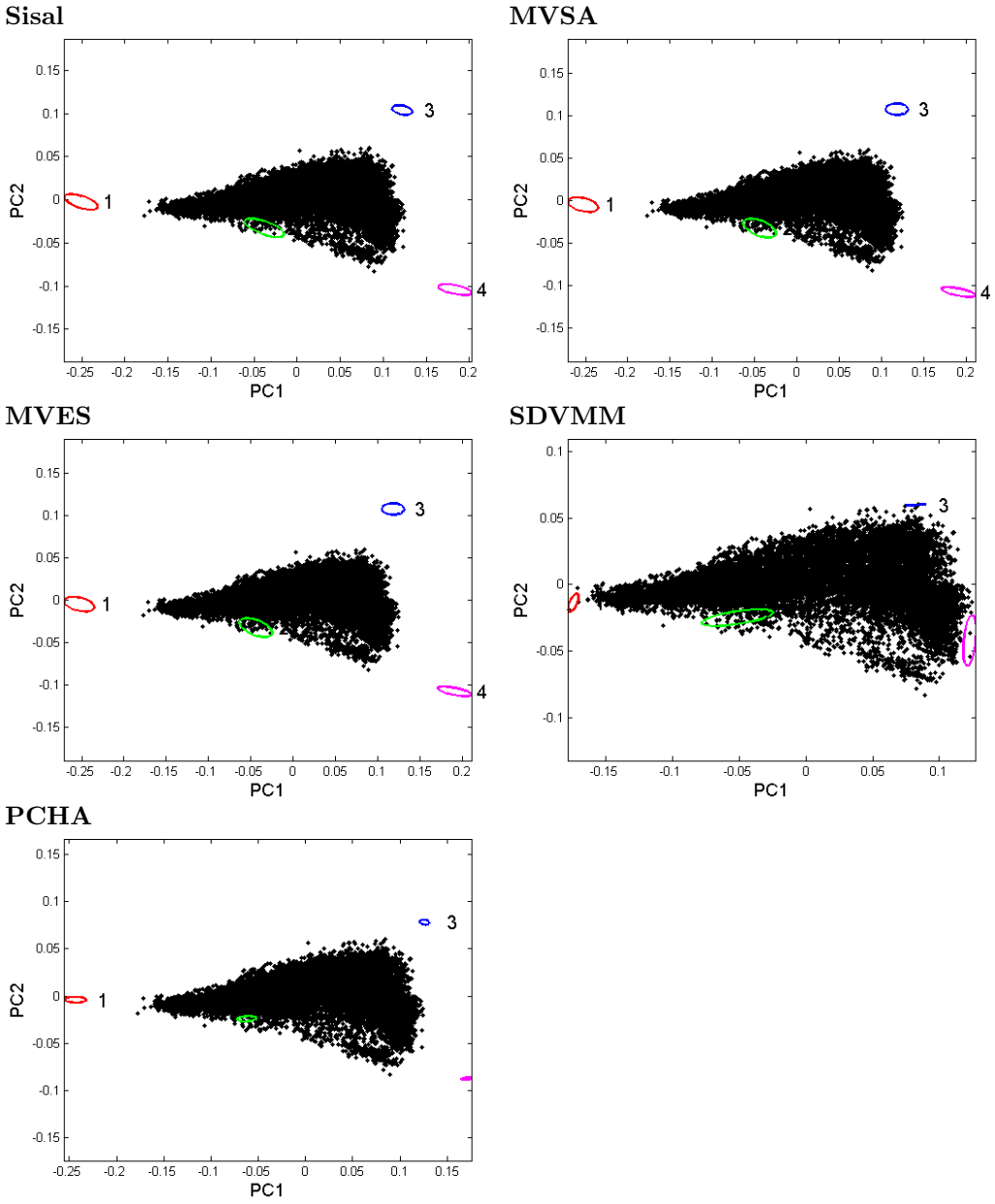

Figura 22: Localización de arquetipos utilizando ParTI en el dataset 'Codones'. Se realiza la búsqueda de 4 arquetipos con cada una de las metodologías de ajuste del politopo. Se muestran los resultados de los diferentes métodos con las ubicaciones de cada arquetipo marcado en color. Los círculos definen un área dentro de un desvío estándar tomado del resultado del bootstrapping.

5.4 Búsqueda de arquetipos

Los ensayos realizados con diferentes algoritmos muestran los arquetipos alrededor del conjunto de organismos (Fig. 22). Sin embargo, la búsqueda de 4 arquetipos se realiza sobre un espacio de 3 dimensiones tras aplicar la reducción de dimensionalidad mediante PCA, por lo que observar la distribución en 3 dimensiones permite apreciar mejor las posiciones relativas (Fig. 23). Los arquetipos encontrados

en el dataset 'Codones', utilizando ParTI y el método 'Sisal', muestran una tendencia general que da cuenta de 4 organismos extremos compatibles con los puntos que representan cada uno de los organismos observados en las bases de datos (Fig. 23A). Los arquetipo hallados mediante este método se encuentran por fuera del conjunto de puntos observados y parecen constituir los organismos arquetípicos de los vértices del politopo. Con el dataset 'Aminoácidos' se aplica la misma metodología en la búsqueda de 4 arquetipos a efectos de localizarlos en el espacio de abundancias de aminoácidos. El método, nuevamente, logra ubicar a los arquetipos por fuera de las observaciones y parecen representar los vértices extremos del conjunto de los puntos en el espacio (Fig. 23B). Finalmente, el mismo método se aplicó sobre el dataset 'Sinónimos' con el objeto de localizar a los 4 arquetipos en este espacio de variables. Al igual que con los otros dos casos, los arquetipos se localizaron rodeando las observaciones por fuera del conjunto de puntos (Fig. 23C).

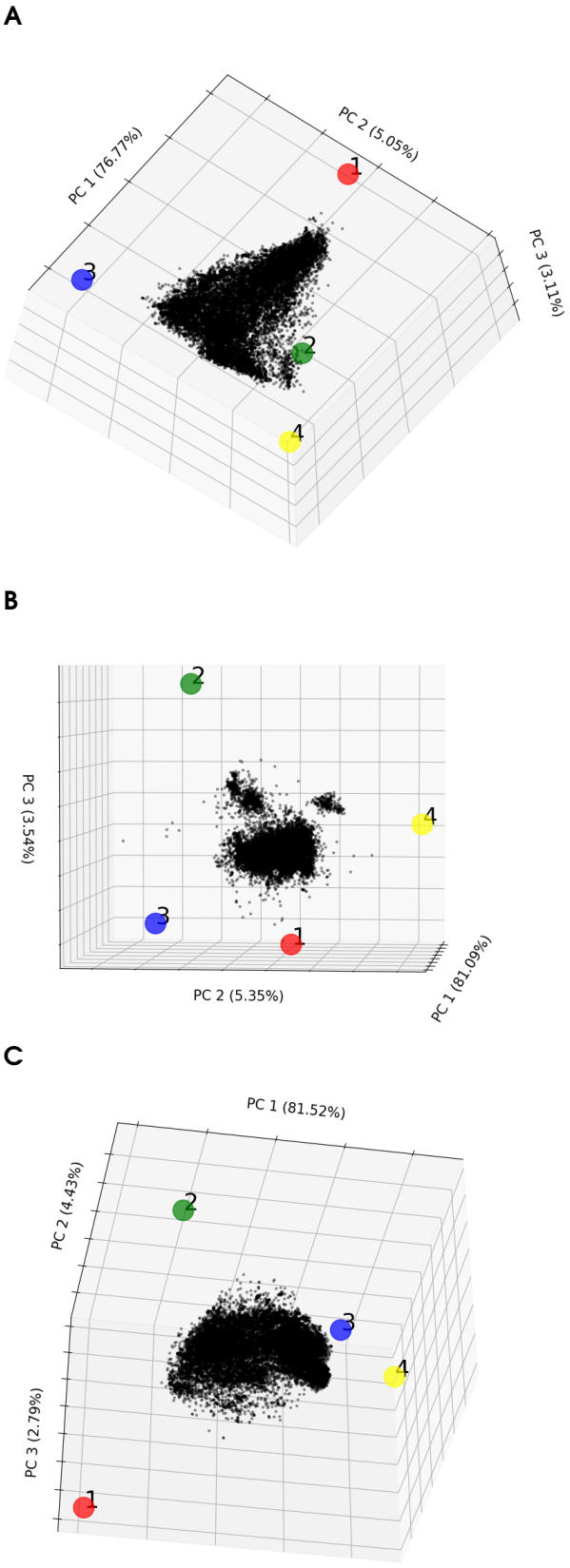

Figura 23: Localización de arquetipos. A: Se observan los 4 arquetipos hallados junto con los organismos proyectados sobre las primeras 3 componentes principales en el dataset 'Codones'. B y C: se aplica el mismo método para los dataset 'Aminoácidos' (B) y 'Sinónimos' (C). Los resultados fueron obtenidos para cada dataset por separado utilizando el método 'Sisal' de ParTI. La búsqueda de 4 arquetipos en cada dataset arroja localizaciones compatibles con las distribuciones de puntos observadas mediante PCA y en la evaluación del método de ParTI (Fig. 22).

La metodología empleada parece poder aplicarse a los 3 datasets de estudio con resultados compatibles con lo esperado. En todos los casos los arquetipos parecen representar los vértices del politopo que se intenta hallar. Por otro lado, las formas obtenidas entre los datasets 'Codones' y 'Sinónimos' sugieren una correlación entre las posiciones de los arquetipos. El dataset 'Aminoácidos' muestra una topología diferente y una distribución de los puntos algo menos continua.

La aplicación de ParTI permitió localizar a los arquetipos en los tres datasets primarios haciendo posible su posterior caracterización. Además, los sucesivos algoritmos de búsquedas presentes en ParTI mostraron que los arquetipos se localizan en las mismas posiciones relativas consistentemente, validando de alguna manera, la aplicación de esta metodología sobre los datasets primarios.

5.5 Análisis de localización de arquetipos

La posiciones de los arquetipos permiten explorar sus variables presentes en los datasets primarios. Además, la presencia de variables adicionales, que caracterizan cada uno de los organismos, puede enriquecer este análisis a partir de evaluar los organismos cercanos a cada arquetipo. Se proponen, entonces, características de cada arquetipo a partir de una gran cantidad de variables de diversas fuentes.

5.5.1 Características primarias de los arquetipos

Cada uno de los arquetipos hallados puede localizarse en el espacio de variables originales permitiendo una primera caracterización. Sin embargo, la localización de los arquetipos no está sujeta a las limitaciones impuestas sobre los organismos, pudiéndose obtener valores de proporciones de suma distinta de uno o valores fuera de los rangos físicos posibles para cada variable. Los valores que toma cada variable para cada arquetipo aporta una primera aproximación de sus características y la posibilidad de observar sus principales diferencias.

Los tres datasets primarios construídos comparten una semántica común asociada al código genético. Las relaciones de traducción y sesgos se relacionan íntimamente con las relaciones entre codones y los aminoácidos que codifican, por lo que se representan los resultados de los tres datasets dentro de la estructura del código genético (Fig. 24, 25 y 26).

Se utilizaron los resultados de la búsqueda de los 4 arquetipos en el dataset

'Codones' para extraer sus valores en cada una de las variables originales, es decir, las abundancias de sus 61 codones codificantes. Cada columna da cuenta de un arquetipo diferente, donde los colores representan las abundancias de cada uno de los diferentes codones. Los arquetipos similares tienen patrones de color similares en sus columnas. De los arquetipos hallados, dos de ellos se muestran similares entre sí con codones de mayor contenido GC en el tercer nucleótido (arquetipos 3 y 4), si bien existen varios codones que no cumplen esta tendencia como ser AGG y UUG. El arquetipo 1 se asocia con abundancias de codones con menor contenido GC en la tercera posición, sobre todo en los codones que contienen C o G en la segunda posición. El arquetipo 2 no parece manifestar valores extremos, con excepción de UUA (Leu) y GCG (Ala), donde toma valores muy bajos. Aparece destacado el codón CUG (Leu) como el que posee valores más extremos, tanto por valores máximos como mínimos en los arquetipos 1 y 3. Otros codones con valores extremos son UUA (Leu), GCC (Ala), GCG (Ala) y AAA (Lys).

El método se aplica nuevamente sobre el dataset 'Sinónimos' a fin de localizar a los arquetipos de este dataset. Luego se extraen los valores de cada variable de los 4 arquetipos hallados y se representan en una estructura similar al dataset 'Codones' para facilitar su interpretación (Fig. 25). En el dataset 'Sinónimos', al igual que en el caso del dataset 'Codones', aparecen 2 arquetipos relacionados (arquetipos 3 y 4). Sin embargo, en este caso el arquetipo 4 aparece con valores más extremos en la tercera posición. Esto se destaca particularmente en los casos donde solo hay dos codones sinónimos por aminoácido. En el arquetipo 2 aparecen valores elevados en los codones CAA (Gln), AAA (Lys) y GAA (Asp) acompañados por valores bajos en sus codones sinónimos CAG, AAG y GAG respectivamente. En el caso del arquetipo 1, se muestran algunos codones con un valor muy bajo en comparación con el resto de sus codones sinónimos como ser los casos de CUG (Leu), ACC (Thr), CGC (Asn) y GCC (Gly).

Por último, se repite la metodología de búsqueda en el dataset 'Aminoácidos' a efectos de ubicar los 4 arquetipos en este espacio de variables. Se extraen los valores de abundancias de cada aminoácido de cada uno de los arquetipos de este dataset para compararse. Nuevamente, se representan en una estructura similar a la utilizada para los otros dos datasets (Fig. 26). En el caso de los arquetipos del dataset 'Aminoácidos', las localizaciones ubican al arquetipo 1 con baja Ala, al arquetipo 3

con alta Ala, al arquetipo 2 alto Ser y un último arquetipo 4 con alta Ala y baja Gln. Además, se observa que el arquetipo 2 presenta un valor de Cys mayor que los demás.

Estos resultados son consistentes con lo observado en las correlaciones de las abundancias de aminoácidos (Fig. 18). Al analizar las correlaciones entre los aminoácidos, se muestran dos claros grupos que correlacionan entre sí y un número limitado de aminoácidos que tienen comportamientos algo más independientes. Esta tendencia general se repite al analizar los arquetipos, que no escapan a estas relaciones de correlación entre aminoácidos. Se destaca el aminoácido Leu en el arquetipo 3 dado que constituye una fuerte diferencia con respecto al arquetipo 4, a pesar de correlacionar en la mayoría de los demás aminoácidos. Los arquetipos hallados en los tres datasets primarios poseen diferentes valores de abundancias. Sin embargo, características más relevantes se hacen evidentes al analizar su posición relativa dentro del conjunto de organismos (ver siguiente sección).

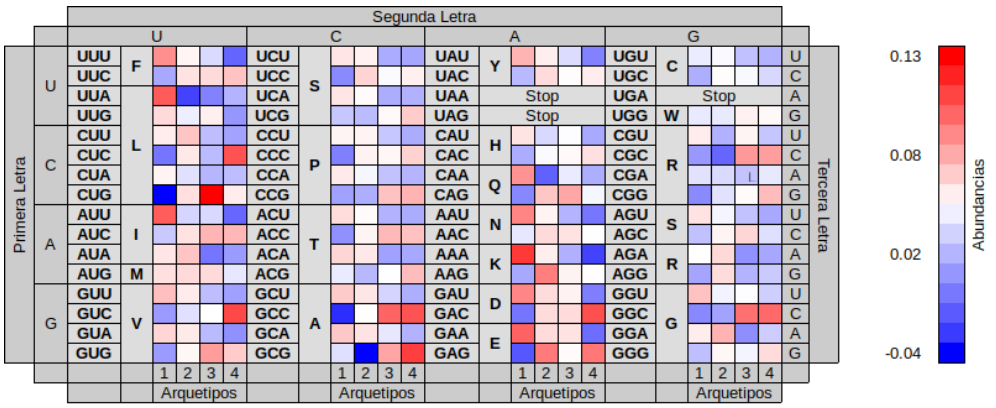

Figura 24: Características primarias de los arquetipos en el dataset 'Codones'. Se muestran los valores que toma cada variable primaria del dataset 'Codones' en las posiciones de cada arquetipo. Se discrimina por cada codón agrupado según su ubicación en el código genético.

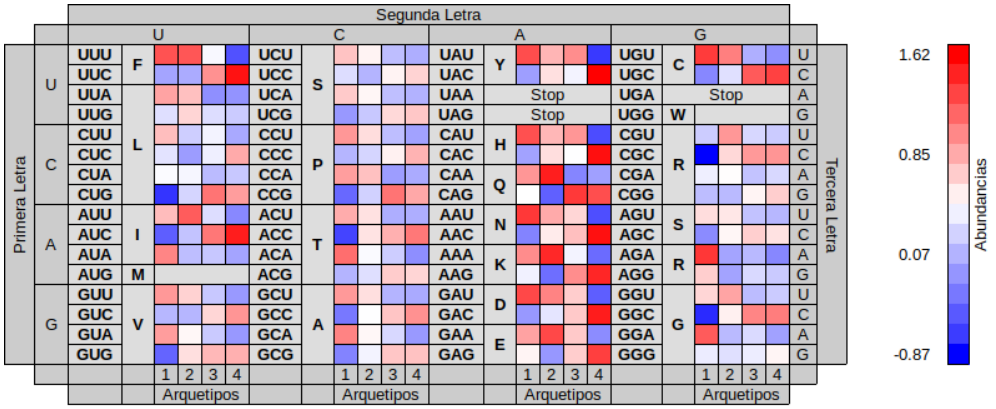

Figura 25: Características primarias de los arquetipos en el dataset 'Sinónimos'. Se muestran los valores que toma cada variable primaria en las posiciones de cada arquetipo, estructurados bajo el código genético. Met (M) y Trp (W) no contienen valores dado que su valor de abundancia relativa r_j siempre es igual a 1 ya que solo tienen un codón cada uno.

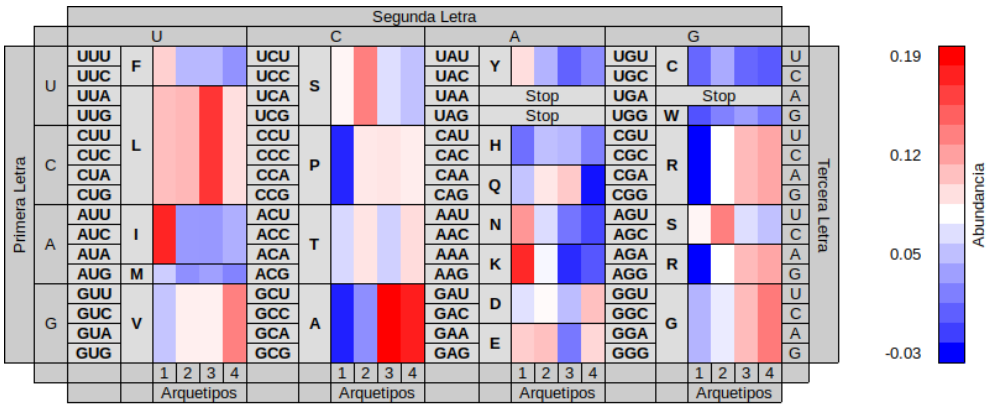

Figura 26: Características primarias de los arquetipos en el dataset 'Aminoácidos'. Se muestran los valores que toma cada variable primaria en el dataset 'Aminoácidos' de abundancias de aminoácidos. Los valores toman forma de mapa de calor en cada aminoácido ordenados según el código genético.

5.5.2 Distancias a los arquetipos

Si bien las características primarias dan una idea general de los arquetipos, las relaciones entre los arquetipos de diferentes datasets no quedan claras del todo. Para poder definir estas relaciones entre los arquetipos, se propone considerar una posición relativa con referencia al conjunto de organismos que sí se encuentran representados en cada uno de los datasets. En cada dataset se hace posible tomar una distancia entre cada organismo y cada arquetipo de forma de poder compararlas.

Para cada arquetipo dentro de cada uno de los 3 datasets primarios, se toma la distancia euclídea a cada uno de los organismos del mismo dataset. De esta manera, cada arquetipo se puede caracterizar como una serie de distancias a todos

los organismos. Arquetipos con una misma posición relativa al grupo de organismos deberán tener una correlación elevada entre sí, de esta manera es posible comparar similitudes en la posición relativa de arquetipos de distintos datasets. A efectos de realizar una comparación entre arquetipos, se calculan los coeficientes de correlación de Spearman entre ellos. Con estos valores de correlación es posible construir un correlograma de los arquetipos de los tres datasets al mismo tiempo (Fig. 27).

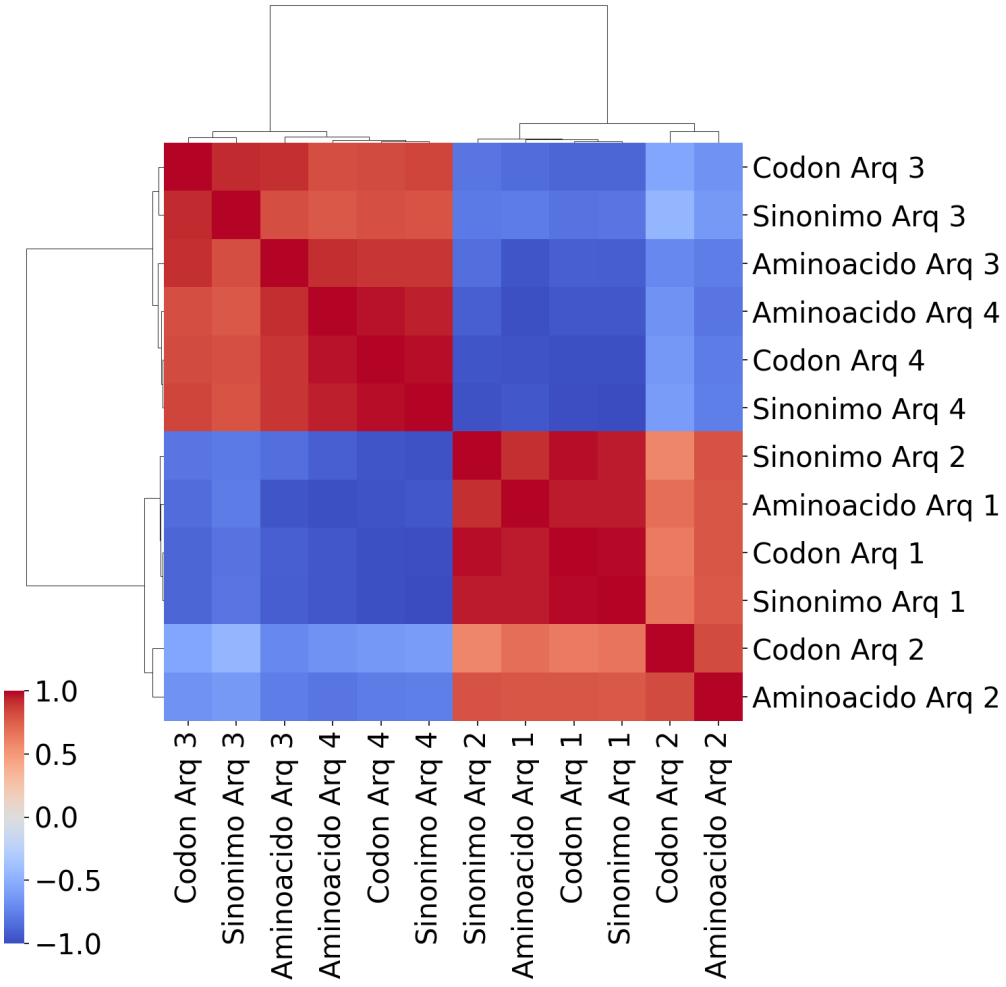

Figura 27: Correlograma de posiciones de arquetipos. Se muestran los coeficientes de correlación de Spearman para las distancias de cada arquetipo a los organismos de su propio dataset. Correlaciones cercanas a 1 indican una posición relativa de los arquetipos frente al conjunto de los organismos. Se observa que los arquetipos 1, 3 y 4 se agrupan juntos por UPGMA, dando cuenta de una consistencia de estos arquetipos en los 3 dataset de trabajo. El arquetipo 2 se halla en posiciones relativas diferentes en los distintos datasets primarios, ya que, si bien los arquetipos correspondientes a los datasets 'Codones' y 'Aminoácidos' se agrupan juntos, el arquetipo 2 correspondiente al dataset 'Sinónimos' se halla fuera del grupo.

Además, se aplica el algoritmo de agrupamiento de UPGMA a fin de dar una mejor idea de parecidos y diferencias entre ellos. Se observan dos grandes grupos anticorrelacionados: arquetipos 1 y 2 por un lado, y los arquetipos 3 y 4 por el

otro. Además de esa tendencia general, se observa que los arquetipos 1, 3 y 4 se agrupan entre sí en los 3 datasets primarios, dando cuenta de la consistencia de estos arquetipos en los tres datasets. En el caso del arquetipo 2, se observa que sus localizaciones son más dispersas y, en el caso del dataset 'Sinónimos', se encuentra disociado del arquetipo 2 en los datasets 'Codones' y 'Aminoácidos'. Este arquetipo no presenta valores de contenido GC extremos, como sí ocurre en los otros 3 arquetipos. Esto hace que las abundancias características del arquetipo 2 tengan menos restricciones debido a este factor de contenido GC, lo que podría justificar, al menos hipotéticamente, la mayor variabilidad de la ubicación del arquetipo 2 en los diferentes datasets primarios.

5.6 Caracterización de los arquetipos

La posiciones de los arquetipos permiten evaluar las características asociadas a los datasets primarios. Sin embargo, variables adicionales, propias de a cada organismo presente en el dataset, también pueden estar relacionadas con los arquetipos hallados. Se analiza, entonces, un conjunto de variables adicionales de 'Enriquecimiento' propias de los organismos y su relación con los arquetipos descriptos en la sección anterior.

5.6.1 Variables categóricas: Análisis general

Las búsquedas en las bases de datos de NCBI de los diferentes TaxID de cada organismo proveyeron el linaje completo de cada organismo. Los registros de los linajes se componen de los taxones anidados de cada especie, por lo que fue posible agruparlos por los taxones más representados en los datasets construidos. Al agrupar los organismos por los tres dominios principales y observando los taxones de mayor jerarquía, se pueden vislumbrar tendencias generales que también se repiten en taxones de menor nivel.

Tomando las primeras 3 componentes principales del PCA realizado, se grafican los diferentes organismos coloreados por su clasificación taxonómica. Del total de organismos, 21.337 corresponden a Bacteria, 1.033 corresponden a Eukaryota y 550 a Archaea. Se agregan los arquetipos dentro de cada dataset para comparar sus posiciones relativas y poder observar las asociaciones entre los taxones y los arquetipos. Diferentes grupos de organismos pueden identificarse en los diferentes

datasets. En los 3 datasets primarios analizados se observa que cada dominio presenta un lugar en los espacios de variables propio (Fig. 28). Esto se observa más nítidamente en el dataset de abundancias de aminoácidos (Fig. 28B) donde aparece una separación más clara entre dominios. Además, las posiciones relativas de los arquetipos frente al conjunto de los organismos permite vislumbrar una potencial asociación entre los arquetipos y dominios taxonómicos. Nuevamente, esto aparece más marcadamente en el caso del dataset 'Aminoácidos'.

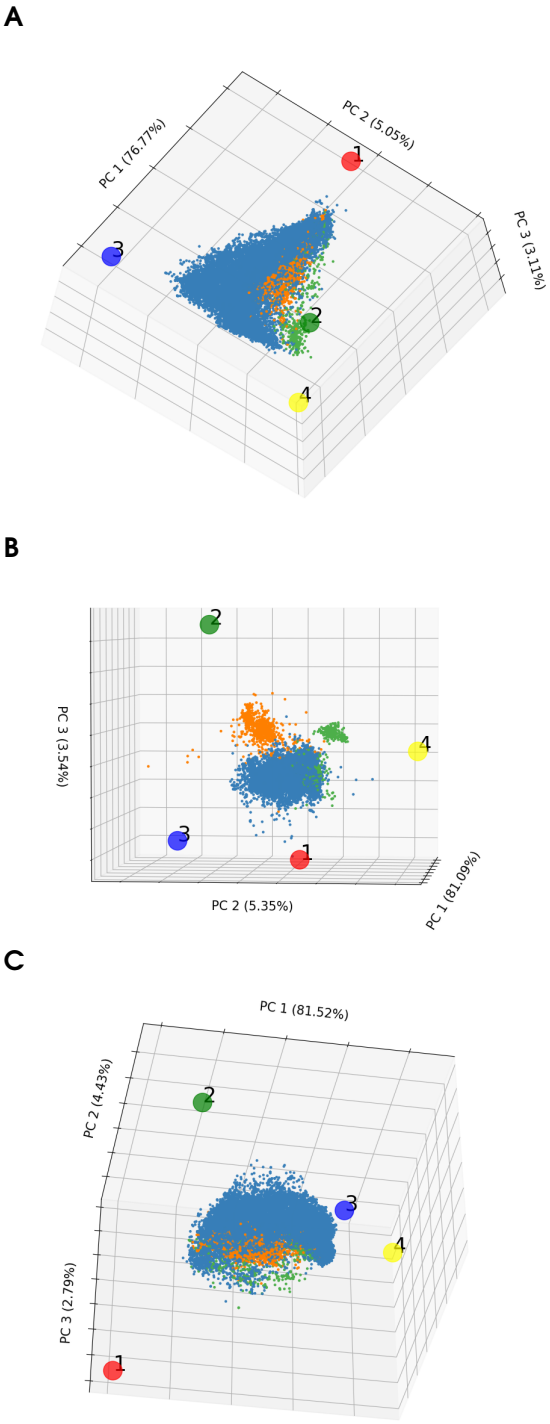

Figura 28: Dominios taxonómicos y arquetipos. A: Se observan los 4 arquetipos hallados junto con los organismos proyectados sobre las primeras 3 componentes principales en el dataset 'Codones'. B y C: se aplica el mismo método para los dataset 'Aminoácidos' (B) y 'Sinónimos' (C). Los resultados fueron obtenidos para cada dataset por separado. La búsqueda de 4 arquetipos en cada dataset arroja localizaciones compatibles con las distribuciones de puntos observadas mediante PCA y en la evaluación del método de ParTI (Fig. 22). Los dominios toman espacios propios, si bien existe superposición entre ellos dando cuenta de las similitudes entre organismos filogenéticamente más cercanos.

Se repite la estrategia para los dominios planteada anteriormente tomando las

primeras 3 componentes principales pero únicamente del dataset 'Codones' y se colorean todos los organismos según su relación al taxón Bacteria. Se seleccionan los 4 taxones mayoritarios de la jerarquía inmediatamente inferior a Bacteria de forma de evaluar su dispersión en el espacio y comparar con lo observado para el caso de los dominios. Al observar los taxones inferiores a Bacteria en el dataset 'Codones' (Fig. 29A), se observa una superposición entre ellos, aunque los taxones mantienen áreas distintivas propias indicando preferencias por el uso de ciertos conjuntos de codones. Si se repite el análisis sobre el dataset 'Aminoácidos', se reproduce lo observado en el dataset 'Codones'.

Al igual que en el caso del dominio Bacteria, se repite la estrategia pero sobre el dominio Eukaryota. Se toman las 3 primeras componentes principales y se colorean todos lo organismos según su relación con el dominio Eukaryota. Se tienen en cuenta los dos phylum mayoritarios para podes categorizar a los organismos dentro de Eukaryota y así poder observar su distribución. En el caso de organismos del dominio Eukaryota, la separación se hace menos evidente (Fig. 29B). Algo similar ocurre en el caso de los oranismos del dominio Archaea, donde se observan los diferentes phylum en grupos difusos (Fig. 29C). Conforme se avanza hacia taxones de menor jerarquía, las diferencias se desdibujan, pero sin dejar de traslucir un rastro filogenético en los usos de codones y aminoácidos evidenciando la preferencia de los taxones por sectores propios de los espacios de variables. Los grupos de organismos se superponen en una gran porción, pero no dejan de ocupar espacios separados, dando cuenta de la influencia de la filogenia sobre el uso de codones y aminoácidos. Aparecen, sin embargo, algunos taxones más fácilmente separables que otros.

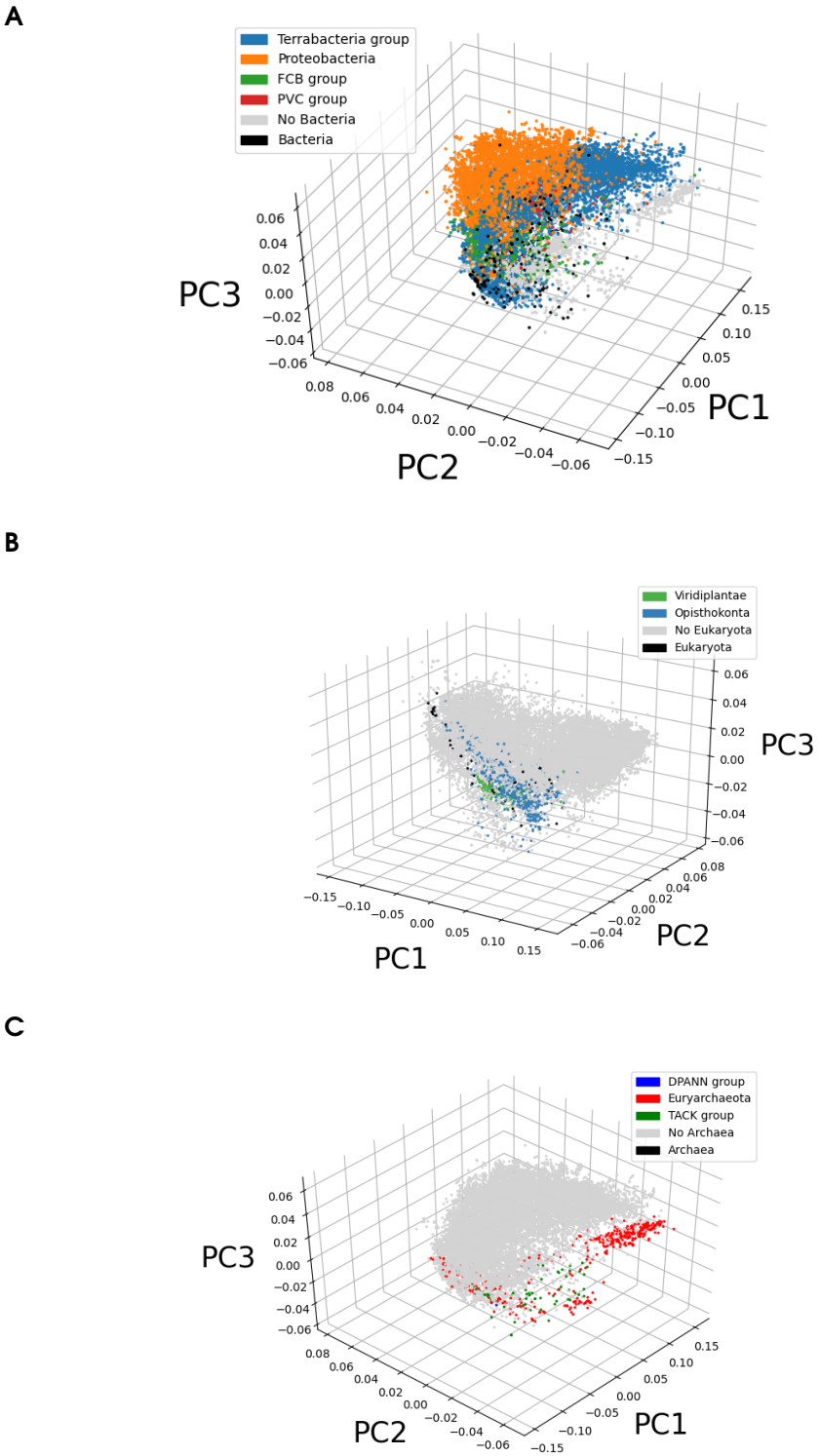

Figura 29: Distribución de taxones de menor jerarquía en dataset 'Codones'. A: Se muestran en las 3 primeras componentes principales los taxones derivados de 'Bacterias': Proteobacteria, Terrabacteria y FCB. El resto de los organismos que corresponden a Bacteria se marcan en negro, los que no corresponden a Bacteria se marcan en gris. B: Al igual que en 'A', se muestran los taxones derivados de Eukaryota: Ophistokonta y Viridiplantae. El resto de los organismos que corresponden a Eukaryota se marcan en negro, los que no corresponden a Eukaryota se marcan en gris. C: Al igual que en 'A' y 'B', se muestran los taxones derivados de Archaea: DPANN, TACK y Euryarchaeota. El resto de los organismos que corresponden a Archaea se marcan en negro, los que no corresponden a Archaea se marcan en gris. En los 3 casos se muestra la totalidad de los organismos de manera de evidenciar el contexto. Los diferentes grupos toman sectores propios en el espacio aunque se vislumbra una fuerte superposición.

5.6.2 Variables categóricas: Distancias entre taxones y arquetipos

Para analizar la asociación entre taxones y arquetipos, las diferentes categorías se evaluaron frente a las distancias a los arquetipos. Para esto se utilizó el dataset 'Codones', no obstante, los otros dos datasets primarios arrojan resultados similares. Tomando las distancias de cada organismo a cada arquetipo, se agruparon los diferentes organismos según su taxonomía y se analizaron las distancias por cada grupo por separado. Se construyeron gráficos con las distancias a cada arquetipo de cada uno de los grupos de forma de mostrar si existe una asociación evidente (Fig. 30 y Fig. 31).

Los tres dominios principales (Fig. 30) permiten ver una tendencia que se repite en taxones de rango inferior. Bacteria se muestra con una amplia dispersión, si bien se asocia más al arquetipo 3. En el caso de Eukaryota, cada una de las distancias a los arquetipos se encuentra con menor dispersión en general. Estos organismos aparecen más cercanos al arquetipo 2. En el caso de Archaea, se repite lo observado para el caso de Bacteria, siendo el arquetipo 1 el más alejado con respecto a los restantes, y el arquetipo 2 el más cercano.

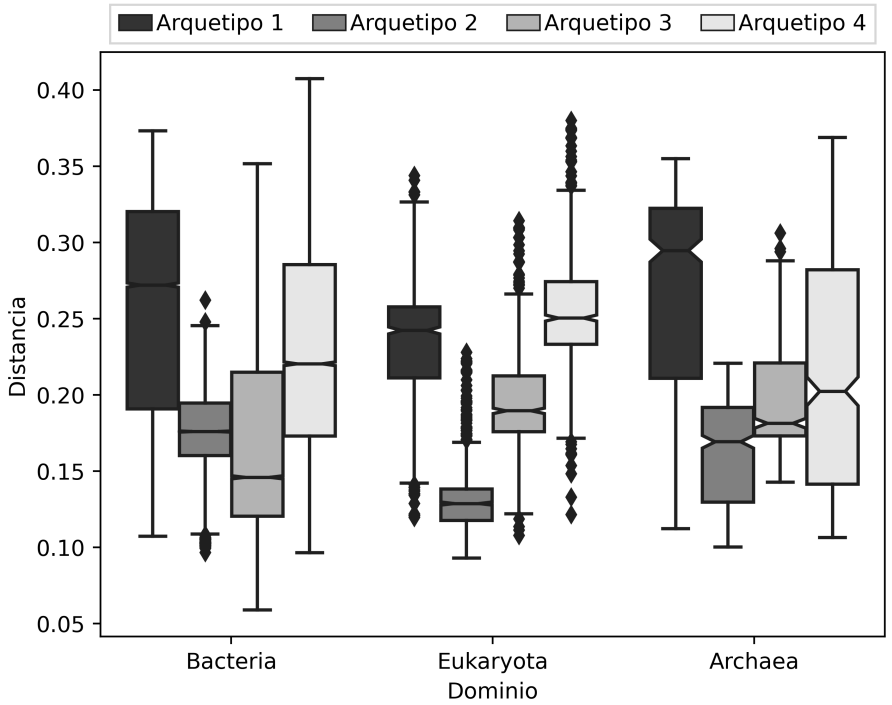

Figura 30: Relación entre dominios y arquetipos. Se muestran las distancias a cada arquetipo para cada dominio, utilizando el dataset 'Codones'. Valores más bajos implican una asociación entre el dominio y el arquetipo. Se muestra el boxplot con el intervalo de confianza (95 %) de las medianas como cuñas. El gran número de organismos involucrados genera que las cuñas sean muy reducidas dada la distribución de los valores.

Utilizando el mismo enfoque, se toma únicamente el conjunto de organismos del dominio Bacteria y se separa en sus phylum mayoritarios. Se analizan las distribuciones de las distancias a cada arquetipo de cada uno de estos grupos (Fig. 31). Si bien las tendencias no son tan fuertes como en el caso de los dominios, se pueden observar algunas tendencias. Aparecen algunos grupos con una gran dispersión de distancias hacia cada arquetipo, como es el caso de 'Terrabacteria'. Este grupo se muestra más cercano a los arquetipos 2 y 3. Por el contrario, 'FCB' muestra una dispersión algo más reducida y se muestra más próximo a los arquetipos 1, 2 y 3. A pesar de esto, los valores límites de todos los taxones evaluados son semejantes, dando cuenta de que, en todos los grupos, existen organismos extremos que se alejan de la mayoría de los organismos de cada taxón. El grupo de 'Proteobacteria' parece asociado al arquetipo 3, aunque posee muchos organismos marcadamente alejados. En el caso de 'PVC', el número reducido de individuos incluidos evidencia un intervalo de confianza más grande para la mediana y se muestran más cercanos a los arquetipos 2 y 3. Estos taxones presentan distribuciones algo más superpuestas que las observadas para los dominios (Fig. 30), si bien una tendencia general parece

conservarse.

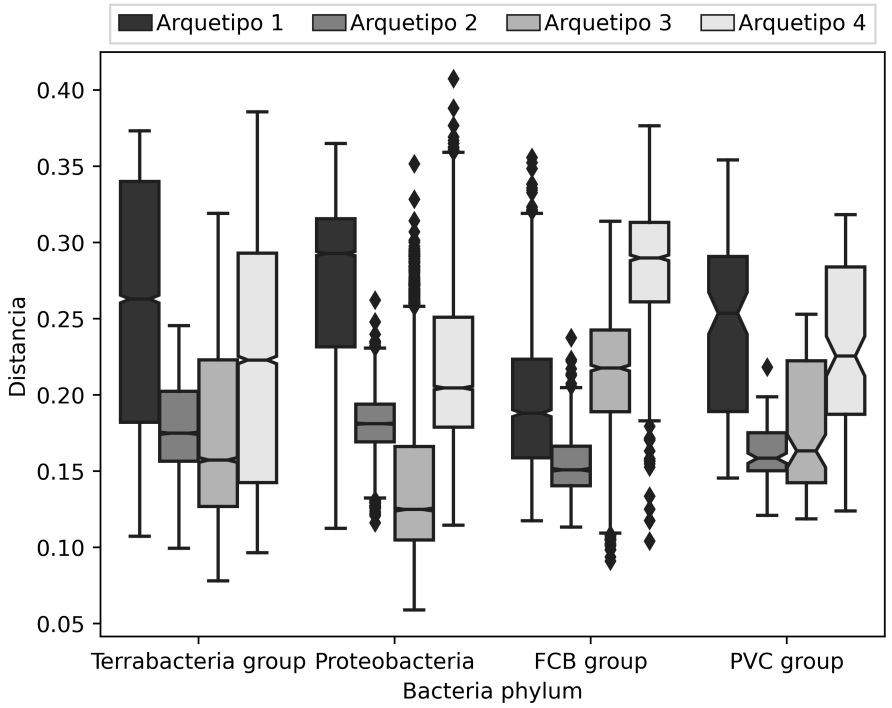

Figura 31: Relación entre Phylum de Bacteria y arquetipos. Se muestran las distancias a cada arquetipo en el dataset 'Codones' de los cuatro grupos de bacterias mayoritarios en dicho dataset. Las distribuciones se muestran superpuestas en varios casos. Se muestra el boxplot con el intervalo de confianza (95 %) de su mediana como cuñas.

5.6.3 Variables continuas: Análisis general

Los análisis sobre las variables de los datasets primarios de los arquetipos dejan afuera propiedades adicionales de los organismos que pueden estar relacionados con las posiciones de los arquetipos. Las variables de enriquecimiento propuestas surgen del uso de los codones o aminoácidos e incluyen determinaciones externas que permiten adicionar dimensiones independientes al análisis.

Para analizar en conjunto las 153 variables continuas de forma simultánea, se determinaron las distancias de correlación, a partir del coeficiente de correlación de Spearman de las diferentes variables de enriquecimiento, y las métricas de distancia d_E y proximidad p_E basadas en las distancias euclídeas a los arquetipos. Variables muy correlacionadas entre sí (correlación cercana a 1) tendrán distancias muy pequeñas (ver sección métodos). Si la correlación es elevada con respecto a p_E, esto implica que la variable en cuestión decrece conforme aumenta la distancia, lo que podría implicar una asociación entre esa variable y el arquetipo. Se construye una matriz de distancias capaz de representar las variables y los arquetipos al mismo

tiempo, y se aplica la metodología de MDS a fin de poder reducir las distancias a dos dimensiones y facilitar así su representación. Se incluyen, para cada arquetipo, las métricas de d_E y p_E. Se ubica en un mismo mapa a todas las variables de enriquecimiento continuas junto con d_E y p_E de los arquetipos de los 3 datasets primarios (Fig. 32).

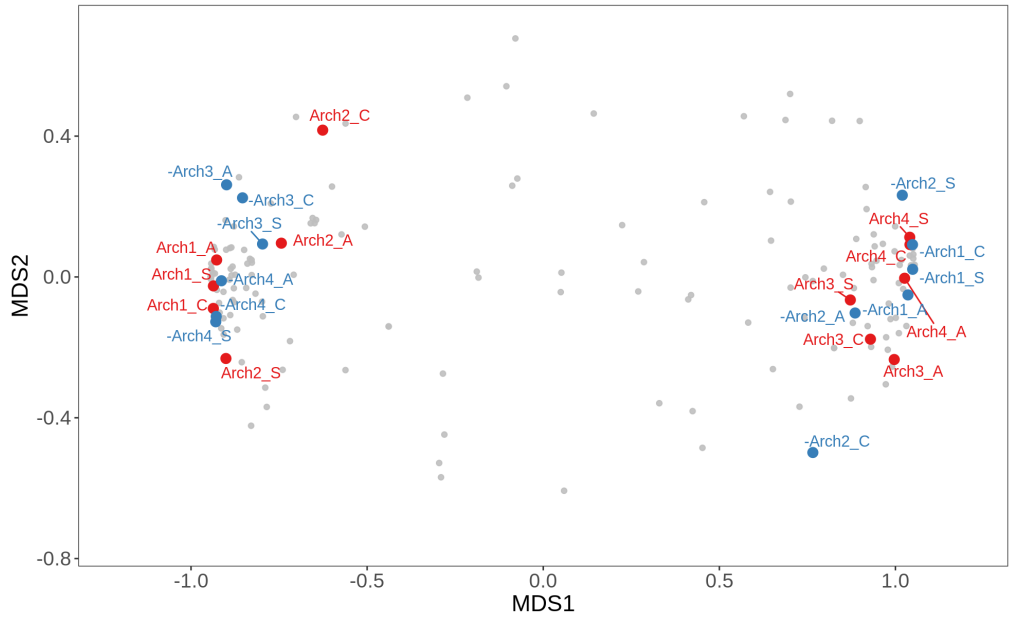

Figura 32: Distribución de los arquetipos de los diferentes datasets. A partir de una distancia de correlación *dSpear* entre los arquetipos, se construye un MDS en donde se muestran las dos mayores componentes del MDS. Para cada arquetipo, se calculan las métricas de distancias d_E y p_E, y se incluyen en el mapa. p_E toma el color rojo y d_E el color azul para cada arquetipo, en gris se muestran la totalidad de las variables de enriquecimiento continuas. Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'. Si una variable se encuentra cerca de otras, su correlación es elevada. Asimismo, si una variable se encuentra cerca de un arquetipo rojo, ese arquetipo se asocia con valores elevados de esa variable. Por el contrario, en caso de que la variable se encuentre cerca de un arquetipo azul, el arquetipo se asocia con valores bajos de esa variable. Finalmente, arquetipos cercanos entre si implica que existe una correlación entre sus distancias hacia el conjunto de organismos y por lo tanto ocupan una posición relativa similar. Si un arquetipo rojo se encuentra cercano a uno azul, se encuentran en puntos opuestos del conjunto de organismos.

Al igual que en el correlograma (Fig. 27), en el mapa construido por MDS (Fig. 32) se observa una concordancia entre los arquetipos 1, 3 y 4 de los tres datasets. En el caso del arquetipo 2 de cada dataset, el correspondiente a 'Sinónimos' aparece cercano al arquetipo 1 mientras que los asociados a 'Aminoácidos' y 'Codones' aparecen más alejados tomando un espacio más propio. Finalmente, los arquetipo 1 y 4 se acomodan en extremos opuesto del gráfico dando cuenta de su correlación negativa. La posición relativa de los arquetipos 1, 3 y 4 al resto de los organismos

en cada dataset parece coincidir lo suficiente como para postular que se localizan por principios comunes. En el caso del arquetipo 2, su posición parece más difusa, lo que coincide, nuevamente, con los observado en el correlograma (Fig. 27).

Tomando cada uno de los organismos, se generaron nuevas variables de enriquecimiento de interés. Con estas nuevas variables, se calcularon las distancias de correlación usando el coeficiente de correlación de Spearman entre cada una de las variables, incluyendo las métricas de distancias de los arquetipos. Utilizando el MDS, se lograron colocar cada una de las variables, y relacionarlas con variables cercanas y arquetipos. Variables cercanas en el mapa poseen una correlación positiva (Fig. 33). Cuanto más se alejen, su correlación será cada vez menor. En las secciones subsiguientes, se analizan los distintos grupos de variables de enriquecimiento según su significado.

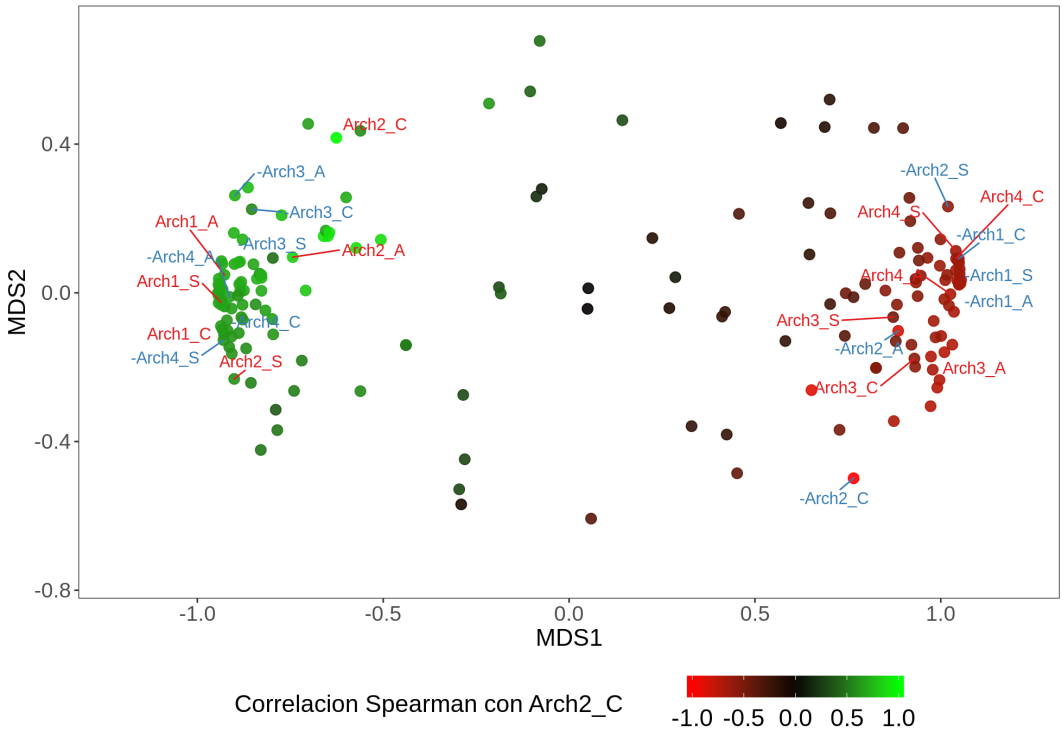

Figura 33: Correlaciones de Spearman en los mapas de variables. Tomando únicamente la métrica de p_E del arquetipo 2 del dataset 'Codones', se calcula la correlación de Spearman para cada variable del dataset 'Enriquecimiento'. Se muestran los valores de correlación en el gradiente de color. El valor máximo de 1 lo posee el arquetipo 2 azul de 'Codones' ya que se mide la correlación consigo misma. El valor mínimo -1 lo posee el arquetipo 2 rojo del dataset 'Codones' ya que d_E se encuentra anticorrelacionado con p_E.

El MDS intenta reproducir las distancias originales en un espacio de menor dimensión. El mapa construido intenta resumir las relaciones entre las variables y todos los arquetipos. Con el fin de evaluar la relación entre la correlación y las distancias observadas en el mapa, se tomaron las distancias existentes en el mapa de variables construido por MDS, entre cada uno de los arquetipos rojos del dataset 'Codones' y la totalidad de las variables (Fig. 34). En todos los casos, las distancias observadas se relacionan fuertemente y de forma negativa con la correlación de Spearman de cada una de las otras variables. Los arquetipos 1 y 4 muestran una relación lineal entre la distancia y el coeficiente de correlación de Spearman que se desea evaluar (Fig. 34A y Fig. 34D). En los casos de los arquetipo 2 y 3, la relación lineal aparece con una mayor dispersión (Fig. 34B y Fig. 34C). En todos lo casos se

evidencia que el mapa producto del MSD representa las correlaciones de Spearman entre variables consistentemente.

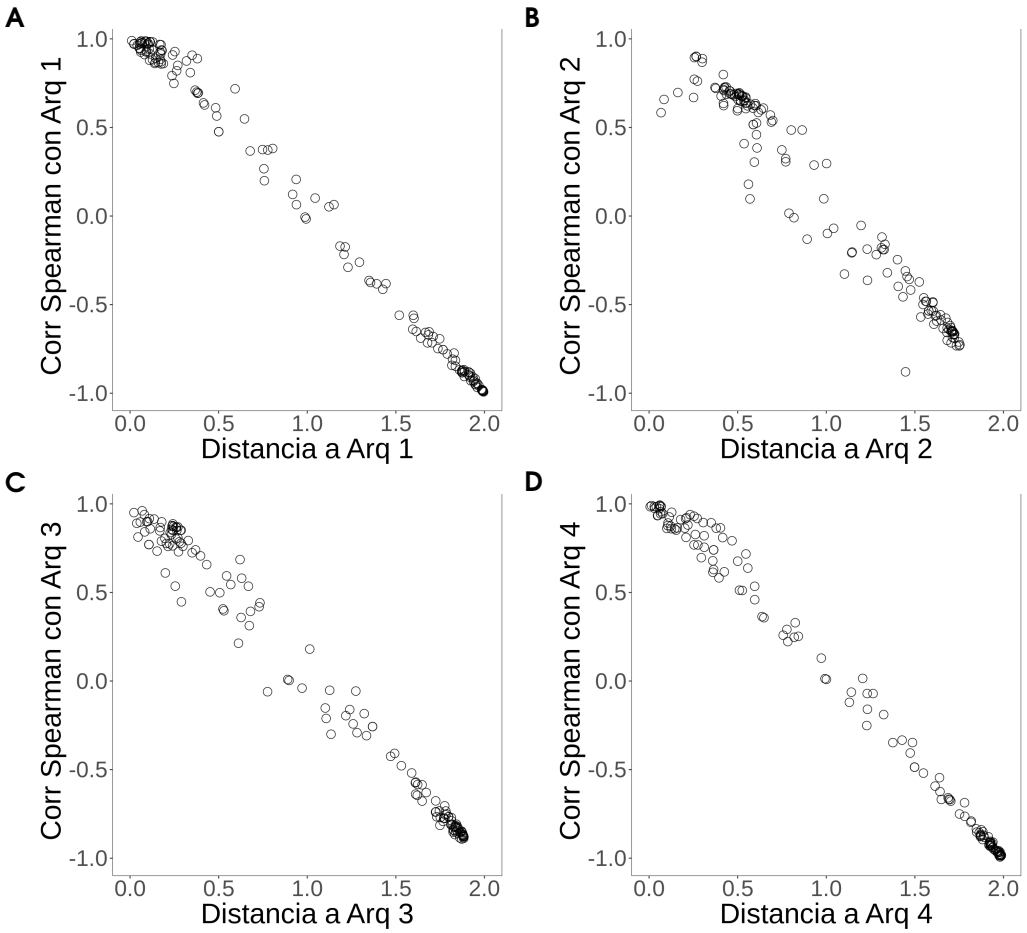

Figura 34: Distancias MDS y Correlación de Spearman. Se determinaron las distancias observadas en las primeras dos componentes del MDS entre cada arquetipo rojo del dataset 'Codones' y todas las variables de enriquecimiento. Se contrastan estas distancias con el índice de correlación de Spearman entre cada uno de los arquetipos planteados y las variables de enriquecimiento. A. Relación entre la correlación de Spearman entre p_E del arquetipo 1 del dataset 'Codones' y el resto de las variables de enriquecimiento en función de la distancia observada entre el arquetipo 1 rojo, correspondiente a p_E, y el resto de las variables observadas en el mapa producto del MDS. B, C y D. Se repite la metodología aplicada, pero para los arquetipos 2, 3 y 4 del dataset 'Codones', respectivamente.

5.6.3.1 Perfiles de Distancias

Cada una de las variables que se localiza en el mapa puede relacionarse con un arquetipo conforme se acerca al mismo. Las distancias de correlación proporcionan una medida de esta relación, aunque puede realizarse un análisis más detallado analizando los perfiles de distancia, es decir, la relación entre cada variable de enriquecimiento y la distancia d_E a cada arquetipo. A efectos de evaluar la validez del mapa de variables, se propone un análisis de los perfiles de algunas variables características (Fig. 35). Si bien el mapa da una idea general del conjunto de

variables, el perfil de cada variable de enriquecimiento en particular representa más fielmente la relación entre cada variable y cada arquetipo.

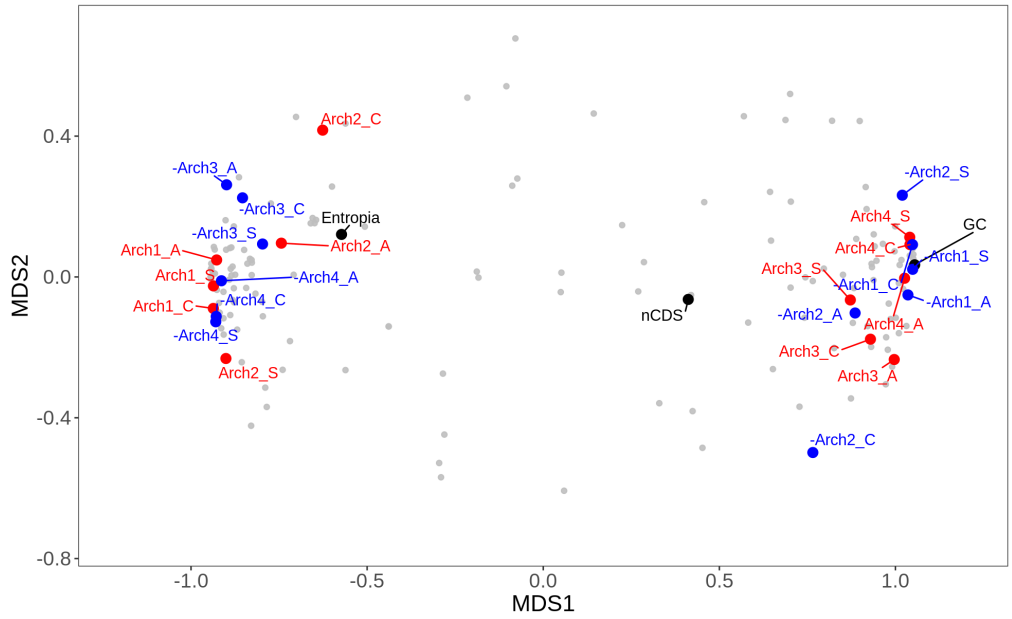

Figura 35: Mapa de variables de enriquecimiento. A partir de las distancias basadas en la correlación de Spearman de todas las variables y las métricas de distancias a los arquetipos (d_E y p_E) se construye un MDS con todas las variables de interés junto con los arquetipos de todos los datasets al mismo tiempo. Se destacan algunas variables para un análisis más detallado de sus perfiles según su ubicación en el mapa. Las métricas p_E toman el color rojo y las d_E, el color azul. Se anota el origen de los datasets de las métricas con 'C', 'A' o 'S' según provengan de los datasets 'Codones', 'Aminoácidos' o 'Sinónimos', respectivamente. En gris figuran el resto de las variables que serán analizadas progresivamente en secciones posteriores.

El contenido GC se conoce como una variable rectora de los usos de codones y aminoácidos. En el mapa de variables se encuentra asociada con el eje principal cercano al arquetipo 4. El perfil analizado corresponde con el dataset 'Codones' y muestra las relaciones con los diferentes arquetipos, especialmente su correlación negativa con la distancia al arquetipo 4. El gradiente de colores permite apreciar más detalladamente su distribución (Fig. 36). La medida de correlación parece representar la relación entre la variable y cada uno de los arquetipos, dado que esta variable se encuentra asociada principalmente con el arquetipo 4.

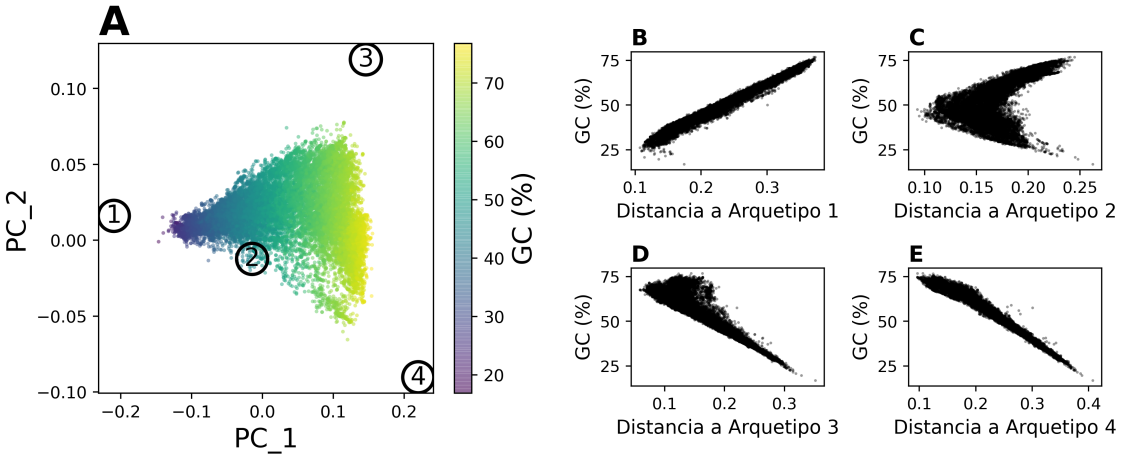

Figura 36: Distribución de contenido GC. A: Se muestra el contenido GC de cada organismo como un gradiente de color en el diagrama de las primeras 2 componentes principales junto con los arquetipos del dataset 'Codones'. B, C, D, E: Para cada arquetipo del dataset 'Codones', se muestra la relación entre la distancia a cada organismo y su contenido GC, pudiendo describir 4 perfiles característicos. Dada la correlación negativa con respecto a la distancia del arquetipo 4, puede suponerse una relación entre el contenido GC y este arquetipo. Correlaciones positivas, como en el caso del arquetipo 1, ubican a la variable GC y al arquetipo 1 en extremos opuesto del mapa.

Otra variable que constituye un buen ejemplo dentro del mapa de variables es la Entropía de Shannon de abundancias de codones (h_C), cercana al arquetipo 2, aunque no tan cercana como la variable GC al arquetipo 4. Nuevamente, se reproduce lo observado con contenido GC en cuanto a la representatividad de la métrica de correlación (Fig. 37). Los perfiles observados en el caso del contenido GC (Fig. 36B y 36E) presentan menor grado de dispersión que los observados en el caso de h_C (Fig. 37C), dando muestra de los perfiles resultantes en los casos en donde una variable y un arquetipo se hallan próximos (arquetipo 2 y h_C) o prácticamente superpuestos (arquetipo 4 y contenido GC). Finalmente, se toma una variable de baja correlación con todos los arquetipos (nCDS), a modo de ejemplificar perfiles de bajo impacto sobre los mismos (Fig. 38). El número de CDS muestra una posición central dentro del mapa, alejada de todos los arquetipos y de la mayoría de las variables. Los perfiles muestran su distribución independiente de las distancias de los arquetipos dando cuenta de la independencia de los arquetipos con esta variable. Tomar una transformación logarítmica del número de CDS no modifica la relación entre esta variable y las distancias a los arquetipos (Fig. 38).

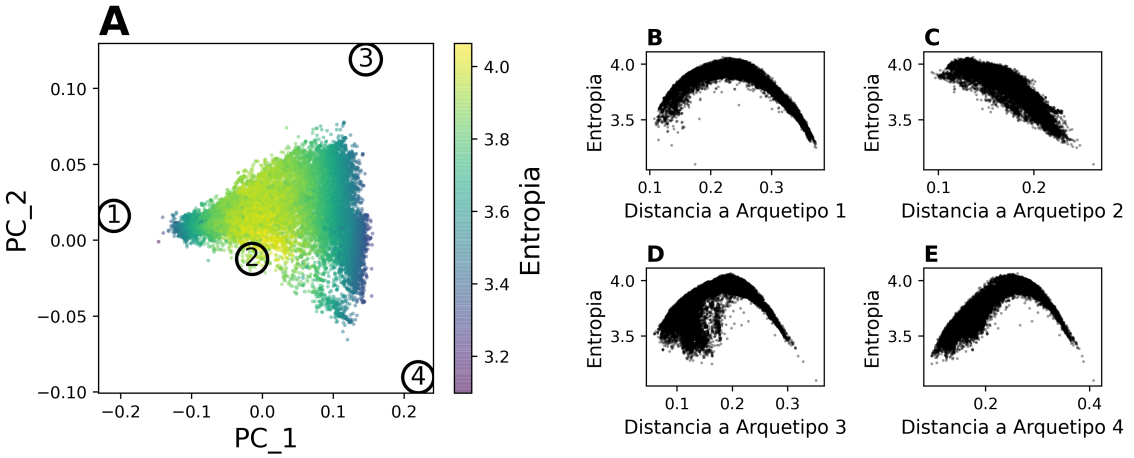

Figura 37: Distribución de entropía. A: Se muestran los valores de 'h_C' (Entropía de Shannon para las abundancias de codones) para cada organismo en las primeras 2 componentes principales del dataset 'Codones'. B, C, D, E: perfiles de 'h_C' en función de distancias a cada arquetipo. La correlación negativa con la distancia aparece en el caso del arquetipo 2. Los perfiles de distancia corroboran lo observado en el mapa de MDS.

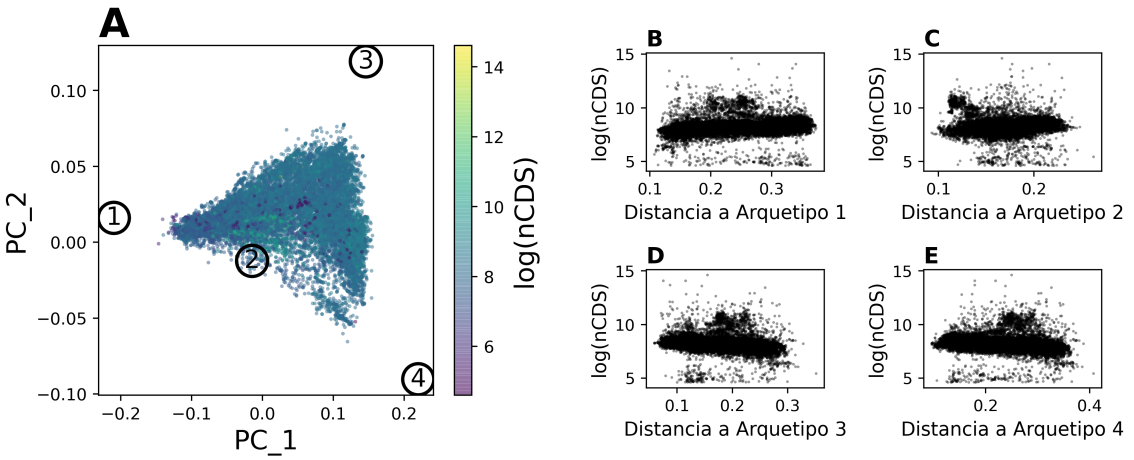

Figura 38: Distribución de cantidad de CDS. A: Se muestran los valores de número de CDS (nCDS) para cada organismo en las primeras 2 componentes principales del dataset 'Codones'. B, C, D, E: perfiles de $log(nCDS)$ en función de distancias a cada arquetipo. Los perfiles de distancia muestran la nula correlación de la variable con la distancia. La transformación logarítmica permite apreciar mejor la distribución de los valores de CDS ya que toman valores muy extremos.

Cada uno de estos casos muestran que el mapa de variables puede asociarse a perfiles entre las distancias a los arquetipos y las variables. Es posible, entonces, observar los mapas en lugar de los perfiles como forma de evaluar las relaciones entre grupos de variables y los arquetipos.

5.6.4 Variables continuas: Mapas de variables

En esta sección, se detallan secuencialmente diferentes grupos de variables continuas. Se emplean mapas de variables descriptos anteriormente, haciendo

especial foco en las relaciones de los arquetipos con las distintas variables de enriquecimiento. Idealmente se intenta encontrar las variables que se encuentren lo más cercanas posibles a cada arquetipo, de forma de asociar 'tareas' o 'propiedades' a cada uno de ellos.

5.6.4.1 Abundancias de codones

- 64 variables representando las abundancias relativas de los 61 codones codificantes y las abundancias de los 3 codones 'stop'.

Las abundancias de codones utilizadas en el dataset 'Codones' se incluyeron como variables de enriquecimiento y se categorizaron para enriquecer el análisis. Cada codón se caracterizó según la cantidad de G o C que posee. Por otro lado, se los caracterizó según si cada codón corresponde con el de máximo o mínimo contenido GC entre sus codones sinónimos. Cada categorización intenta revelar patrones y tendencia de los diferentes grupos de codones (Fig. 39 y Fig. 40).

Desde un punto de vista general, los codones se encuentran distribuidos en todo el mapa, aunque concentrados en los extremos del eje MDS1 (Fig. 39). Se observan los codones con contenidos GC de 0 cercanos al arquetipo 1 y los de contenido GC 3 más cercanos al arquetipo 3 y 4. Esto concuerda con el perfil observado para la variable 'GC'. Los codones de contenido GC de 2 y 3 toman posiciones intermedias y con mucha más dispersión en el centro del mapa, pudiéndose hallar tanto cerca del arquetipo 1 y 2 como de los arquetipos 3 y 4. El arquetipo 1 se puede asociar a valores elevados de abundancias de los codones AUU, UUA, UUU, AAA, AAU, UAU, GUU, UCA, ACU, UCU, ACA, AGA, GAU, AGU, GCU, CCU, AUA, CUA, GCA y CCA. En el caso del arquetipo 3, se observan cercanos los codones GCG, CGC, GUG, CUG, AGC, UCG, GGC y CCG. En el caso del arquetipo 4, los codones que se encuentran en sus inmediaciones son ACC, CAC, GCC, GAC, GUC, CAC, CGG y UCC. En el caso del arquetipo 2, su localización se encuentra algo más dispersa dado que cada dataset lo tiene en una posición diferente. En el caso del arquetipo 2 del dataset 'Codones', el codón GGA se encuentra muy cerca, en el caso del datset 'Aminoácidos', el codón CUU se encuentra muy cerca y en el caso del dataset 'Sinónimos', los codones más próximos son UAA, CAA y CAU. El resto de los codones se encuentran más dispersos en el mapa, no pudiéndose establecer una relación clara con ninguno de los arquetipos. Estos codones son AGG, AAG, UAG,

CGA, AAC, CGU, AUG, UUG, AGC, UAC, GGG, UCC y GAG. Los 3 codones 'stop' se presentan distribuidos a lo largo del eje principal del mapa, hallándose UAA cercano al arquetipo 2 correspondiente al dataset 'Sinónimos', UAG en el centro del mapa y UGA muy cercano al arquetipo 4. Esto muestra que el contenido GC de cada codón por sí mismo no alcanza a explicar del todo la relación entre los codones y los arquetipos.

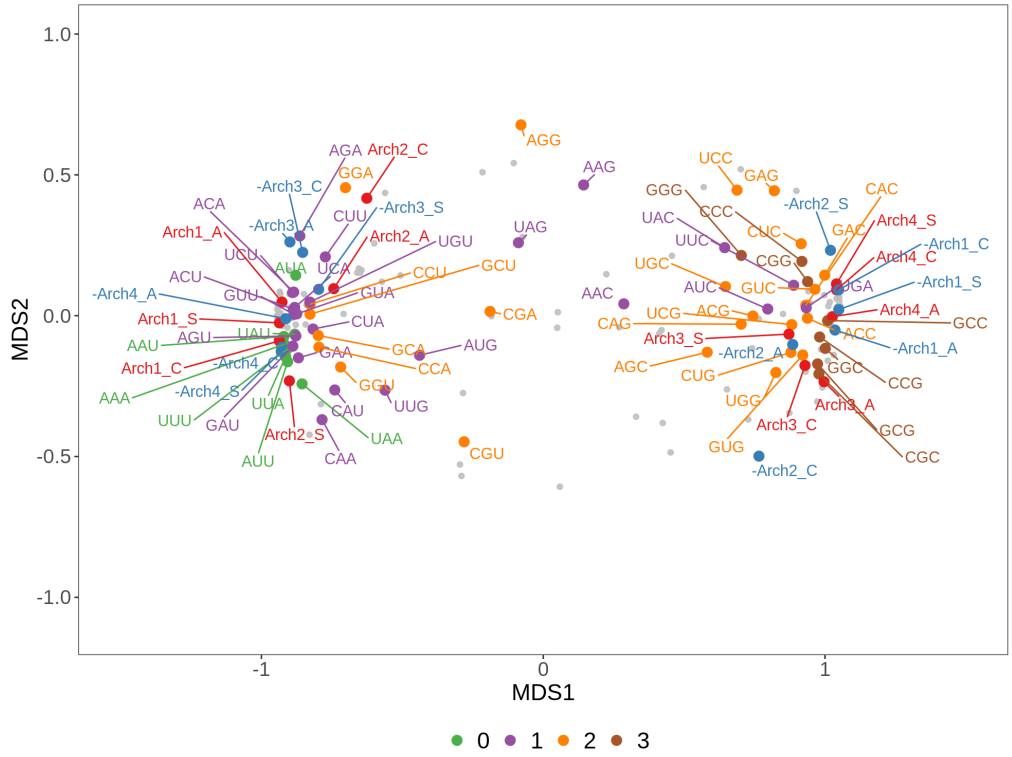

Figura 39: Abundancias de codones y su contenido GC. Se muestran las ubicaciones de cada uno de los 61 codones codificantes y los 3 codones 'stop' en el mapa de variables construido por MDS. Se incluyen los arquetipos de los 3 datasets en rojo y azul según se trate de p_E o d_E respectivamente. Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'.

Por otro lado, se pueden analizar la distribución de los codones según su contenido GC dentro de cada grupo de codones sinónimos (Fig. 40). En esta caso se observan los codones más asociados al eje principal de MDS1, donde los codones de contenido GC máximo entre sinónimos se hallan más cercanos al arquetipo 3 y 4, y los de GC mínimo entre sinónimos más cercanos al arquetipo 1. Sin embargo, los codones de GC mínimo se encuentran muy concentrados en las inmediaciones del arquetipo 1 mientras que los de GC máximo parecen tener una distribución más amplia en el mapa. De hecho, los codones de GC intermedio parecen más cercanos al arquetipo 1

y 2 que a los arquetipos 3 y 4. El contenido GC dentro de codones sinónimos aparece como un rasgo más asociado al eje principal del mapa y determina, al menos en parte, la relación entre los codones y los arquetipos hallados.

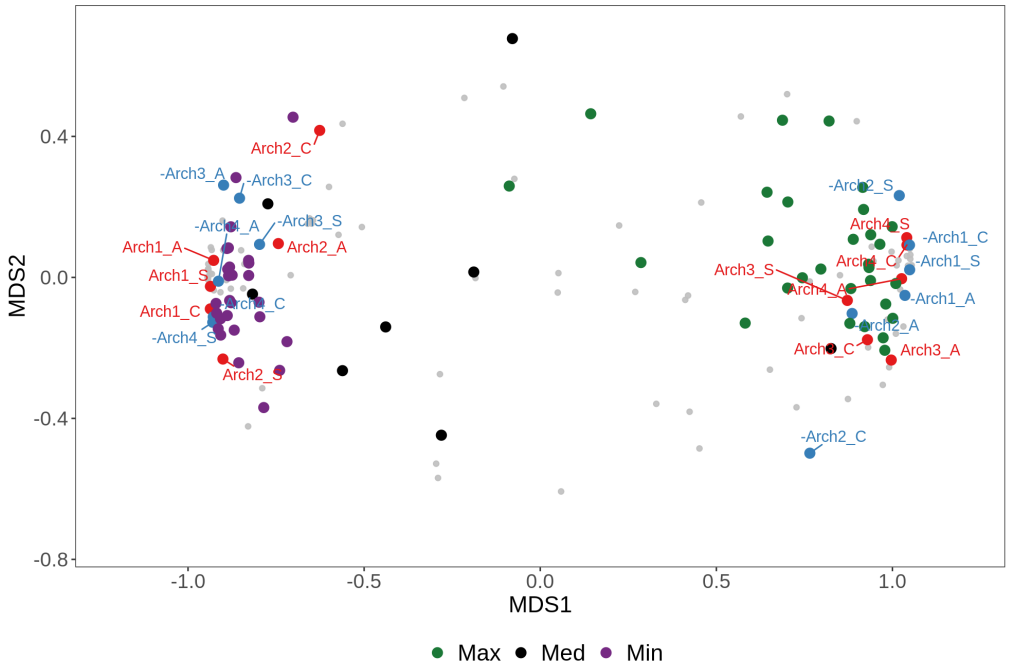

Figura 40: Contenido GC entre codones sinónimos. Se muestran los 61 codones codificantes y los 3 codones 'stop' dentro del mapa de variables construido a partir de MDS. Se destaca cada codón según el uso de contenido GC entre sus codones sinónimos, siendo máximo ('Max') o mínimo ('Min') si contiene el contenido GC máximo o mínimo dentro de su grupo de codones sinónimos. Se anotan como 'Med' el resto de los codones, es decir, los que poseen un contenido GC intermedio o bien existe un solo codón por aminoácido. Se dejan en gris el resto de las variables. Para cada arquetipo, se calculan las métricas p_E (color rojo) y d_E (color azul). Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'.

5.6.4.2 Abundancias de nucleótidos

- Abundancia de Adenina
- Abundancia de Timina
- Abundancia de Guanina
- Abundancia de Citosina
- Abundancia de Adenina + Guanina (Purinas)
- Abundancia de Guanina + Timina (GT-Keto)

- Contenido GC

Al igual que con los codones, se incluyeron las abundancias de nucleótidos o conjuntos de ellos (Fig. 41). En primer lugar, se destaca al contenido GC asociado al eje horizontal, ya analizado en secciones anteriores, con valores elevados asociados principalmente al arquetipo 4 y con valores reducidos asociados al arquetipo 1. Las bases (Adenina, Guanina, Timina y Citosina) determinadas por separado no parecen alterar este patrón. Valores elevados de Guanina o Citosina se asocian al arquetipo 4 mientras que valores elevados de Adenina o Timina se asocian al arquetipo 1 (Fig. 41). En cuanto a las variables surgidas de considerar las sumatorias de las abundancias de bases, la variable de Adenina + Guanina aparece cercana al arquetipo 1 y al arquetipo 2 del dataset 'Aminoácidos' por lo que valores elevados de esta variable podría asociarse a estos arquetipos. En el caso de la variable que suma Guanina + Timina (GT-Keto), se ubica en una posición central lejos de los grandes agregados de variables y lejos de cualquier arquetipo (Fig. 41).

Los valores de contenido GC de los organismos, si bien íntimamente relacionadas con las abundancias de codones y aminoácidos, permiten ilustrar y corroborar los resultados hasta aquí obtenidos (Fig. 36). El gradiente de contenido GC mantiene 2 arquetipos en el extremo de máximo contenido GC y un arquetipo en el extremo de menor valor. El cuarto arquetipo se localiza en un área de GC cercana a 0,5 y plantea una dimensión más a la reconocida influencia del contenido GC de cada organismo en su distribución de codones. El contenido GC evidencia una clara correlación con los arquetipos 1, 3 y 4 (Fig. 36). Adicionalmente, se destaca la variable 'AG-purinas' cercana al arquetipo 1, por lo que valores elevados de esta variable pueden asociarse a este arquetipo (Fig. 41). Esta variable podría asociarse a la contribución que las purinas hacen a la estabilidad del ADN ante la falta de G o C [21].

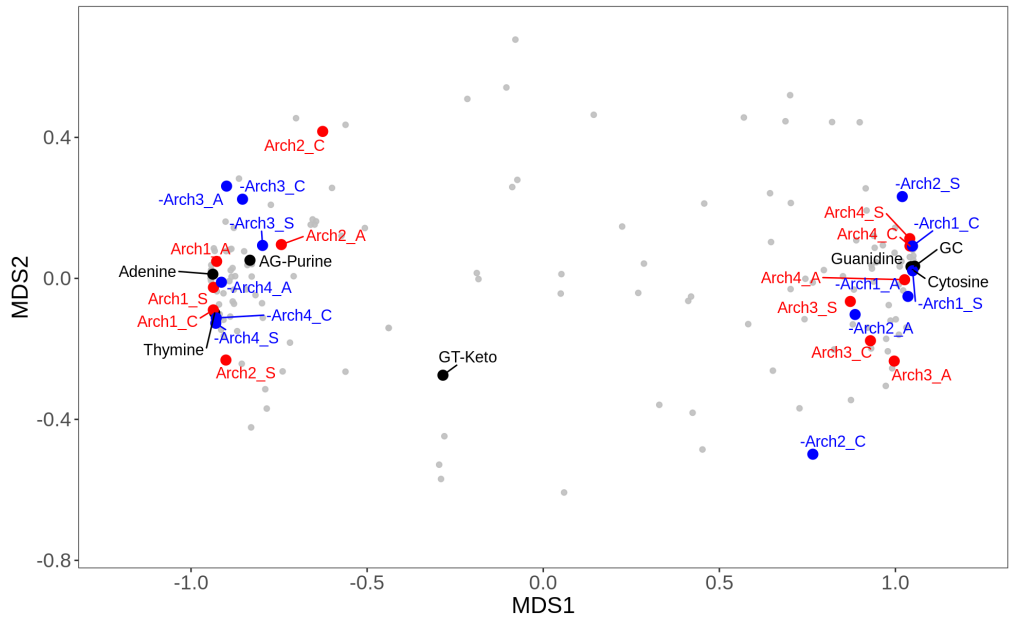

Figura 41: Abundancias de nucleótidos. Se muestran las variables de las proporciones de cada nucleótido o conjunto de ellos (color negro). También se muestran el resto de las variables (gris). Para cada arquetipo, se calculan las métricas p_E (color rojo) y d_E (color azul). Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'.

5.6.4.3 Abundancias de aminoácidos

- Abundancias de cada uno de los 20 aminoácidos

De forma similar a los realizado con los codones y conjuntos de nucleótidos, se incluyeron las abundancias de los aminoácidos como variables de enriquecimiento (Fig. 42). En cuanto a los aminoácidos individuales, se observan abundancias elevadas de aminoácidos asociadas con el arquetipos 1 (Ile, Lys, Phe, Asn, Tyr), con el arquetipo 2 (Ser, Glu), con el arquetipo 3 (Trp, Val, Ala) y con el arquetipo 4 (Arg, Pro, Gly). Siete aminoácidos se observan en posiciones alejadas de los arquetipos, dando cuenta de su independencia (Met, Cys, Thr, Asp, His, Leu, Gln). Estos grupos correlacionan con lo observado en el correlograma de abundancias de aminoácidos (Fig. 18), donde His y Leu formaban un grupo separado del resto y aparecían dos grupos marcados de aminoácidos con una fuerte correlación entre si: Ile, Lys, Phe, Asn y Tyr, por un lado y Trp, Val, Ala, Arg, Pro y Gly por el otro.

Al evaluar la cantidad de codones que codifican a cada uno de los aminoácidos, se observa un patrón en donde aminoácidos con mayor número de codones codificantes se orientan hacia los arquetipos 3 y 4. Algunos aminoácidos constituyen excepciones como Ser o Trp que se localizan en las inmediaciones de los arquetipos 2 y 3

respectivamente. La cantidad de codones codificantes, entonces, no alcanza para explicar por si misma la distribución de los aminoácidos en el mapa.

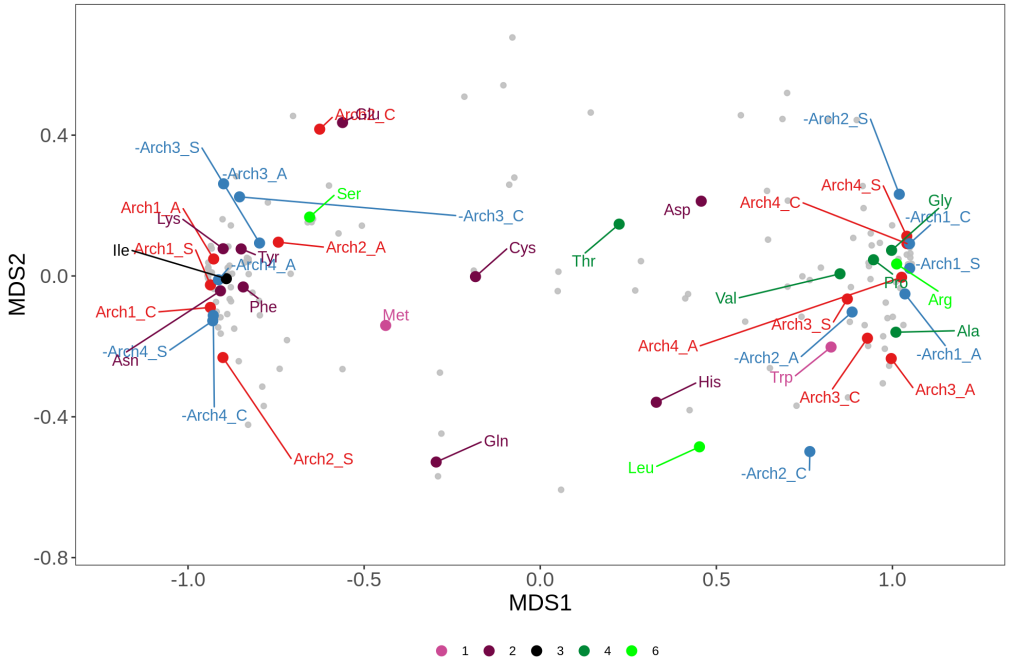

Figura 42: Abundancias de aminoácidos. Se muestran las variables de abundancias de cada aminoácido coloreado según la cantidad de codones codificantes que tiene cada aminoácido en el código genético estándar (1, 2, 3, 4 o 6). También se muestran el resto de las variables (gris) a modo de referencia. Para cada arquetipo, se calculan las métricas p_E (color rojo) y d_E (color azul). Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'. Las distribuciones de los aminoácidos se relacionan con lo observado en el correlograma (Fig. 18)

5.6.4.4 Propiedades originales

- Cantidad de CDS (nCDS)
- Cantidad de codones (nCodons)

Estas dos variables surgen de las bases de datos originales y dan cuenta de las cantidades y de los tamaños de las secuencias utilizadas paras las determinaciones de abundancias. Valores elevados muestran organismos con genomas grandes y ampliamente investigados, donde múltiples secuencias fueron identificadas. Indirectamente dan cuenta de la representatividad de las abundancias calculadas para cada especie. En el caso de 'nCDS', esta variable ya fue analizada en secciones anteriores con mayor detalle (Fig. 38). Ambas variables se encuentran fuertemente correlacionadas entre si y toman una posición en el centro del mapa por lo que no pueden asociarse con ningún arquetipo (Fig. 43).

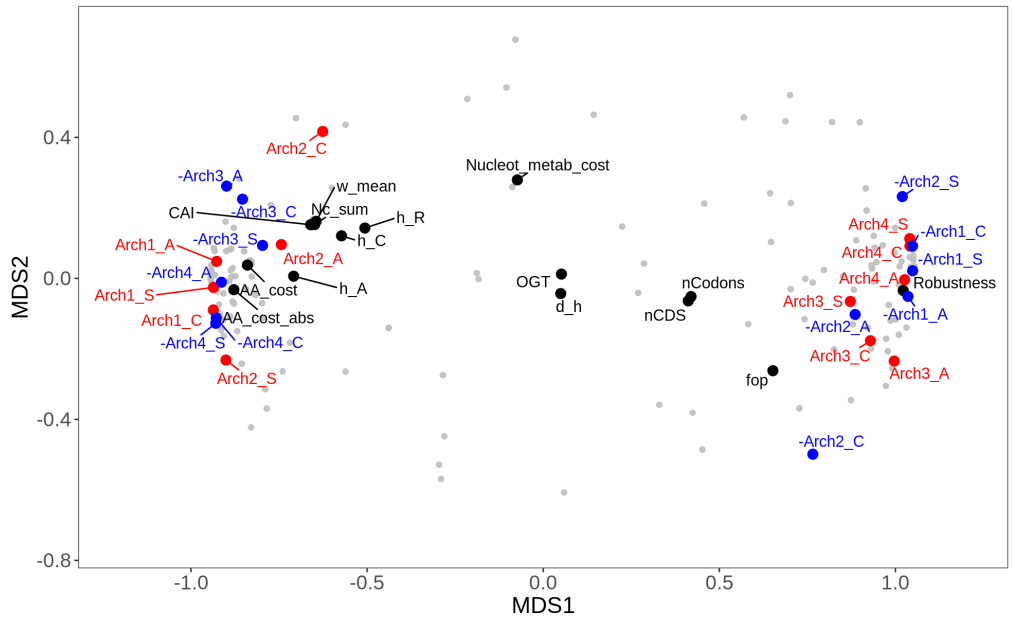

Figura 43: Variables originales, de distribución, fisiológicas y de crecimiento bacteriano. Se detallan variables de distribución de codones, variables fisiológicas, variables de crecimiento bacteriano y algunas variables originales. Para cada arquetipo, se calculan las métricas p_E (color rojo) y d_E (color azul). Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'. El resto de las variables de enriquecimiento se dejan en color gris.

5.6.4.5 Propiedades geométricas de dinucleótidos

- Rise
- Roll
- Shift
- Slide
- Tilt
- Twist
- Major Groove Depth
- Major Groove Width
- Minor Groove Depth
- Minor Groove Width
- Persistance length

De las diferentes variables asociadas a dinucleótidos, se seleccionaron las propiedades de conformación en el espacio que las bases toman entre ellas. Cada una de las propiedades surgen de diferentes fuentes bibliográficas, aunque en algunos casos son determinaciones del mismo parámetro. Para facilitar el análisis, se agruparon las características según los parámetros medidos, de modo que diferentes valores del mismo parámetro que difieren en la fuente bibliográfica tengan el mismo nombre (Fig. 44). Estas variables fueron tomadas tanto de ADN desnudo como de ADN asociado a proteínas, agregando una dimensión más a nuestro análisis. Además, se incluyeron algunas variables asociadas a la estrucutra general del ADN, como la profundidad y ancho de los surcos mayores ('Major Groove') y surcos menores ('Minor Groove'), o la rigidez de las hebras ('Persistance length').

Las 4 variables 'Twist' (en marrón en la Fig. 44) y las 4 variables 'Shift' (en naranja en la Fig. 44), tomadas de múltiples fuentes, se muestran con una amplia distribución. En el caso de las 4 variables 'Roll' (en violeta en la Fig. 44), las variables determinadas con unión a proteínas se encuentran cercanas al arquetipo 4 mientras que las medidas sobre ADN libre se encuentran más dispersas. Las 4 variables 'Tilt' (en verde en la Fig. 44) tiene una comportamiento similar. Las 2 variables 'Tilt' determinadas sobre ADN unido a proteínas se concentran cercanas al arquetipo 3, mientras que las variables determinadas sobre ADN libre toman posiciones más dispersas. En el caso de las 4 variables 'Rise' (en rosa en la Fig. 44), se observan cercanas al arquetipo 4, por lo que valores elevados de esta variable puede asociarse con este arquetipo. Finalmente, las 3 variables 'Slide' (en marrón claro en la Fig. 44) se hallan en las inmediaciones del arquetipo 3, y se observa que las variables de determinaciones realizadas sobre ADN asociado a proteínas se agrupan entre si y toman posiciones más alejadas.

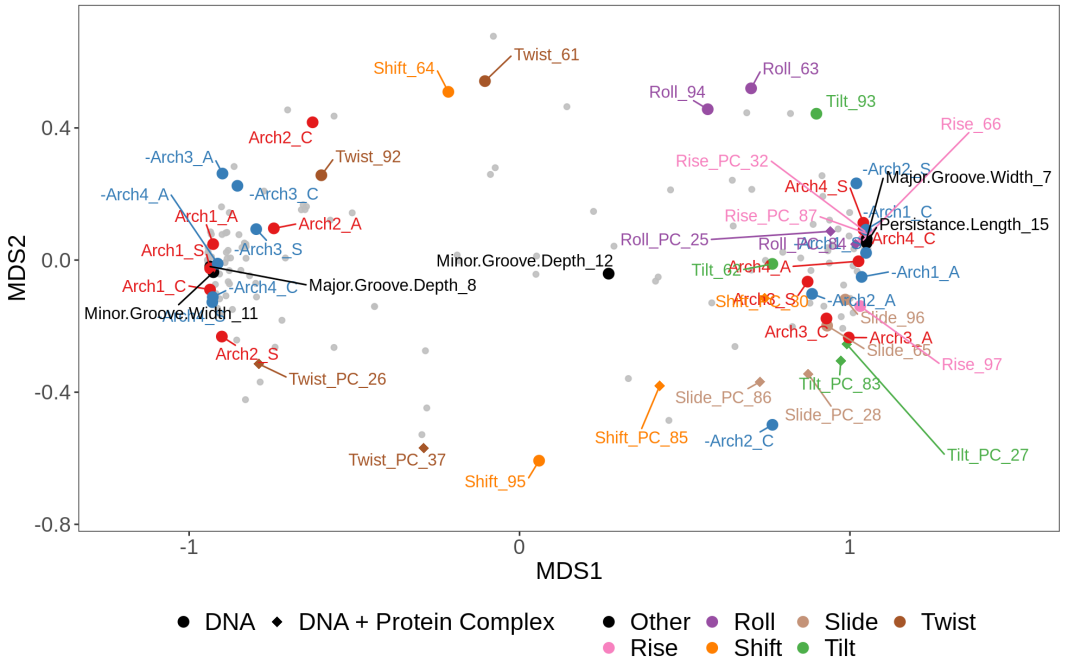

Figura 44: Propiedades geométricas. Se muestran las variables geométricas asociadas con las posiciones de los nucleótidos en las hebras de ADN. Se diferencian entre el ADN libre y el asociado a proteínas. Se muestran en 6 parámetros (Rise, Roll, Slide, Shift, Tilt, Twist) donde se incluyen variables de diferentes fuentes, y representan posiciones o ángulos relativos de las bases en las hebras de ácidos nucleicos. Se adiciona la categoría 'Other' donde se incluyen características estructurales del ADN, dado que también constituye un descriptor de la geometría del ADN. Se anotan los parámetros determinados sobre ADN unido a proteínas como 'PC' ('Protein Complex') Para cada arquetipo, se calculan las métricas p_E (color rojo) y d_E (color azul). Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'. Se deja el resto de las variables de enriquecimiento en gris.

En cuanto a las variables referentes a los surcos del ADN, la profundidad del surco mayor y el ancho del surco menor se encuentran muy cercanas entre si, por lo que se encuentran altamente correlacionadas. Además, valores elevados de estas dos variables pueden asociarse con el arquetipo 1 dada su proximidad en el mapa. Con respecto al ancho del surco mayor y a la rigidez de las hebras ('Persistance length'), esas dos variables se encuentran correlacionadas y próximas al arquetipo 4, por lo que valores elevados de estas variables pueden asociarse a este arquetipo. Finalmente, la variable de profundidad de surco menor se halla en una zona central del mapa sin poder correlacionarse con ninguna de las otras variables analizadas.

Entre las variables geométricas podemos hallar algunas con un comportamiento más consistente, como los casos de 'Rise' y 'Slide', y otras con distribuciones más amplias como 'Shift' o 'Twist'. Se logran identificar las variables con claras asociaciones a algunos arquetipos, y se facilita el análisis de las relaciones que existen

entre las variables de este grupo.

5.6.4.6 Propiedades asociadas a energías de dinucleótidos

- Entropía (Entropy)
- Entalpía (Enthalpy)
- Energía Libre (Free Energy)
- Movilidad hacia surco mayor (Mobility Major Groove)
- Movilidad hacia surco menor (Mobility Minor Groove)
- 'Melting Temperature'
- Energía de apilamiento (Stacking Energy)

Al igual que en el caso de las variables geométricas, se agruparon varias variables relacionadas con las energías involucradas en el acoplamiento de nucleótidos y hebras de ADN (Fig. 45). La variable entropía, entalpía y energía libre hacen referencia al proceso de formación del duplex de ADN. La entropía definida en este caso refiere a la magnitud del proceso termodinámico de formación del duplex de ADN y no a una entropía de Shannon como se analiza en secciones posteriores. Las variables que cuantifican movilidad refieren a la contribución de los nucleótidos o conjuntos de ellos a facilitar que se deforme la estructura del ADN. La energía de apilamiento da cuenta de la energía involucrada en el solapamiento de las bases de los nucleótidos de una misma hebra de ADN. Finalmente, 'melting temperature' da cuenta de la temperatura a la cual la mitad de una determinada cantidad de ADN se desnaturaliza (Fig. 45).

Desde un punto de vista general, las propiedades energéticas asociadas a nucleótidos se muestran muy polarizadas en el mapa de variables. Una de las variables de movilidad y las variables de 'melting temperature' se acercan al arquetipo 4, pudiendo asociarse valores elevados de estas variables con este arquetipo. Las 2 variables correspondientes a la movilidad ('Mobility') se encuentran en lugares opuestos del mapa, la correspondiente al surco mayor cercana al arquetipo 1 y la correspondiente al surco menor cercana al arquetipo 4.

El resto de este grupo de variables se agrupan fuertemente en las inmediaciones del arquetipo 1, por lo que valores altos de todas estas variables pueden asociarse

con este arquetipo (Fig. 45). La energía libre ('Free Energy') muestra una elevada consistencia asociada al arquetipo 1, dando cuenta de la concordancia en la literatura de este parámetro, ya que valores elevados de energía libre implican una menor estabilidad de la doble hélice de ADN, frecuentemente asociada a la baja proporción de GC en la secuencia nucleotídica. La energía libre también se relaciona con la energía de apilamiento ('Stacking Energy'), por lo que también es esperable que se hallen fuertemente correlacionadas y cercanas en el mapa de variables. Algo semejante ocurre con la entalpía, que forma parte de la energía libre y se correlaciona fuertemente con ella, localizándose en la misma región del mapa (Fig. 45).

A diferencia de las variables geométricas, en este grupo de variables de enriquecimiento, las variables correspondientes al mismo parámetro pero tomadas de diferente fuentes, parecen guardar coherencia en cuanto a correlación. Además, la mayoría de ellas parecen responder a principios parecidos, por lo que se hallan fuertemente correlacionadas entre sí.

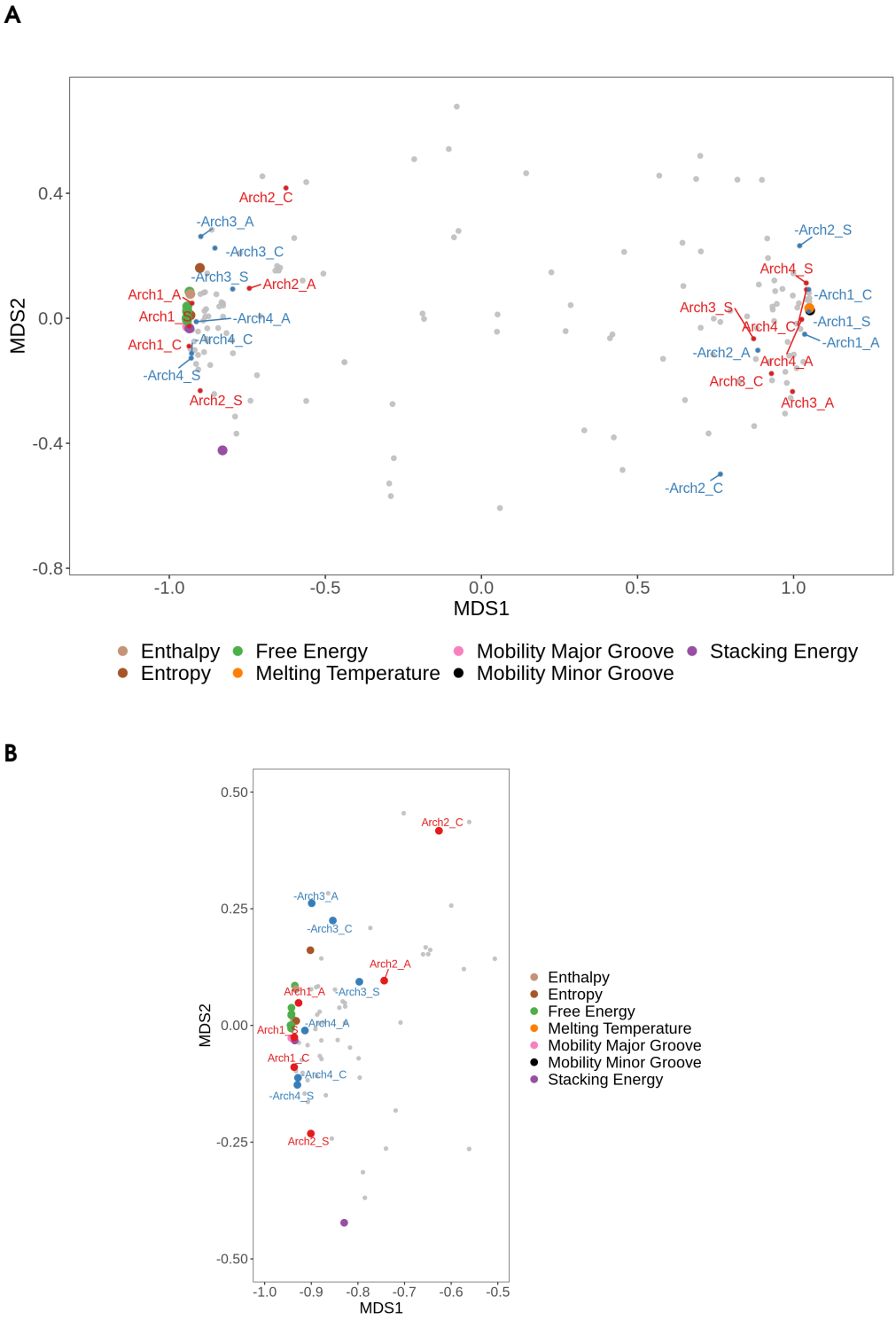

Figura 45: Variables de energías. Se muestran los conjuntos de variables asociadas a diferentes energías o a parámetros íntimamente relacionados con las energías de interacción (ej. 'Melting Temperature') entre las bases de los nucleótidos. A: se muestra el mapa completo junto con la totalidad de los arquetipos. B: detalle de área cercana al arquetipo 1, donde aparecen la mayoría de las variables de energía analizadas. Se incluyen la totalidad de las variables de enriquecimiento (gris) y las métricas de distancias a los arquetipos p_E (rojo) o d_E (azul). Se anota cada arquetipo según su dataset de origen, siendo los arquetipos 'C' localizados en el dataset 'Codones', 'A' los localizados en el dataset 'Aminoácidos', y 'S' los localizados en el dataset 'Sinónimos'.

5.6.4.7 Propiedades de distribución

- Entropía de abundancias de codones (h_C)
- Entropía de abundancias de aminoácidos (h_A)
- Entropía de sinónimos (h_R)
- Adaptabilidad relativa promedio (w_mean)
- Índice de adaptabilidad de codones (CAI)
- Frecuencia de codones óptimos (fop)
- Suma de número efectivo de codones (Nc_sum)

Las propiedades de distribución muestran distintas medidas del sesgo presente en las abundancias de los datasets primarios. En todos los casos se construyen variables que ayudan a interpretar el conjunto de abundancias ya sea de codones o de aminoácidos. Estas variables se incluyen en el mapa de variables, construido mediante MDS, de forma de poder evaluar su potencial relación con los arquetipos hallados, o con otras variables de interés (Fig. 43). Valores elevados de las variables h_C, h_A, h_R, w_mean, CAI y Nc_sum implican una distribución de abundancias más homogénea (menos sesgada), mientras que un valor elevado en fop implica una distribución más asimétrica (fuertemente sesgada).

Las variables de entropía de abundancias (h_C y h_A) dan cuenta de las distribuciones generales de abundancias de los datasets de trabajo 'Codones' y 'Aminoácidos' respectivamente. Valores elevados de entropía indican mayor homogeneidad en la distribución de abundancias. La entropía de uso de codones se muestra en las inmediaciones del arquetipo 2, dando cuenta que un valor elevado de entropía puede asociarse con este arquetipo (Fig. 43). En el caso de la entropía de uso de aminoácidos (h_A), su posición es similar, hallándose muy cerca del arquetipo 2 del dataset 'Aminoácidos', indicando que este arquetipo puede asociarse a valores altos de entropía en las abundancias de aminoácidos. En el caso de la entropía de sinónimos (h_R) su localización es similar a las otras dos, dando cuenta de su correlación y comportamiento similar (Fig. 43).

En cuanto a las variables 'CAI', 'w_mean' y 'Nc_sum', se localizan en el mismo punto, por lo que su comportamiento es similar. Se hallan en las inmediaciones

del arquetipo 2 por lo que valores altos de estas variables pueden asociarse a este arquetipo. La variable 'fop' se halla en el lugar opuesto del mapa. De hecho, se halla cerca del arquetipo 3 pero también del arquetipo 2 azul. Esto implica que valores elevados de esta variable podrían asociarse al arquetipo 3, o bien valores bajos de esta variable podrían asociarse al arquetipo 2. Esto muestra, además, una relación inversa con las variables 'CAI', 'w_mean' y 'Nc_sum' mencionadas previamente (Fig. 43).

Las variables de distribución se muestran con una clara relación con el arquetipo 2. Cada una de las variables, si bien próximas entre si, ocupan lugares propios en el mapa. El arquetipo 2 aparece, entonces, con características asociadas a un bajo sesgo en las abundancias de cualquiera de los tres datasets primarios.

5.6.4.8 Propiedades fisiológicas

- Costo metabólico de nucleótidos (Nucleot_metab_cost)
- Costo metabólico de aminoácidos absoluto (AA_cost_abs)
- Costo metabólico de aminoácidos (AA_cost)
- Robustez del uso de codones (Robustness)

Estas propiedades se relacionan con procesos del metabolismo celular. Los metabolismos de aminoácidos o nucleótidos involucran costos energéticos asociados a su síntesis. Estos costos se reflejan en las variables 'AA_cost_abs' y 'Nucleot_metab_cost'. La estabilidad de las moléculas también afecta el costo global, ya que moléculas más estables requerirán menor tasa de síntesis debido a que pueden reutilizarse más veces, por lo que su costo global será menor. Este factor de estabilidad se incluye en la variable 'AA_cost'. Finalmente, el uso de codones más robustos implica que las posibles mutaciones no afectarán significativamente a las proteínas resultantes, por lo que se conservará su función y eficiencia. Cambios de aminoácidos por otros muy diferentes o hacia un codón 'stop' afectará severamente la función proteica, aumentando las exigencias a la funcionalidad celular. Este factor de robustez de los diferentes codones se incluye en la variable 'Robustness'.

En el caso del costo metabólico de nucleótidos, esta variable se encuentra en una región céntrica del mapa (Fig. 43), no pudiendo asociarse a ningún arquetipo en particular. Sin embargo, en el caso del los costos asociados a aminoácidos, las dos

métricas elegidas se encuentran cercanas entre si, orientadas hacia los arquetipos 1 y 2. Particularmente, la variable de costo de aminoácido 'AA_cost' se encuentra cercana al arquetipo 2 del dataset de 'Aminoácidos', por lo que un valor alto de estos costos se podrían asociar con este arquetipo en particular (Fig. 43). La variable 'AA_cost' resulta más relevante que 'AA_cost_abs', ya que involucra un término adicional relacionado con la labilidad de los aminoácidos. De hecho, demuestra ser más efectiva a la hora de explicar las distribuciones naturales de abundancias de aminoácidos [20].

En el caso de la variable de robustez, esta se halla cercana al arquetipo 4, especialmente al arquetipo del dataset 'Aminoácidos'. Esto implica que valores elevados de esta variable, es decir, el uso de codones más 'robustos', podría asociarse con este arquetipo (Fig. 43).

5.6.4.9 Propiedades de crecimiento bacteriano

- Temperatura óptima de crecimiento (OGT)
- Tiempo de duplicación bacteriano (d_h)

Por último, estas dos variables representan determinaciones relacionadas con el crecimiento bacteriano y se relacionan indirectamente con la administración de la energía en las células. Se hallaron los valores de ambas variables para 132 organismos del dominio Bacteria y Archaea. En el caso de 'OGT', las temperaturas afectan a la estabilidad molecular y a la funcionalidad de las macromoléculas. El tiempo de duplicación bacteriano hace referencia a la velocidad a la cual las células pueden multiplicarse [67]. Ambas variables podrían dar muestras de la presión selectiva que pudiese aplicarse sobre las secuencias, sobre todo en cuanto a las preferencias de codones sinónimos o aminoácidos de funcionalidad similar. Si bien ambas variables correlacionan fuertemente entre si, su posición en el centro del mapa impide poder relacionarlas con algún arquetipo en particular (Fig. 43).

6 Discusión

6.1 Procesamiento de variables

Como se analizó en la sección 5.1, la base de datos CoCoPUTs dispone, desde un inicio, de un número de organismos que representan fundamentalmente a los dominios de Bacteria y Eukaryota (Fig. 13, 14 y 15). El dominio de Archaea, si bien menos representado, pareciera alcanzar para poder evaluar las diferencias en sus distribuciones y características. Luego de la aplicación de los diferentes filtros, se obtiene un conjunto de registros aceptables para los análisis posteriores. La base de datos CoCoPUTs posee registros provenientes tanto de Genbank como de Refseq en diferentes cantidades, a pesar de ser Genbank una base con datos no necesariamente verificados. Las secuencias obtenidas provenientes del dominio Eukaryota provienen casi exclusivamente de registros provenientes de Genbank, mientras que en el caso de Bacteria, tanto Refseq como Genbank tienen gran cantidad de información. Esto vuelve a Genbank una base de datos esencial para este estudio ya que brinda un gran aumento en la variedad de organismos de Eukaryota, donde Refseq solo cuenta con 996 organismos. Así las dos bases de datos se complementan y permiten balancear, al menos en parte, el número de organismos pertenecientes a dominios diferentes. La inclusión de bases de datos adicionales, o de mayor número y variedad de registros, puede ayudar a expandir el análisis ya realizado.

Las distribuciones, con respecto al contenido GC, muestran una leve tendencia de organismo de mayor número de CDS hacia valores de mayor contenido GC (Fig. 15). Este sesgo aparece marcadamente en el dominio Bacteria, aunque parece reproducirse en el dominio Archaea. Si bien este hecho está extensamente estudiado en genomas procariotas [79], en este trabajo se destaca la cantidad y variedad de los organismos involucrados.

Los datos tomados originalmente podían producir algún sesgo en el análisis, por lo que aplicar filtros sobre los datos originales fue necesario para alcanzar datos de mejor calidad y confianza (Fig. 11 y 16). Se dejaron afuera un gran número de registros, con el fin de garantizar una muestra homogénea de datos capaces de representar los fenómenos a ser estudiados. Los filtros se aplicaron de manera secuencial de forma de poder evaluar el impacto sobre el conjunto de organismos. Se dejaron de lado registros provenientes de genomas virales o con códigos genéticos

distintos del estándar. También se excluyeron registros provenientes de organelas o plásmidos. Todos estos grupos constituyen objetos de estudio en sí mismos, pudiéndose plantear nuevos interrogantes relacionados con el enfoque planteado en este trabajo. Se intentaron establecer criterios que reduzcan lo menos posible el número de registros, y que permitan utilizar organismos con cantidades de CDS y codones con cierta representatividad. Los registros originales de las bases de datos mostraban organismos con cantidades de CDS muy bajas, pudiendo ocasionar un sesgo en el análisis, por lo que se intentó remover estos organismos poco representados, para mantener cantidades de CDS representativas que den cuenta de las abundancias de codones y aminoácidos esperables para organismos naturales. Así, los datos sin filtrar poseen organismos con cantidades menores de 20 CDS, mientras que la aplicación de cotas inferiores deja únicamente organismos con cantidades de CDS superiores a 100. Los filtros planteados se aplicaron de forma de mejorar la representatividad de los datos, sin embargo, es posible expandir el análisis incluyendo parte de los datos excluídos de los datasets de trabajo, así como analizar subconjuntos puntuales basados en algún criterio puntual (alto o bajo contenido GC, secuencias provenientes de organelas, organismos con códigos genéticos por fuera del estándar, por ejemplo).

Como se menciona en la sección 4.3.2, la construcción de las variables de enriquecimiento logró integrar diferentes fuentes de datos, y generar una lista de características relevantes para el análisis (Tabla 2). Estas características forman un objeto de estudio en si mismas, ya que la relación de cada una con los arquetipos forma parte de su comportamiento y de su relación con el resto de las variables. Tomando variables de diferentes orígenes, se intenta explorar diferentes aspectos de los procesos celulares que pudieran relacionarse con los usos de codones y aminoácidos. Las diferentes variables utilizadas abarcan un gran espectro de significados diferentes (metabolismo celular, taxonomía, distribuciones de abundancias, sesgos de uso de codones), con escalas muy variadas, que pueden asociarse con los arquetipos hallados. Se seleccionaron un grupo de variables de interés, pero el análisis puede expandirse hacia un número mucho mayor de características. La bibliografía dispone de características de un número limitado de organismos, aminoácidos y codones. Mayor cantidad de datos permitirá construir nuevas variables y poder enriquecer el análisis ya realizado.

6.2 Datasets y metodologías de trabajo

En la sección 5.2 se analizan las características generales de los datasets de trabajo. Intentan representar aspectos diferentes, pero complementarios de los sesgos en los usos de codones y aminoácidos, teniendo en cuenta la relación establecida por el código genético (Fig. 4). Además, los datasets por separado mantienen un significado propio. En cada dataset las variables involucradas son interdependientes entre si, y las diferentes observaciones ayudan a construir una zona de organismos observables que dé cuenta de las interacciones existentes. Son estas interacciones las que constituyen el objeto de estudio principal de este trabajo.

Los datasets muestran en su conjunto un eje principal asociado al contenido GC (Fig. 17). Esta variable rige los tres datasets y agrupa sus componentes, al menos en una primera aproximación. Los aminoácidos muestran una clara formación de dos grupos principales, donde el GC de sus codones toma un rol determinante de este ordenamiento (Fig. 18). Aparecen, sin embargo, varios aminoácidos cuyo comportamiento no puede explicarse únicamente por esta variable.

Los datasets de trabajo obtenidos resultan de pequeña escala en comparación con aplicaciones más intensivas donde se construyen modelos mediante aprendizaje no supervisado sobre grandes cantidades de información. Si bien es posible encontrar ejemplos de aplicaciones de estrategias similares en conjuntos de datos más pequeños [41, 42, 43], los métodos actuales permiten construir modelos muchos más complejos con volúmenes de datos mucho mayores [49]. Los volúmenes de datos provenientes de fuentes biológicas pueden variar su tamaño según su naturaleza, pero en este caso se alcanza un dataset que, si bien no es particularmente extenso, abarca una gran diversidad de organismos en cuanto a su clasificación taxonómica. Ahora bien, la cantidad de datos genómicos disponibles se encuentra en un crecimiento exponencial [35], por lo que se vuelve esencial poder repetir este mismo análisis con un número mayor de datos. La metodología empleada está preparada para conjuntos de datos muchos mayores, por lo que es escalable a datasets más completos y con un mayor número de variables.

Este trabajo involucra metodologías de aprendizaje no supervisado aplicado a datasets conteniendo información biológica. Además, incluye la aplicación de estrategias de reducción de dimensionalidad y de construcción de variables. Se incluyen métodos de muestreo, manejo de bases de datos y análisis de distribuciones.

Finalmente, se aplican las diferentes etapas propias de los procesos de minería de datos alcanzando a llegar a una mejor interpretación del conjunto de datos originales.

En cada una de las etapas de este trabajo se emplearon diferentes herramientas de software y se implementaron las diferentes metodologías, tomando como base y en caso de que estuviese disponible, lo publicado en la bibliografía. Las etapas de exploración de datos y filtrado involucraron la confección de un pipeline que permita tomar los datos iniciales desde las bases de datos de CoCoPUTs y posibilite verificar y cruzar sus referencias con las bases de datos de taxonomía de NCBI. Las bases de datos secundarias son muy abundantes en las disciplinas biológicas y facilitan la disponibilidad de la información. En este trabajo se utilizaron bases de datos secundarias de abundancias como CoCoPUTs, de dinucleótidos, y de registros taxonómicos. Además, se tuvieron que incorporar datos de fuentes adicionales para la confección de las diferentes variables de enriquecimiento. Esta etapa fue realizada utilizando las implementaciones existentes en R y Python, dependiendo de los paquetes existentes que facilitaran tanto la interacción con las distintas bases de datos como la confección de los gráficos. Las implementaciones de las metodologías de búsqueda de arquetipos fueron tomadas de sus respectivos autores y adaptadas, según el caso, a los datasets de trabajo. AAnet presenta su método a partir de un paquete implementado en Python que fue tomado y adaptado. Por su parte, ParTI mantiene su software implementado en MATLAB por lo que, nuevamente, se adaptó para aplicarlo a los datasets de trabajo. Finalmente, las etapas de análisis de variables de enriquecimiento se realizaron utilizando los paquetes ya existentes de Python y R, según el caso. Durante el desarrollo de la tesis, la mayor dificultad radicó en poder asociar diferentes herramientas de distintas fuentes en un pipeline robusto que tuviera en cuenta todas las variables al mismo tiempo. Poder tomar datos de entrada de diferentes fuentes, e incluirlos en un análisis en conjunto, implicó sucesivos pasos de manipulación de datos.

Todo el análisis pudo realizarse en una PC de escritorio (i5-2310 CPU 2.90GHz, RAM 8Gb), si bien algunos de los pasos demoraron horas de cómputo. Los métodos de búsqueda de arquetipos mediante AAnet, y las estimaciones de los intervalos de confianza de las localizaciones de los arquetipos mediante ParTI, fueron las etapas de más requerimiento computacional. En todos los casos, los procesos no demoraron más de un día en completarse. El volumen de los datos de trabajo y la complejidad

de los algoritmos empleados, simplificaron la operatoria durante la realización de este trabajo.

6.3 Modelo de búsqueda y localización de arquetipos

Como se describe en la sección 5.3, se utilizan dos métodos alternativos para la búsqueda de arquetipos tomados de publicaciones recientes. Dentro de las publicaciones de las disciplinas biológicas, el método ParTI aparece como el más aplicado para la búsqueda de arquetipos [41]. Sin embargo, este método fue publicado hace más de 5 años, por lo que se decidió ensayar el método AAnet que constituye la última metodología de búsqueda publicada al momento del inicio de este trabajo [49] (Fig. 12).

La metodología empleada de AAnet mostró capacidad de hallar arquetipos en el dataset testeado, si bien fue diseñada para trabajar con un mayor volumen de datos. A pesar de esto, mostró la capacidad de localizar arquetipos en ubicaciones compatibles con las distribuciones de organismos. Además, el número de arquetipos óptimo según el método resulta comparable con el obtenido con el método ParTI, lo cual resulta esperable dado que los arquetipos se infieren a partir de los mismos datos en ambos métodos. El método AAnet posee mayor requerimiento de cómputo, ya que requiere entrenar un modelo constituido por un gran número de parámetros. Es un método preparado para trabajar con imágenes y grandes conjuntos de datos, si bien es posible adaptar modelos con menor número de parámetros para datasets más chicos. Resultó posible adaptarlo a los datasets de este estudio, si bien resulta difícil hallar una estructura del autoencoder satisfactoria. La estructura del autoencoder impacta fuertemente en el proceso de entrenamiento. En primer lugar, impacta en el requerimiento de cómputo, ya que redes más complejas requieren más tiempo de entrenamiento. Para los datasets trabajados, esto no constituye un problema mayor, ya que se obtienen redes entrenadas tras, como mucho, 1 hora de cálculo. En segundo lugar, afecta a los resultados de la función de pérdida y, consecuentemente, a los gráficos determinantes del número de arquetipos. Esto hace que el número de arquetipos se vea fuertemente afectado por el modelo elegido. Finalmente, la estructura afecta la ubicación de los arquetipos y su reproducibilidad. Si bien el modelo intenta resolver el número de arquetipos y su ubicación al mismo tiempo, se requiere un método que pueda balancear estos aspectos y poder así

optimizar la mejor estructura del autoencoder para un dataset dado. El proceso de optimización de los parámetros permite atacar todos los problemas al mismo tiempo, pero no siempre resulta en una estructura compatible con la distribución de las observaciones. El método de AAnet resultó complejo de poner a punto y poder emplearse en la determinación del número óptimo de arquetipos, ya que se pretende emplear los mismos parámetros para correr diferentes número de arquetipos y lograr, en todos los casos, resultados aceptables. Finalmente, es importante aclarar que la publicación del método AAnet no finalizó su proceso de revisión, por lo que constituye una metodología aún en desarrollo.

La metodología AAnet, si bien más actualizada y con un enfoque más integral, no logró adaptarse totalmente al problema planteado. Si bien es capaz de manejar volúmenes de datos más grandes, para los problemas biológicos donde los datasets son más reducidos, un método más simple puede ser capaz de producir mejores resultados.

El método ParTI constituye el más usado en los últimos años para aplicaciones biológicas, y fue utilizado en datasets de dimensiones comparables a los utilizados en este trabajo. El comportamiento de ParTI es comprable a lo descripto en la bibliografía [49] ya que, tanto las ubicaciones como el número de arquetipos, son esperables para los datasets utilizados (Fig. 19 y 20). Los controles aplicados mostraron el comportamiento del método en conjuntos de datos sin estructura (Fig. 21). Las diferencias entre los perfiles de los datasets control y los perfiles de los datasets primarios validan los resultados obtenidos. El método ParTI produce perfiles de varianzas explicadas marcadamente más elevadas en los datasets primarios con respecto a los datasets control. Se hace evidente la necesidad de evaluar los perfiles en su conjunto y compararlos entre si, para luego poder evaluar el punto de quiebre de los perfiles y determinar el número óptimo de arquetipos. El método cuantitativo propuesto facilita la determinación del número óptimo de arquetipos, a la vez que facilita el análisis de las curvas. Los diferentes algoritmos para la búsqueda de los arquetipos mostraron concordancia, y localizan a los arquetipos en posiciones con baja variabilidad (Fig. 22). Las metodologías de búsqueda mantienen un enfoque común, mientras aplican estrategias similares y logran hallar arquetipos en locaciones comunes y consistentes con lo buscado [44].

ParTI, utiliza un modelo más simple que AAnet por lo que requiere menor tiempo

de cómputo. Además, dado que los datasets utilizados son reducidos, se adapta mejor al problema planteado. Este método resultó robusto en cuanto a los datos de trabajo y permitió ubicar eficientemente a los arquetipos en los datasets. Dado que el método involucra la forma de las observaciones en el conjunto de variables, se vuelve robusto frente a grupos de observaciones muy concentradas. Esto se vuelve importante al trabajar con datasets de organismos ya que, por ejemplo, en el caso del dominio Bacteria, es frecuente hallar grupos muy grandes de organismos, muy similares entre si, que podrían afectar la ubicación de los arquetipos. Al trabajar sobre las observaciones en los límites y el politopo que puede ajustarse a esta forma, estas variaciones del dataset no impactan significativamente en las localizaciones de los arquetipos. Dado que se trabaja con organismos relacionados filogenéticamente, esta propiedad se vuelve particularmente relevante. Esto se aplica, de forma similar, en los ejemplos donde se utiliza este método para conjuntos de líneas celulares que se diversifican a partir de células madre [41]. El enfoque de ParTI, con sus algoritmos de búsqueda, mantiene su vigencia para estudios como el descripto en ese trabajo. Por estos motivos, se descartó el método AAnet y sólo se lo utilizó como una confirmación de lo obtenido por el método ParTI.

Como se exhibe en la sección 5.4, la aplicación del método ParTI en los tres datasets de trabajo mostró arquetipos compatibles con las distribuciones de los organismos. La reducción de la dimensionalidad en los tres datasets reveló, por medio de PCA, que las primeras 3 componentes son capaces de representar una buena proporción de la variabilidad de los datos (80 % - 90 %). Esto resulta en que las distribuciones de puntos son representadas adecuadamente en las primeras componentes principales. Los 4 arquetipos hallados, en cada dataset, se ubican por fuera del conjunto de puntos experimentales, remarcando su carácter de organismos extremos.

Como se muestra en la sección 5.5, las ubicaciones de los arquetipos permitieron dar cuenta de sus características en los conjuntos de variables propios de los datasets de trabajo. En todos los casos se lograron reconstruir los arquetipos, con todas sus variables, a partir de su ubicación en el espacio de dimensionalidad reducida (Fig. 24, 25 y 26). Las variables de cada dataset en su totalidad resultan complejas de analizar en su conjunto, sin embargo, logran mostrar las características principales de los arquetipos.

Las posiciones muestran que las abundancias de algunos codones y aminoácidos toman valores negativos (Fig. 24, 25 y 26). Esto ocurre debido a que el método elegido ubica los arquetipos por fuera del conjunto de observaciones, corriendo el riesgo de llegar a esta tipo de inconsistencias. Además, la sumatoria de las abundancias de cada uno no necesariamente es igua a 1. No existen, en ninguno de los métodos, restricciones que garanticen que las abundancias sean superiores a 0 ni su sumatoria sea igual a 1. Dado que existen otras restricciones al momento de localizar los arquetipos en algunos métodos, sería posible pensar en restricciones adicionales que garanticen estas dos condiciones para resolver este problema en particular. Los algoritmos de ParTI contienen funciones de perdida que ajustan las observaciones como sumas ponderadas de los arquetipos, aunque, en algunos casos, esta restricción no es estricta. Se hace posible entonces, adicionar términos a la función de ajuste que incluya que las abundancias sean superiores a 0 o que su sumatoria sea igual a 1. Por otro lado, se hace posible hallar el espacio entre las variables de abundancias donde se satisfagan estas propiedades y realizar la búsqueda de arquetipos en este espacio en particular. Sin embargo, estos hechos no impiden poder interpretar los valores y caracterizas los arquetipos. Las abundancias resultantes de las ubicaciones halladas deben interpretarse solo como características de fenotipos extremos ideales que ayudan a comprender la totalidad de la distribución de organismos en cada uno de los datasets.

Agrupar los diferentes arquetipos según las distancias al resto de los organismos confirmó la presencia de arquetipos comunes en los diferentes datasets (Fig. 27). Los grupos de arquetipos de diferentes datasets muestran posiciones consistentes en 3 de los 4 casos, dando cuenta de que los organismos plantean una distribución similar en los tres datasets, y que los arquetipos responden a esas distribuciones consistentemente. Los arquetipos 1 mantienen una posición asociada a bajo contenido GC, mientras que los arquetipos 3 y, especialmente, los arquetipos 4 se ubican como arquetipos caracterizados por un bajo contenido GC. Estos tres casos muestran consistencia en sus ubicaciones en los 3 datasets de trabajo. El arquetipo 2 es el que tiene un comportamiento más complejo y constituye el caso más interesante de estudio, dado que rompe la aparente correlación entre las abundancias de aminoácidos y sus codones que parece primar en el resto de los casos. El arquetipo 2 junto con el arquetipo 3 rompen con la distribución de los

organismos a lo largo del eje de mayor a menor GC constituído entre los arquetipo 1 y 4. El arquetipo 2 se ubica marcadamente cercano a organismos de GC intermedio, mientras que el arquetipo 3 solo aparece como alternativo al arquetipo 4. En el dataset 'Aminoácidos', el arquetipo 2 evidencia un eje de naturaleza diferente, que permite discriminar marcadamente los organismos del dominio Eukaryota (Fig. 28). Esto también aparece en los otros dos datasets, aunque con menor intensidad. Por otro lado, el agrupamiento de los arquetipos (Fig. 27) muestran que el arquetipo 2 también se asocia más al arquetipo 1 que a los otros dos. Las posiciones relativas hacia el resto de los organismos funciona como puente entre cada uno de los diferentes datasets, de forma de poder comparar arquetipos localizados en diferentes grupos de variables. Dado que los datasets están vinculados en cuanto a su constitución a partir de las relaciones establecidas por el código genético, es esperable cierto grado de consistencia en estas posiciones relativas. Tanto la consistencia en las posiciones como su elevada variabilidad, tal como sucede con el arquetipo 2, pueden brindar pistas sobre las relaciones de abundancias de aminoácidos y codones.

6.4 Caracterización de arquetipos

6.4.1 Variables taxonómicas

Como se menciona en la sección 5.6, las variables taxonómicas utilizadas muestran un patrón general en cuanto a su distribución en los diferentes datasets primarios. Los taxones ocupan espacios particulares en los grupos de variables, y los arquetipos dan cuenta de ello analizando sus posiciones relativas (Fig. 28 y 29). En la sección 5.6 se analizaron los taxones individualmente (Fig. 30 y 31) evaluando los arquetipos cercanos. Realizar un análisis desde los arquetipos para hallar los taxones cercanos (Fig. 46 y 47) permitió caracterizar los arquetipos de manera evidente. Si bien es posible relacionar arquetipos con taxones particulares, sus roles más generales pueden vislumbrarse con los dominios de mayor jerarquía. Son los principios generales los más relevantes, dado el limitado número de arquetipos y su relación con un gran conjunto de organismos de taxones diversos (Fig. 30 y 31).

Los arquetipo 3 y 4 poseen sus distribuciones de distancias a los diferentes dominios de forma similar. Al observar las medianas, se denota al arquetipo 3 asociado al dominio Bacteria, y al arquetipo 4 asociado a los dominios Bacteria y Archaea (Fig. 46). En el arquetipo 1 los valores de las medianas muestran a los

organismos del dominio Eukaryota como los más cercanos. En el caso del arquetipo 2 las distribuciones de distancias muestran una clara asociación con los organismos del dominio Eukaryota. Además, presenta distancias con menor dispersión dentro de cada grupo. Esto muestra que los diferentes dominios ocupan espacios más reducidos en la dirección del arquetipo 2, y se distribuyen a lo largo de los ejes delimitados por los otros arquetipos, en donde la distribución de distancias es más grande. Esto concuerda con lo observado en las distribuciones de organismos (Fig. 28).

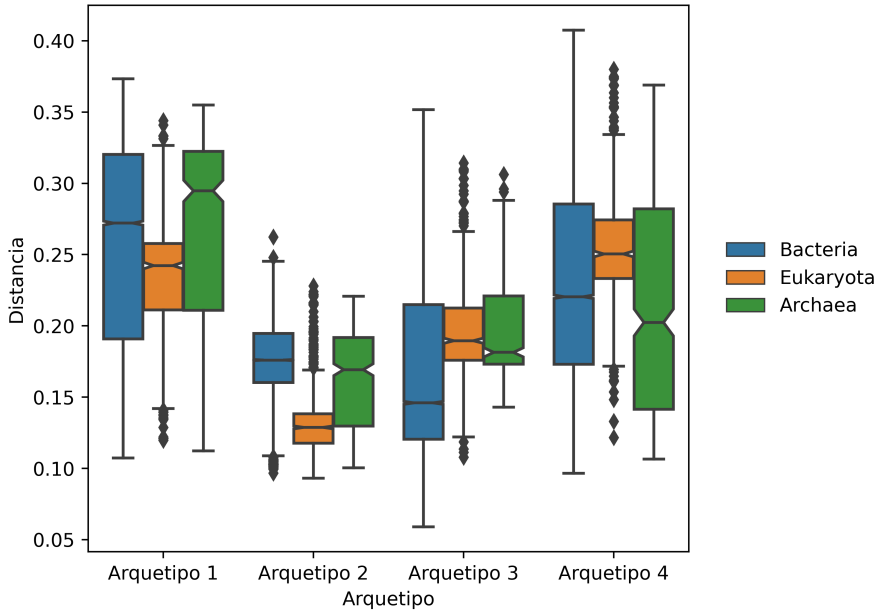

Figura 46: Relación entre dominios y arquetipos. Se muestran las distancias a cada arquetipo, halladas mediante el método de ParTI en el dataset 'Codones', para cada dominio. Esta gráfica repite lo expuesto en la Fig. 30 pero agrupados por arquetipo, de forma de hacer énfasis en las características de cada uno de ellos. Las muescas dan cuenta de los valores de las medianas con sus respectivos intervalos de confianza (95 %).

Los taxones de menor jerarquía muestran un patrón comparable a los observados en los dominios (Fig. 47). En el arquetipo 1 las medianas de las distribuciones de distancias muestran al grupo FCB como el más cercano. En cuanto al arquetipo 2, se observa una tendencia comparable en la baja dispersión de las distancias. Las medianas de las distribuciones muestran a los grupos FCB y PVC como los más cercanos. Las distribuciones de los phylum en los casos de los arquetipos 3 y 4 se observan semejantes, en donde el grupo de Proteobacteria posee una mediana menor que los otros tres (Fig. 47).

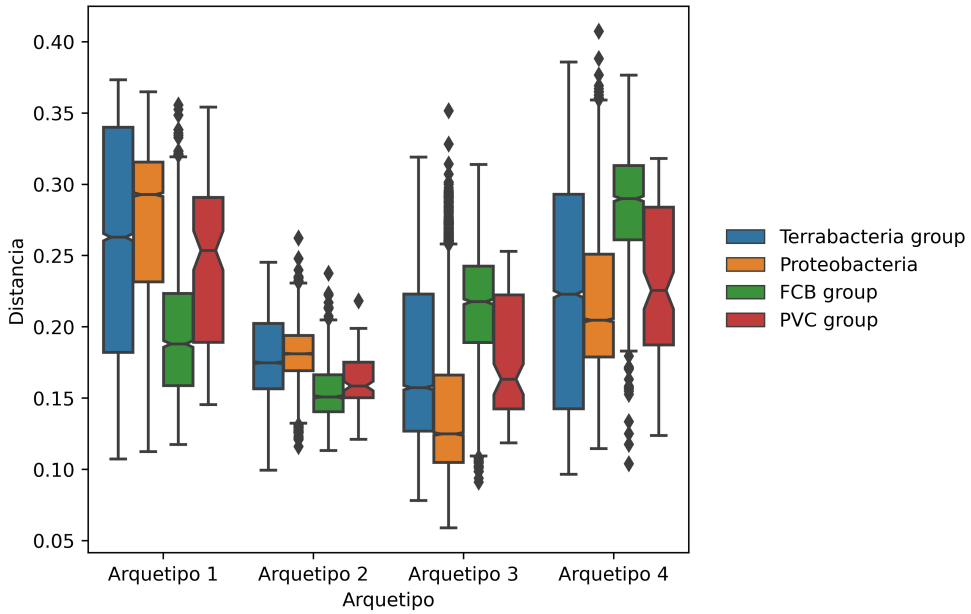

Figura 47: Relación entre phylum de Bacteria y arquetipos. Se muestran las distancias a cada arquetipo, halladas mediante el método ParTI en el dataset 'Codones', para cada phylum del dominio Bacteria. Esta gráfica repite lo expuesto en la Fig. 31 pero agrupados por arquetipo, de forma de hacer énfasis en las características de cada uno de ellos. Las muescas dan cuenta de los valores de las medianas con sus intervalos de confianza (95 %). Se repite, en cierta medida, lo observado para los dominios.

Los análisis muestran que lo observado en los dominios y phylum de Bacteria se repiten, en cierta forma, en otros taxones y permite dar cuenta de cómo los diferentes taxones se diferencian mejor en la dirección del arquetipo 2. Los grupos de organismos de un mismo dominio o phylum se orientan en sentido de los arquetipo 1 y 4 (Fig. 28 y 29). El arquetipo 2 señala, entonces, un eje sobre el cual los organismos se discriminan en diferentes taxones. Las asociaciones entre los arquetipos y los taxones particulares se discuten con más detalle junto con las variables continuas (ver más adelante).

6.4.2 Variables continuas

Como se describe en la sección 5.6, las variables de enriquecimiento constituyen el aspecto más rico en el análisis de los arquetipos, ya que se incluyen dimensiones adicionales con significados más diversos (metabolismo celular, propiedades moleculares, taxonomía, distribución de abundancias de codones o aminoácidos), que son utilizados para localizar a los arquetipos (datasets primarios). Como se mencionó anteriormente, se intentan asociar tareas o características generales a los arquetipos más allá de los dataset sobre los cuales se hallaron. Los mapas de variables

constituyen un elemento simple para poder asociar las características seleccionadas, si bien pueden aplicarse a otras características de igual forma, siempre que se defina a cada organismo con un único valor. Los mapas demostraron representar las relaciones de correlación entre las variables (Fig. 34).

Entre las variables incluidas, se hallan algunas típicamente asociadas con los sesgos en el uso de codones en la bibliografía disponible, como el contenido GC o la clasificación taxonómica (ver sección 3.2.2). Ambas variables mostraron patrones claros en su distribución dentro del conjunto de organismos. Por otro lado, no se incluyeron variables relacionadas con las abundancias de tRNA debido a la complejidad en la expresión de este tipo de secuencias, y en la baja disponibilidad de información en diferentes organismos. Además, tampoco se lograron incluir variables relacionadas con la tasa de expresión de las secuencias. Los niveles de expresión se asocian típicamente con genes particulares, por lo que se vuelve complejo poder incluir estas variables en un estudio a nivel de organismo. Es, sin lugar a duda, un aspecto relevante en caso de realizar un estudio comparable a lo realizado, pero a nivel de genes.

Debido al trabajo realizado en esta tesis, será posible realizar mapas con variables adicionales muy fácilmente, aplicando el MDS a partir de matrices de distancias de correlación entre variables, por lo que el enfoque planteado es escalable a un dataset de enriquecimiento mayor. Se agrupan las características principales de cada uno de los arquetipos, a modo de síntesis, de lo observado en la sección de resultados (Fig. 24, 25, 26, 39, 41, 42, 43 y 44) (Tabla 6).

Propiedades	Arquetipo 1	Arquetipo 2	Arquetipo 3	Arquetipo 4
Nucleótidos	Adenina Timina Purinas		GC (parcialmente)	Citosina Guanina GC GC sinónimo
Codones	UUA (Leu) AUU (Ile) AAU (Asn) AAA (Lys)	CAA (Gln) AAA (Lys) GAA (Asp) CUU (Leu) AGA (Arg)	CUG (Leu)	ACC (Thr) CAC (His) GCC (Ala) GAC (Asp)
Aminoácidos	Baja Ala Alta Ile Alta Lys Alta Asn Alta Phe Alta Tyr	Alta Ser Alta Glu	Alta Ala Alta Leu Alta Val	Alta Ala Baja Gln Alta Arg Alta Gly
Geométricas	Major Groove Depth Minor Groove Width		Tilt Slide	Roll Rise Major Groove Width Persistance Length
Energías	Baja Estabilidad: Alta Free Energy Alta Enthalpy Alta Entropy Alta Stacking Energy Alta Mobility Major Groove			Alta Estabilidad: Alta Melting Temperature Alta Mobility Minor Groove
Distribución		Alta h_C Alta h_A Alta h_R Alta CAI Alta w_mean Alta NC_sum Bajo fop		
Fisiológicas	AA_cost_abs	AA_cost		Robustness
Taxonomía	FCB group	Eukaryota Menor dispersión de taxones	Bacteria Proteobacteria	Bacteria Archaea

Tabla 6: Características de arquetipos. Se describen las características principales de los arquetipos, hallados mediante el método ParTI, tras relacionar las posiciones de cada arquetipo con las características de 'Enriquecimiento' de los organismos.

El arquetipo 1 muestra muchas características asociadas comúnmente a la baja proporción de contenido GC (Fig. 36 y 35). Elevados valores de Adenina y Timina, así como valores elevados de las variables del grupo de energías, le son propias y dan cuenta de su posición (Fig. 44 y 40). Aparecen, sin embargo, algunas características propias más allá de esta tendencia general. Valores elevados de Ile y Lys, así como abundancias muy bajas de Ala, son las abundancias más marcadas. También se destacan los aminoácidos Asn, Phe y Tyr (Fig. 44). Los codones más relacionados con este arquetipo son UUA (Leu), AUU (Ile), AAU (Asn) y AAA (Lys). Aparecen

variables relacionadas a los surcos del ADN que le son propias (Minor Groove Depth y Major Groove Width), y la mayoría de las variables relacionadas con la energía, las cuales dan cuenta de una baja estabilidad. También se observa la variable de costo de aminoácidos absoluto ('AA_cost_abs') (Fig. 43). Este arquetipo se opone a los arquetipos 3 y 4, pero sobre todo marca el eje principal junto con el arquetipo 4, donde el arquetipo 1 funciona como el correspondiente a bajo contenido GC y el arquetipo 4 como el de alto contenido GC. A diferencia del arquetipo 4, el arquetipo 1 no posee otro arquetipo tan cercano (Fig. 32). El arquetipo 2, si bien tiene algunas características comparables y está en una posición vecina, se asocia a variables muy diferentes.

En el caso del arquetipo 4, este se asocia principalmente a variables comúnmente relacionadas con un elevado contenido GC, y se correlaciona negativamente con el arquetipo 1. Este contenido GC se hace muy evidente en la tercera posición de sus codones, y se manifiesta en el contenido GC de codones sinónimos (Fig. 40). Entre los aminoácidos, la Ala aparece muy abundante en este arquetipo y se observan valores muy bajos de Gln. Al constituirse como polo principal de alto contenido GC, valores de Guanina y Citosina, son muy elevados. Además, entre las variables geométricas, aparecen las relacionadas con las variables 'Roll', 'Rise' y algunas relacionadas a los surcos como 'Major Groove Width'. Como es de esperarse, dado el elevado contenido GC, también tiene profunda asociación con la alta estabilidad del ADN, resultando en valores elevados de 'Melting Temperature' y 'Mobility Minor Groove'. Este arquetipo aparece asociado a valores elevados de 'Robustness', dando cuenta de una posible relación entre la robustez en el uso de codones y el resto de las variables. Finalmente, este arquetipo se halla cercano a regiones de organismos caracterizadas por pertenecer a Bacteria o Archaea (ver más adelante).

El arquetipo 3 muestra propiedades similares al arquetipo 4, compartiendo, parcialmente, una asociación con valores altos de contenido GC, si bien se denotan algunas diferencias. Se hallan cercanos en el espacio y en el mapa de variables, por lo que también se asocia a valores elevados de contenido GC. Aparecen algunas características propias como ser las abundancias del codón CUG (Leu) y los aminoácidos Ala, Val y Leu. En el caso de las variables geométricas 'Tilt' y 'Slide', se muestran como las más cercanas en los mapas de variables. Finalmente, organismos del dominio Bacteria aparecen más cercanos a este arquetipo. Este arquetipo se

asocia de forma distintiva a variables geométricas ('Tilt' y 'Slide') y a grupos taxonómicos como Proteobacteria.

Por último, el arquetipo 2 se constituye como un arquetipo intermedio entre el eje formado por los arquetipos 1 y 4 (Fig. 28 y 32). Se localiza en un sector intermedio entre estos dos arquetipo, y se asocia a valores intermedios de contenido GC. Muestra abundancias de codones CAA (Gln), AAA (Lys) y GAA (Asp), y una asociación con CUU (Leu) y AGA (Arg), según los mapas de variables. Además, se asocia con valores elevados del aminoácido Ser, y muestra buena correlación con valores elevados de Glu. Toma una posición ligeramente más céntrica en los mapas de variables, por lo que no se lo pudo asociar a variables geométricas o relacionadas con la energía. En el caso de las distribuciones, da cuenta de valores elevados de entropías de Shannon (h_C, h_A y h_R), número efectivo de codones (Nc_sum) e índices de adaptabilidad (CAI, w_mean), y valores bajos de 'fop'. Este arquetipo se muestra como un representante de una distribución homogénea de las abundancias, ya sea de codones como de aminoácidos. Además, se relaciona con valores elevado de 'AA_cost'. Finalmente, este arquetipo se localiza en las inmediaciones de conjuntos de organismos del dominio Eukaryota (ver más adelante). En los datasets primarios, el arquetipo 2 aparece en posiciones relativas más dispersas. Cada uno de los datasets lo coloca en lugares diferentes, dando cuenta de su sensibilidad a las distribuciones en cada uno de los conjuntos de variables primarias. Este hecho lo vuelve de particular interés, ya que da cuenta de las diferencias entre las distribuciones de codones y de los aminoácidos. Este arquetipo muestra las variables de distribución como las más distintivas. Además, se asocia particularmente con el dominio Eukaryota y con una baja distribución en las distancias de cada uno de los taxones.

Todos los arquetipos, en mayor o menor medida, pudieron asociarse con un conjunto de características. Los arquetipos 1 y 4 mantienen el eje principal guiado con el contenido GC sobre el cual se alinean los organismos y la mayoría de las variables. Los arquetipos 2 y 3 aportan ejes alternativos que pueden asociarse a otros conjuntos de variables. Cada arquetipo puede tomar características propias, donde los arquetipos 1 y 4 se asocian a organismos con menor y mayor contenido GC respectivamente, el arquetipo 2 con organismos más complejos y de GC intermedio, y el arquetipo 3 como alternativa al arquetipo 4, aunque con características particulares.

La caracterización de los arquetipos muestra un aspecto adicional a los análisis presentes en la bibliografía sobre los usos de codones y aminoácidos. Este enfoque aporta un nuevo marco a las descripciones existentes de los sesgos en los usos de codones y aminoácidos. Por otro lado, mayor cantidad de variables de enriquecimiento ayudarán a mejorar la caracterización de los arquetipos, sobre todo si las variables adicionales incluyen significados que vayan más allá de los analizados en este trabajo. Poder construir variables nuevas de fuentes adicionales complementará la caracterización realizada hasta ahora, y creemos que abrirá el camino hacia nuevos interrogantes.

6.5 Discusión general

A través de este trabajo, se logró realizar un análisis de arquetipos sobre las distribuciones de abundancias de los codones y aminoácidos de un gran número de organismos vivos. Se aplicaron 2 metodologías de búsquedas de arquetipos distintas, ParTI y AAnet, seleccionando al método ParTI como el más idóneo para este fin. Se lograron hallar 4 arquetipos en posiciones esperables para el conjunto de datos utilizado, y asociarles características particulares. Se utilizaron todos los datos de abundancias de codones y aminoácidos disponibles en las bases de datos, incluyendo una gran diversidad de organismos. Se incluyó una gran cantidad de variables de enriquecimiento de diferente naturaleza, aportando múltiples perspectivas al análisis, y dando lugar a la inclusión de mayor cantidad de características para expandir el estudio.

El análisis realizado complementa lo existente en la bibliografía, ya que se aplican métodos novedosos al estudio de los sesgos de codones, donde además, se incluyen una diversidad de organismos y variables de enriquecimiento, ausentes en estudios previos. Si bien en este trabajo se logran relacionar algunas variables con los arquetipos hallados, también se vislumbran potenciales oportunidades de análisis adicionales como el estudio de subgrupos de organismos, o la construcción de nuevas variables.

Los métodos utilizados para las búsqueda de arquetipos podrían complementarse con restricciones adicionales, que mantengan los valores de abundancias coherentes entre sí, de forma de poder armar organismos arquetípicos más consistentemente. También, la búsqueda de arquetipos fue realizada sobre los datasets de forma

independiente, por lo que lograr un métodos capaz de buscar arquetipos en los tres datasets primarios al mismo tiempo podría arrojar resultados adicionales. Además, la aplicación de los métodos sobre las abundancias podría tener un correlato en las variables de enriquecimiento, por lo que aplicar las metodologías de búsqueda de arquetipos sobre las variables de enriquecimiento resultaría en una dimensión complementaria del análisis.

Por otro lado, la disponibilidad de secuencias biológicas se encuentra en constante aumento, de esta manera se vuelve factible expandir el análisis con un mayor volumen de datos en un futuro. El mismo enfoque puede utilizarse para análisis de uso de codones de secuencias excluidas en este análisis como plásmidos, virus u organelas. Puede incluirse, además, información relacionada con la transcriptómica y proteómica, como son las abundancias de ARN mensajeros o de proteínas, y relacionarlas con las abundancias ya incluidas en los datasets primarios. También pueden analizarse los sesgos de codones a nivel metagenómico, de cepas o a nivel de genes individuales, expandiendo lo ya estudiado a nivel de especie. Otros aspectos de interés incluyen al uso de codones 'Stop', el uso de códigos genéticos distintos del estándar, la relación de las abundancias con el árbol filogenético, y las relaciones entre las abundancias de los virus o parásitos y sus hospedadores. Por último, las variables de enriquecimiento representan solo una muestra de potenciales variables que pudiesen aportar mayor claridad sobre la funcionalidad de los arquetipos y el ordenamiento de los diferentes organismos en los conjuntos de variables determinados por cada datatset [80, 81, 82].

La biodiversidad se ve reflejada en las diferentes secuencias biológicas que se hallan en todos los organismos vivos. El uso de metodologías de aprendizaje automático, como el análisis de arquetipos, permite describir mejor el comportamiento de la información contenida en las secuencias biológicas de los diferentes grupos de organismos. El análisis de arquetipos funciona como una herramienta invaluable para la condensación de un gran y variado volumen de información, en un marco conceptual simple y generalizable, contribuyendo a una comprensión más acabada del fenómeno en estudio.

Referencias

[1] Curtis H, Schnek A. 2008 *Curtis. Biología.* Ed. Médica Panamericana.

[2] Barbieri M. 2008 *The Codes of life: the rules of macroevolution*, vol. 45. (doi:10.5860/choice.45-5558).

[3] Bobay LM, Ochman H. 2018 Factors driving effective population size and pan-genome evolution in bacteria. *BMC evolutionary biology* **18**, 1, 153. (doi:10.1186/s12862-018-1272-4).

[4] Darwin C. 1859 *On the origin of species.* UK: John Murray.

[5] Woese CR, Kandler O, Wheelis ML. 1990 Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 12, 4576–4579. (doi:10.1073/pnas.87.12.4576).

[6] Albarracín A, Teulón AA. 1993 *La teoría celular en el siglo XIX*, vol. 42. Ediciones AKAL.

[7] Nelson DL, Cox MM. 2017 *Lehninger Principles of Biochemistry 7th*, vol. 2.

[8] Alberts B, Bray D, Hopkin K, Johnson AD, Lewis J, Raff M, Roberts K, Walter P. 2013 *Essential cell biology.* Garland Science.

[9] Carey N. 2015 *Junk DNA: a journey through the dark matter of the genome.* Columbia University Press.

[10] Lodish H, Berk A, Kaiser CA, Krieger M, Scott MP, Bretscher A, Ploegh H, Matsudaira P, *et al.* 2008 *Molecular cell biology.* Macmillan.

[11] Jukes TH, Osawa S. 1993. Evolutionary changes in the genetic code. (doi:10.1016/0305-0491(93)90122-L).

[12] WikimediaCommons. 2008. Central dogma of molecular biochemistry with enzymes.

[13] Hershberg R, Petrov DA. 2008 Selection on codon bias. *Annual Review of Genetics* **42**, 1, 287–299. (doi:10.1146/annurev.genet.42.110807.091442).

[14] López JL, Lozano MJ, Lagares A, Fabre ML, Draghi WO, Papa MFD, Pistorio M, Becker A, Wibberg D, Schlüter A, *et al.* 2019 Codon Usage Heterogeneity in the Multipartite Prokaryote Genome: Selection-Based Coding Bias Associated with Gene Location, Expression Level, and Ancestry. *mBio* **10**, 3. (doi:10.1128/MBIO.00505-19).

[15] Forcelloni S, Giansanti A. 2020 Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome. *Journal of Molecular Evolution* **88**, 2, 164–178. (doi:10.1007/s00239-019-09921-4).

[16] Sharp PM, Emery LR, Zeng K. 2010 Forces that influence the evolution of codon bias. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **365**, 1544, 1203–12. (doi:10.1098/rstb.2009.0305).

[17] Tuller T. 2014 Challenges and obstacles related to solving the codon bias riddles. *Biochemical Society Transactions* **42**, 1, 155–159. (doi:10.1042/BST20130095).

[18] Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010 An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 2, 344–354. (doi:10.1016/j.cell.2010.03.031).

[19] Musto H. 2016 What We Know and What We Should Know About Codon Usage. *Journal of Molecular Evolution* **82**, 6, 245–246. (doi:10.1007/s00239-016-9742-z).

[20] Krick T, Verstraete N, Alonso LG, Shub DA, Ferreiro DU, Shub M, Sánchez IE. 2014 Amino acid metabolism conflicts with protein diversity. *Molecular Biology and Evolution* (doi:10.1093/molbev/msu228).

[21] Goncearenco A, Berezovsky IN. 2014 The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. *Biology Direct* **9**, 1, 29. (doi:10.1186/s13062-014-0029-2).

[22] Elofsson A. 2019 The Use of GC-, Codon-, and Amino Acid-frequencies to Understand the Evolutionary Forces at a Genomic Scale. *bioRxiv* p. 863142. (doi:10.1101/863142).

[23] Ikemura T. 1985 Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**, 1, 13–34. (doi:10.1093/oxfordjournals.molbev.a040335).

[24] Wei Y, Silke JR, Xia X. 2019 An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Scientific reports* **9**, 1, 3184. (doi:10.1038/s41598-019-39369-x).

[25] Bulmer M. 1987 Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 6106, 728–730. (doi:10.1038/325728a0).

[26] Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**, 1, 143–155. (doi:10.1016/S0378-1119(99)00225-5).

[27] Yannai A, Katz S, Hershberg R. 2018 The Codon Usage of Lowly Expressed Genes Is Subject to Natural Selection. *Genome biology and evolution* **10**, 5, 1237–1246. (doi:10.1093/gbe/evy084).

[28] Labella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2019 Variation and selection on codon usage bias across an entire subphylum. *PLOS Genetics* **15**, 7, e1008304. (doi:10.1371/journal.pgen.1008304).

[29] Liu X, He T, Guo Z, Ren M. 2019 Predicting Essential Genes of Escherichia coli based on Clustering Method. In: *Proceedings of the 2019 11th International Conference on Bioinformatics and Biomedical Technology - ICBBT'19*, pp. 45–49. New York, New York, USA: ACM Press. (doi:10.1145/3340074.3340080).

[30] Novoa EM, Jungreis I, Jaillon O, Kellis M. 2019 Elucidation of Codon Usage Signatures across the Domains of Life. *Molecular Biology and Evolution* (doi:10.1093/molbev/msz124).

[31] Malakar AK, Halder B, Paul P, Deka H, Chakraborty S. 2019 Genetic evolution and codon usage analysis of NKX-2.5 gene governing heart development in some mammals. *Genomics* (doi:10.1016/J.YGENO.2019.07.023).

[32] Mahajan S, Agashe D. 2018 Translational Selection for Speed Is Not Sufficient to Explain Variation in Bacterial Codon Usage Bias. *Genome biology and evolution* **10**, 2, 562–576. (doi:10.1093/gbe/evy018).

[33] Knight RD, Freeland SJ, Landweber LF. 2001 A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome biology* **2**, 4, RESEARCH0010. (doi:10.1186/gb-2001-2-4-research0010).

[34] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013 GenBank. *Nucleic Acids Research* **41**, D1. (doi:10.1093/nar/gks1195).

[35] Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019 GenBank. *Nucleic Acids Research* **47**, D1, D94—-D99. (doi:10.1093/nar/gky989).

[36] Nasko DJ, Koren S, Phillippy AM, Treangen TJ. 2018 RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biology* **19**, 1, 165. (doi:10.1186/s13059-018-1554-6).

[37] Holcomb DD, Alexaki A, Katneni U, Kimchi-Sarfaty C. 2019. The Kazusa codon usage database, CoCoPUTs, and the value of up-to-date codon usage statistics. (doi:10.1016/j.meegid.2019.05.010).

[38] Lemey P, Salemi M, Vandamme AM. 2009 *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing.* Cambridge University Press.

[39] Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, Dekel E, Kavanagh K, Alon U. 2012 Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. *Science* **336**, 6085, 1157–1160. (doi:10.1126/science.1217405).

[40] Sheftel H, Shoval O, Mayo A, Alon U. 2013 The geometry of the Pareto front in biological phenotype space. *Ecology and Evolution* **3**, 6, 1471–1483. (doi:10.1002/ece3.528).

[41] Hart Y, Sheftel H, Hausser J, Szekely P, Ben-Moshe NB, Korem Y, Tendler A, Mayo AE, Alon U. 2015 Inferring biological tasks using Pareto analysis of

high-dimensional data. *Nature Methods* **12**, 3, 233–235. (doi:10.1038/nmeth.3254).

[42] Szekely P, Korem Y, Moran U, Mayo A, Alon U. 2015 The Mass-Longevity Triangle: Pareto Optimality and the Geometry of Life-History Trait Space. *PLoS Computational Biology* **11**, 10. (doi:10.1371/journal.pcbi.1004524).

[43] Seward EA, Kelly S. 2018 Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. *Genome Biology* **19**, 1, 1–11. (doi:10.1186/s13059-018-1480-7).

[44] Cutler A, Breiman L. 1994 Archetypal Analysis. *Technometrics* **36**, 4, 338–347. (doi:10.1080/00401706.1994.10485840).

[45] Bauckhage C, Thurau C. 2009 Making Archetypal Analysis Practical. pp. 272–281. Springer, Berlin, Heidelberg. (doi:10.1007/978-3-642-03798-6_28).

[46] Frankino WA, Zwaan BJ, Stern DL, Brakefield PM. 2005 Natural selection and developmental constraints in the evolution of allometries. *Science* **307**, 5710, 718–720. (doi:10.1126/science.1105409).

[47] Thøgersen JC, Mørup M, Damkiær S, Molin S, Jelsbak L. 2013 Archetypal analysis of diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics* **14**, 1, 279. (doi:10.1186/1471-2105-14-279).

[48] Mørup M, Hansen LK. 2012 Archetypal analysis for machine learning and data mining. *Neurocomputing* **80**, 54–63.

[49] Dijk DV, Burkhardt DB, Amodio M, Tong A, Wolf G, Krishnaswamy S. 2019 Finding Archetypal Spaces Using Neural Networks. In: *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 2634–2643. Institute of Electrical and Electronics Engineers Inc. (doi:10.1109/BigData47090.2019.9006484).

[50] Wirth R, Hipp J. 2000 Crisp-dm: Towards a standard process model for data mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29–39. Springer-Verlag London, UK.

[51] Marbán Ó, Mariscal G, Segovia J. 2009 A data mining & knowledge discovery process model. In: *Data mining and knowledge discovery in real life applications*. IntechOpen.

[52] Wold S, Esbensen K, Geladi P. 1987 Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**, 1-3, 37–52. (doi:10.1016/0169-7439(87)80084-9).

[53] Abdi H, Williams LJ. 2010 Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 4, 433–459. (doi:10.1002/wics.101).

[54] Shannon CE. 1948 A Mathematical Theory of Communication **27**, 3. (doi:10.1002/j.1538-7305.1948.tb01338.x).

[55] Zeeberg B. 2002 Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome Research* **12**, 6, 944–955. (doi:10.1101/gr.213402).

[56] Sharp PM, Li WH. 1987 The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**, 3, 1281–1295. (doi:10.1093/nar/15.3.1281).

[57] Carbone A, Zinovyev A, Képès F. 2003 Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* **19**, 16, 2005–2015. (doi:10.1093/bioinformatics/btg272).

[58] Lee S, Weon S, Lee S, Kang C. 2010 Relative Codon Adaptation Index, a Sensitive Measure of Codon Usage Bias. *Evolutionary Bioinformatics* **6**, EBO.S4608. (doi:10.4137/EBO.S4608).

[59] Lavner Y, Kotlar D. 2005 Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345**, 1 SPEC. ISS., 127–138. (doi:10.1016/j.gene.2004.11.035).

[60] Wright F. 1990 The 'effective number of codons' used in a gene. *Gene* **87**, 1, 23–29. (doi:10.1016/0378-1119(90)90491-9).

[61] Chen WHH, Lu G, Bork P, Hu S, Lercher MJ. 2016 Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nature Communications* **7**, 1, 11334. (doi:10.1038/ncomms11334).

[62] Akashi H, Gojobori T. 2002 Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6, 3695–3700. (doi:10.1073/pnas.062526999).

[63] Grantham R. 1974 Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)* **185**, 4154, 862–4. (doi:10.1126/science.185.4154.862).

[64] Shenhav L, Zeevi D. 2019 Resource conservation manifests in the genetic code. *bioRxiv* p. 790345. (doi:10.1101/790345).

[65] Friedel M, Nikolajewa S, Sühnel J, Wilhelm T. 2009 DiProDB: a database for dinucleotide properties. *Nucleic acids research* **37**, Database issue, D37–40. (doi:10.1093/nar/gkn597).

[66] Ghorbani M, Mohammad-Rafiee F. 2011 Geometrical correlations in the nucleosomal DNA conformation and the role of the covalent bonds rigidity. *Nucleic Acids Research* **39**, 4, 1220–1230. (doi:10.1093/nar/gkq952).

[67] Vieira-Silva S, Rocha EPC. 2010 The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genetics* **6**, 1, e1000808. (doi:10.1371/journal.pgen.1000808).

[68] Li J, Bioucas-Dias JM. 2008 Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data. In: *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 3. (doi:10.1109/IGARSS.2008.4779330).

[69] Bioucas-Dias JM. 2009 A variable splitting augmented Lagrangian approach to linear spectral unmixing. In: *WHISPERS '09 - 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing.* (doi:10.1109/WHISPERS.2009.5289072).

[70] Chan TH, Chi CY, Huang YM, Ma WK. 2009 A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing* **57**, 11, 4418–4432. (doi:10.1109/TSP.2009.2025802).

[71] Chan TH, Liou JY, Ambikapathi AM, Ma WK, Chi CY. 2012 Fast algorithms for robust hyperspectral endmember extraction based on worst-case simplex volume maximization. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 1237–1240. (doi:10.1109/ICASSP.2012.6288112).

[72] Ringnér M. 2008 What is principal component analysis? *Nature Biotechnology* **26**, 3, 303–304. (doi:10.1038/nbt0308-303).

[73] Chen Ch, Härdle W, Unwin A. 2008 *Handbook of Data Visualization*. (doi:10.1007/978-3-540-33037-0).

[74] Deza MM, Deza E. 2009 *Encyclopedia of distances*. (doi:10.1007/978-3-642-00234-2).

[75] Chen Ch, Härdle W, Unwin A, Cox MAA, Cox TF. 2008 Multidimensional Scaling. In: *Handbook of Data Visualization*, pp. 315–347. Springer Berlin Heidelberg. (doi:10.1007/978-3-540-33037-0_14).

[76] Dawyndt P, De Meyer H, De Baets B. 2006 UPGMA clustering revisited: A weight-driven approach to transitive approximation. *International Journal of Approximate Reasoning* **42**, 3, 174–191. (doi:10.1016/j.ijar.2005.11.001).

[77] Murtagh F, Contreras P. 2012 Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 1, 86–97. (doi:10.1002/widm.53).

[78] Jain AK. 2010 Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **31**, 8, 651–666. (doi:10.1016/J.PATREC.2009.09.011).

[79] Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. (doi:10.1016/j.bbrc.2006.06.054).

[80] Reimer LC, Vetcininova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, Overmann J. 2019 BacDive in 2019: Bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Research* **47**, D1, D631–D636. (doi:10.1093/nar/gky879).

[81] Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, Engqvist MK, Fierer N, Geoghegan JL, Gillings M, *et al.* 2020 A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data* **7**, 1. (doi:10.1038/s41597-020-0497-4).

[82] Barberán A, Caceres Velazquez H, Jones S, Fierer N. 2017 Hiding in Plain Sight: Mining Bacterial Species Records for Phenotypic Trait Information. *mSphere* **2**, 4. (doi:10.1128/msphere.00237-17).