UNIVERSITAT ROVIRA i VIRGILI

ESTUDIO DEL FENÓMENO DE SOLAPAMIENTO DE GENES EN GENOMAS BACTERIANOS MEDIANTE HERRAMIENTAS BIOINFORMÁTICAS

Edgar Pedro Chacón Sánchez

TRABAJO FINAL DE GRADO BIOTECNOLOGÍA

Tutor académico | **Dr. Gerard Pujadas Anguiano**
Grado en Biotecnología
Departamento de Bioquímica y Biotecnología
Universidad Rovira i Virgili, Tarragona
gerard.pujadas@urv.cat

En cooperación con | ***Cheminformatics and Nutrition Research Group***
Departamento de Bioquímica y Biotecnología
Universidad Rovira i Virgili, Tarragona

Supervisor | **Dr. Santiago Garcia-Vallvé**
Grado en Biotecnología
Departamento de Bioquímica y Biotecnología
Universidad Rovira i Virgili, Tarragona
santi.garcia-vallve@urv.cat

Fecha de convocatoria | Junio 2023

Yo, Edgar Pedro Chacón Sánchez, con DNI 39942567-P, soy conocedor de la guía de prevención del plagio en la URV *"*Prevención, detección y tratamiento del plagio en la docencia: guía para estudiantes" (aprobada en julio 2017) (https://www.crai.urv.cat/es/servicios/apoyo-aprendizaje/plagio/) y afirmo que este TFG no constituye ninguna de las conductas consideradas como plagio por la URV.

Tarragona, 6 de junio de 2023

Firma | 

1 | Índice

2 | Datos del centro

Este trabajo de final de grado parte de un estudio realizado en colaboración con el grupo de investigación Quimioinformática y nutrición (Cheminformatics and Nutrition Research Group) de la Universidad Rovira i Virgili, Tarragona (https://www.cheminformatics-nutrition.recerca.urv.cat/).

En su mayoría, el estudio ha podido ser realizado de manera telemática dado el carácter informático que tiene, no requiriendo de ningún equipo especial durante gran parte del desarrollo de este, a excepción de un punto del estudio, donde se utilizaron equipos de la universidad para poder trabajar la cantidad de datos que se requerían, los cuales eran excesivos para su manipulación con un ordenador de uso común.

No obstante, durante la realización del estudio ha habido reuniones periódicas y una constante comunicación con el Dr. Santiago Garcia-Vallvé, uno de los directores del grupo junto al Dr. Gerard Pujadas Anguiano, ambos profesores de la universidad.

El grupo de investigación se especializa en el uso de herramientas computacionales aplicadas a bases de datos de productos naturales para encontrar nuevos ingredientes bioactivos y diseñar fármacos. El proyecto de los genes solapados es un estudio independiente a su área principal de investigación, sin embargo, mantiene en común su base bioinformática.

3 | Abstract / Resumen

The phenomenon of gene overlap occurs when two coding sequences of two genes share at least one nucleotide. This phenomenon has been extensively studied in viruses, but to a much lesser extent in bacteria. In this study, the different types of overlaps and their characteristics are analysed in 5.223 bacterial genomes. These genomes have been bioinformatically processed using a program developed and supervised by the authors of this study, resulting in over 3 million pairs of overlapping genes, which have been analysed for further study.

Keywords | **gene overlap, genomics, bioinformatics, start codons, STOP codons.**

El fenómeno de solapamiento de genes ocurre cuando dos secuencias codificantes de dos genes comparten al menos un nucleótido. Este fenómeno está extensamente estudiado en virus, pero en bastante menor medida en bacterias. En este estudio, se analizan los diferentes tipos de solapamientos y sus características en 5.223 genomas bacterianos, que han sido tratados bioinformáticamente mediante un programa codificado y supervisado por los propios autores de este estudio, resultando en más de 3 millones de pares de genes solapados, los cuales han sido analizados para su estudio.

Palabras clave | **solapamiento de genes, genómica, bioinformática, codones de inicio, codones de STOP.**

4 | Introducción

La naturaleza de triplete del código genético y la configuración de doble cadena del ADN implican que seis secuencias de aminoácidos pueden estar codificadas conceptualmente dentro de una misma secuencia de nucleótidos, pero en diferentes marcos de lectura (ORFs, *open reading frames* en inglés) (Wichmann et al., 2021), tal como queda ilustrado en la Figura 1.

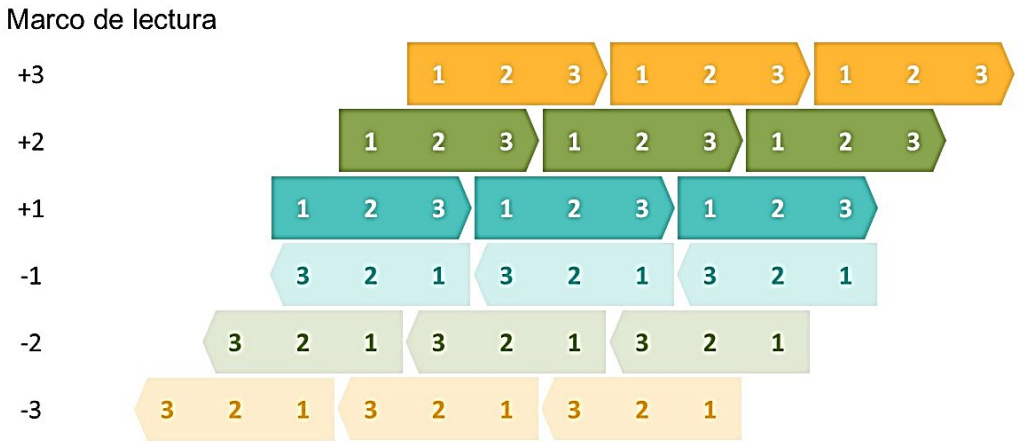

Figura 1 | Ilustración de los 6 posibles ORFs, siendo el "+1" el marco estándar o de referencia, con “+2” y “+3” como alternativas y el “-1” la cadena antisentido, con “-2” y “-3” como alternativas. Figura adaptada de Wichmann et al. (2021).

Cuando dos genes (dos ORFs codificantes para proteínas) se encuentran en el mismo locus, pero entre ellos mantienen una pauta de lectura diferente, se consideran genes solapados (Kreitmeier et al., 2022) y, por lo tanto, una sección de ADN forma parte de ambos ORFs (Huvet & Stumpf, 2014).

Si ambos genes se encuentran en el mismo marco de lectura, el solapamiento será en fase 0, mientras que, si hay un desfase ente ellos de +1 o +2 nucleótidos, respecto el inicio del primer codón compartido total o parcialmente, el solapamiento será de fase 1 o 2 respectivamente.

De esta forma, y tal como se puede observar en la Figura 2, los solapamientos de 1, 4, 7, 10, … nucleótidos corresponden a una Fase 2; los solapamientos de 2, 5, 8, 11, … nucleótidos a una Fase 1 y, finalmente, los solapamientos de 3, 6, 9, 12, … nucleótidos corresponden a una Fase 0.

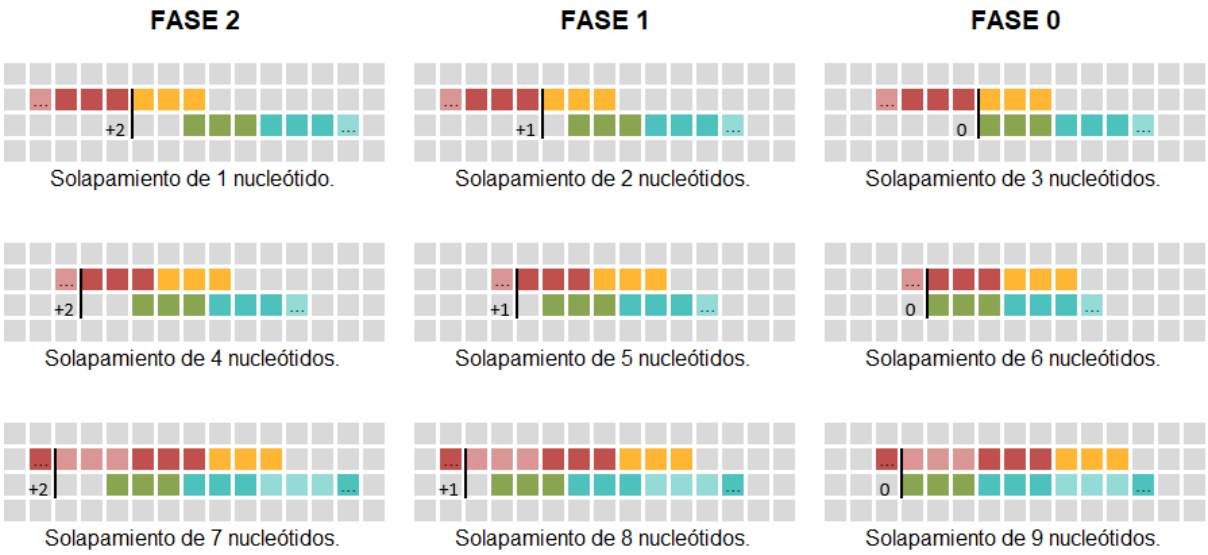

Figura 2 | Esquema representativo de las 3 posibles fases en las que se puede encontrar un gen respecto al otro en un solapamiento con los 3 primeros casos de cada uno a modo de ejemplo. Todos ellos siguen una direccionalidad estándar 5' → 3', siendo representado en naranja los codones de STOP del gen 1 y en verde los codones de inicio del gen 2 y, las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente). Estos ejemplos no son indicativos que no puedan existir otras combinaciones de direccionalidades entre los genes implicados, para más información ver apartado 4.2 | Clasificación más adelante.

El fenómeno de solapamiento de genes se lleva observando desde los inicios de la secuenciación y de la genómica (Wright et al., 2022). No obstante, hasta hace poco, la mayoría de los conocimientos sobre el fenómeno se centraban en los genomas virales (Wright et al., 2022). Sin embargo, contrariamente a lo que se esperaba, este fenómeno presenta cada vez más evidencia de que ocurre también tanto en organismos unicelulares como pluricelulares (Wichmann et al., 2021).

Por consiguiente, estas estructuras solapadas pueden ser observadas tanto en virus como en procariotas y eucariotas (Huvet & Stumpf, 2014). Pero, a pesar de la creciente evidencia de su abundancia, en general, la codificación de genes solapados aún no se considera un fenómeno significativo más allá de los virus, quizás debido a las dificultades percibidas en su evolución para uno o varios ORFs (Wichmann et al., 2021), o al hecho de que los genes solapados en organismos no víricos normalmente son rechazados categóricamente y a los que ya se encuentran anotados se les considera como un error en la anotación (Kreitmeier et al., 2022).

Para que dos ORFs sean designados como genes solapados, el gen madre (más información en el apartado 4.3 | Origen más adelante) debe encontrarse bien anotado y debe codificar por una proteína (Kreitmeier et al., 2022).

La mayoría de los solapamientos de genes caracterizados en detalle presentan una longitud inferior a los 200 codones (Kreitmeier et al., 2022).

4.1 | Significado y Función

Puesto que este documento se centra en los solapamientos ocurridos en genomas bacterianos, es decir procariotas, se utilizará la definición de Wright et al. (2022), que denomina solapamiento de genes al hecho de que dos secuencias codificantes (CDSs, *coding sequences* en inglés) de dos genes compartan al menos un nucleótido, ya sea en la misma cadena o en la complementaria. Esta definición es la que se aplica generalmente en la literatura de virus y procariotas, mientras que, en la literatura sobre eucariotas, el solapamiento se considera entre los límites de la transcripción primaria, tal como se puede observar en la Figura 3.

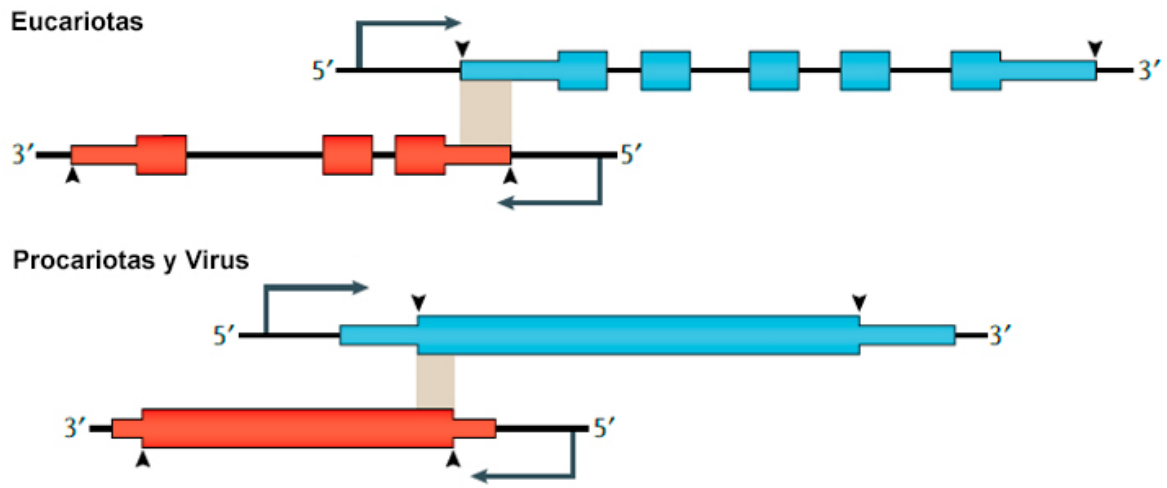

Figura 3 | Diferencias de las definiciones de solapamiento entre eucariotas y procariotas + virus ilustradas. Figura adaptada de Wright et al. (2022).

Distintos estudios anteriores han intentado caracterizar el solapamiento de genes en bacterias, revelando una correlación entre el número de genes que contiene el genoma del organismo y el número de solapamientos que este presenta: se estima

que aproximadamente un tercio de los genes se encuentra involucrado en algún tipo de solapamiento (Huvet & Stumpf, 2014; Wright et al., 2022).

Diversas hipótesis han surgido intentando explicar el beneficio o propósito de la formación de estas estructuras:

- Para **mejorar la compactación del genoma**, hecho que se encuentra fuertemente relacionado con los primeros genomas en los que se observó este fenómeno, los virus, los cuales presentan un gran número de genes solapados, hecho que les confiere una reducción significativa del número de nucleótidos necesarios para codificar genes, y reduciendo el tiempo y material necesario para generar nuevo material genético en la creación de nuevos virus (Huvet & Stumpf, 2014). No obstante, evidencia reciente sugiere que la hipótesis de reducción del genoma presenta una escasa capacidad explicativa (Wichmann et al., 2021).

- Como implicación en la **regulación de la traducción** mediante mecanismos de emparejado traduccional, basándose en las propiedades moleculares de la maquinaria de traducción (Huvet & Stumpf, 2014). Los genes codificantes para proteínas son transcritos a ARN, para ser traducidos, tras una posible modificación, a proteínas por esta maquinaria de traducción (Huvet & Stumpf, 2014). En bacterias, algunos genes codificantes para proteínas se encuentran formando sets (operones) (Huvet & Stumpf, 2014). Los cuales son transcritos juntos dando lugar a un ARN policistrónico (Huvet & Stumpf, 2014). Este será traducido en tantas proteínas como genes codificantes contenga el operón (Huvet & Stumpf, 2014). Diversas líneas de evidencia sugieren que la producción de estas proteínas puede no ser independiente y que, tras la traducción de la secuencia correspondiente a un gen, la misma maquinaria traduccional puede continuar y comenzar directamente la traducción del siguiente gen (Huvet & Stumpf, 2014). Este proceso, conocido como emparejamiento traduccional (*translational coupling* en inglés), es dependiente de la distancia entre las regiones codificantes para proteínas sucesivas dentro del ARN mensajero (ARNm) policistrónico y se encuentra muy extendido en los dominios Arquea y Bacteria (Huber et al., 2019; Huvet & Stumpf, 2014). La

presencia de genes solapados puede tener un impacto en la viabilidad de este proceso (Huvet & Stumpf, 2014).

Estos efectos en la regulación génica pueden ser relevantes en dominios taxonómicos, puesto que es posible que dos genes solapados en la misma cadena de ADN puedan ser co-expresados de manera más efectiva si están codificados dentro del mismo ARNm, mientras que los genes que presentan una superposición antisentido también podrían afectarse mutuamente, de manera similar a lo que se ha denominado recientemente "operón no contiguo", donde los genes que no se superponen pero se codifican en antisentido entre sí, se expresan juntos como un operón (Huvet & Stumpf, 2014).

- Y, por último, la posibilidad de **generar genes *de novo*** en el caso de un solapamiento de ORFs a genes existentes, que pueden ser traducidos y dar lugar a nuevos genes (Wichmann et al., 2021).

A pesar de toda la información que tenemos sobre el fenómeno de solapamiento de genes, las posibles implicaciones biológicas aún se encuentran lejos de ser comprendidas (Huvet & Stumpf, 2014), no obstante, la prevalencia de genes solapados en todo tipos de genomas puede presentar aplicaciones potenciales en biotecnología en el momento que se obtenga una mejor comprensión de sus mecanismos y exista una mayor investigación respecto a sus implicaciones biológicas fundamentales (Wichmann et al., 2021).

4.2 | Clasificación

La clasificación de los genes solapados se basa en su direccionalidad, siendo tres las posibles topologías que pueden presentar (Wright et al., 2022). Los solapamientos codireccionales (también conocidos como unidireccionales) ocurren entre genes codificados en la misma cadena. Mientras que las otras dos topologías ocurren en solapamientos entre genes de cadenas opuestas, siendo denominados convergentes o divergentes (Wright et al., 2022). Estas tres topologías que dan lugar a su clasificación quedan representadas en la siguiente Figura 4:

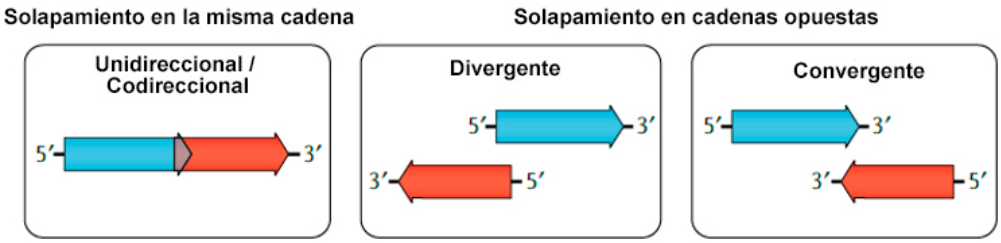

Figura 4 | Representación gráfica de los tres tipos de topologías que pueden presentar los solapamientos de genes/ORFs. Figura adaptada de Wright et al. (2022).

El solapamiento de genes codireccional representa la inmensa mayoría de los genes solapados en procariotas, representando el 84% del total de CDSs solapados (Huvet & Stumpf, 2014; Wright et al., 2022). Este tipo de solapamiento se origina por la pérdida del codón de inicio o del codón de STOP, resultando en que uno de los genes adyacentes (inicialmente no solapados) extienda su secuencia codificante hasta la del gen adyacente (Wright et al., 2022). Se estima que más del 98% de este tipo de solapamientos presenta una longitud inferior a los 60 pb, concentrando su inmensa mayoría en 1 pb o 4 pb solapados (Wright et al., 2022).

A diferencia de los codireccionales, los solapamientos convergentes presentan una menor incidencia.

Finalmente, el solapamiento de tipo divergente es el que menor incidencia presenta, siendo sustancialmente más raros que los solapamientos convergentes (Huvet & Stumpf, 2014; Wright et al., 2022).

Los solapamientos codireccionales son los más frecuentes en bacterias y virus, mientras que los solapamientos divergentes y convergentes son más frecuentes en eucariotas (Wright et al., 2022).

No obstante, las interacciones entre dos genes no solo pueden ser como solapamientos, con una parte de cada ge nen común, sino que también pueden ser en anidados (*nested* en inglés), donde la extensión completa de un gen se encuentra entre los límites del otro. Los pares de genes correspondientes a este último caso no se tendrán en consideración para la realización de este estudio.

4.3 | Origen

El origen y evolución de los genes solapados han sido un tema de discusión desde hace muchos años (Kreitmeier et al., 2022). El origen de un nuevo gen dentro de uno ya existente se denomina sobreimpresión (*overprinting* en inglés) (Kreitmeier et al., 2022).

Inicialmente se pensaba que la mayoría de los genes surgían a partir de la duplicación y divergencia de genes más antiguos (Kreitmeier et al., 2022).

Los solapamientos codireccionales pueden surgir por extensiones de los extremos 3' y 5' del gen madre. El extremo 3' requiere una mutación que afecte al codón de STOP, mientras que el extremo 5' únicamente requiere que el gen *downstream* adopte un codón de inicio *upstream* en la misma fase, este último mecanismo es muy simple y fácilmente ocurrente a lo largo de la evolución, contribuyendo al hecho de que los solapamientos codireccionales sean los más frecuentes (Pallejà Caro, 2009).

En el caso de los solapamientos convergentes, ya requiere de un mecanismo más complejo para realizar una extensión en el extremo 3' (Pallejà Caro, 2009).

Finalmente, los solapamientos divergentes requieren que las secuencias Shine–Dalgarno (unos RBSs, *ribosome binding sites* en inglés) se encuentren en la región de solapamiento, causando que su frecuencia sea mucho menor, pero, al igual que pasa con los codireccionales, solo requieren de la generación de un nuevo codón de inicio en la secuencia codificante *upstream* para su formación (Pallejà Caro, 2009).

Se ha visto que el perfil de hidrofobicidad que tiene una secuencia en un marco de lectura desplazado tiende a ser similar que el de la secuencia no desplazada, en los solapamientos ocurridos en una misma cadena (unidireccionales/codireccionales) (Bartonek et al., 2020; Kreitmeier et al., 2022). En cambio, en solapamientos en la pauta de lectura "+1" y "-1" tienden a tener un perfil de hidrofobicidad en sus aminoácidos contrario entre ellos (Kreitmeier et al., 2022). Además, se conservan aminoácidos similares en el marco antisentido "-1" después de una mutación sinónima en el marco de referencia, lo que facilita el mantenimiento de los genes superpuestos (Kreitmeier et al., 2022). El área de investigación en desarrollo de los orígenes de genes superpuestos complementa los hallazgos recientes de muchos genes

taxonómicamente restringidos ("huérfanos") y genes cortos no anotados en procariotas (Kreitmeier et al., 2022).

Otro punto es que se ha observado que los genes solapados pueden tener una mayor conservación durante el curso de la evolución, puesto que las restricciones funcionales intrínsecas que presentan previenen la rotura de la unión de ambos genes (Luo et al., 2006).

4.4 | Interés biotecnológico

Los solapamientos en genomas naturales son un fenómeno complejo, de los que emergen cada vez más casos registrados (Wright et al., 2022). A su vez, en biología sintética, las características funcionales de estos solapamientos son cada vez más importantes, puesto que utilizan material genético ya existente de fuentes diversas para crear nuevas rutas metabólicas, actividades enzimáticas y dispositivos genéticos complejos, entre otras cosas (Wright et al., 2022). Incluso en el campo de la genómica sintética han comenzado a reconstruir genomas enteros desde cero (Wright et al., 2022). En todos estos casos, es importante decidir cómo actuar ante secuencias solapadas.

En los últimos 15 años, varios proyectos de ingeniería genética donde han modificado CDSs solapados han resultado en pérdidas de viabilidad y de eficiencia en el producto final (Wright et al., 2022). Estos proyectos pretendían hacer una refactorización (*refactoring*) del genoma, es decir, reorganizar su arquitectura génica manteniendo su funcionalidad deshaciendo el solapamiento entre genes para que cada uno se encuentre codificado en fragmentos de DNA separados (Wright et al., 2022). El hecho de separar dos genes solapados podía interrumpir la función de elementos reguladores, como promotores, o elementos importantes de la estructura secundaria del RNA, entre otros (Wright et al., 2022).

En otros proyectos actuales donde completaron la refactorización donde se encontraban involucrados CDSs solapados, compensaron las deficiencias funcionales generadas mediante una sintonización empírica de cada sitio de unión al ribosoma (RBSs) y la regulación transcripcional (Wright et al., 2022).

En otros casos, en vez de buscar cómo deshacer un solapamiento, se ha estudiado cómo generarlos para proteger los CDs solapados de la deriva genética (Wright et al.,

2022). Recientemente, también se ha investigado como generar un solapamiento entre el CDS de un gen esencial y el CDS del gen de interés para que este último sea protegido ante posibles mutaciones (Blazejewski et al., 2019).

Aún quedan secretos fundamentales por ser revelados a pesar de los avances en genómica bacteriana de las últimas décadas, como el descubrimiento de muchos más genes solapados, escondidos en las sombras de genes conocidos, de los que se conocen hoy en día (Graf et al., 2023; Kreitmeier et al., 2022)

Por lo tanto, es importante conocer las características de cada posible solapamiento para tener un mejor control sobre un gen concreto que se ve afectado por este fenómeno o para proteger a un gen, no solapado inicialmente, de la influencia de posibles mutaciones o de la deriva genética.

Otro punto de interés es que el solapamiento de genes puede aportar nuevas perspectivas en las relaciones filogenéticas, puesto que presentan una mayor conservación y pueden servir, junto con otros, como marcadores genómicos (Luo et al., 2006).

4.5 | Antecedentes

La idea de este nuevo estudio parte de la tesis doctoral hecha por el Dr. Albert Pallejà (Pallejà Caro, 2009), que en uno de sus capítulos estudió este mismo fenómeno.

En el capítulo 2 de dicha tesis, estudió el fenómeno de solapamientos analizando, bioinformáticamente, 678 genomas. De entre los solapamientos presentes en esos genomas analizados, el 87% resultaron ser codireccionales, un 11% convergentes y un 3% divergentes.

5 | Objetivos

El objetivo principal de este trabajo es realizar un estudio en profundidad del fenómeno de solapamiento de genes en diferentes tipos de genomas bacterianos, clasificando los solapamientos según su topología en: codireccionales, convergentes y divergentes. De esta manera analizar las frecuencias que presentan cada tipo de solapamiento y algunas características asociadas a estos.

Para ello será necesario realizar una preparación previa, mediante programación bioinformática, de los datos existentes sobre cada genoma bacteriano provenientes de una base de datos (RefSeq del NCBI), como también el tratamiento y análisis de los datos resultantes.

6 | Hipótesis

Los solapamientos codireccionales serán más comunes que los solapamientos convergentes y divergentes, tal como prevé la literatura, no obstante, pueden surgir diferencias entre los resultados obtenidos al partir cada estudio de datos diferentes.

7 | Materiales y Métodos

Para la realización de este estudio, se han utilizado los genomas bacterianos que se encuentran agrupados en un listado de genomas procariotas representativos de la base de datos Refseq del NCBI (https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/). Mediante la programación de un código bioinformático, hecho con lenguaje *Python*, se filtraron de tal manera que únicamente se utilizaron los genomas en forma de cromosomas, excluyendo todos los plásmidos.

Los genomas completos en formato GenBank (Gb) se descargaron automáticamente mediante programas propios de Python, utilizando la librería *Biopython* (Chapman & Chang, 2000). Los ficheros de los genomas completos son la fuente principal de información de la que se extrajo información diversa para formar tres archivos csv (*Comma Separated Values* en inglés):

- Un primer archivo identificado como "*ID_genes*", siendo ID el genoma correspondiente, que contenía todos los genes del genoma indexados por filas y una gran cantidad de información referente a cada gen en las columnas como, por ejemplo, las posiciones de inicio y final de cada gen respecto al genoma completo, los identificativos que determinan cada gen (nombre propio y *locus tag*), codones de inicio y STOP, la función del gen y su secuencia, entre otros datos. Estos datos permiten formar el segundo archivo.

- El segundo archivo, identificado como "*ID_overlapping_genes*", siendo ID el genoma correspondiente, es resultado de una filtración de cada gen del primer archivo de tal manera que, si el final del primer gen corresponde a una posición

con un valor numérico mayor a la posición de inicio del segundo gen, determinan la presencia de un solapamiento. Este archivo recoge todos los solapamientos existentes en el correspondiente, además de una gran cantidad de información referente a cada uno de ellos como, por ejemplo, los identificativos de ambos genes implicados y sus respectivas posiciones, la distancia de solapamiento (valor positivo de la distancia entre genes), los respectivos codones de inicio y final para cada gen implicado, el tipo de solapamiento presente en ese par de genes (ver el apartado introductorio 3.2 | Clasificación) y la secuencia del respectivo solapamiento, entre otros datos.

- Y, finalmente, el tercer archivo, identificado como "*ID_info*", siendo ID el genoma correspondiente, que recoge información taxonómica del organismo, además de otros datos referentes al propio genoma, como el tipo de cromosoma (circular o lineal), la medida de este en pares de bases (pb), el contenido de guanina y citosina en porcentaje, el recuento de los codones de inicio y STOP, etc. Estos archivos de información posteriormente se tratan para que resulten en un único archivo que contiene dicha información de todos los genomas tratados de forma que pertenecen a un único *DataFrame* global.

Una vez se han obtenido todos los resultados, se procedió al análisis de resultados mediante herramientas bioinformáticas. Para ello se utilizaron las librerías *Matplotlib* (https://matplotlib.org/) y *Plotly* (https://plotly.com/python/), librerías de *Python* especializadas en la creación de gráficos para analizar los resultados obtenidos previamente y de presentar-los de forma adecuada para su correcta comprensión y visualización.

Todos los programas bioinformáticos utilizados en este estudio han sido creados *de novo*, utilizando el lenguaje de programación *Python*, y las librerías *Biopython* y *Pandas* (https://pandas.pydata.org/).

8 | Resultados y Discusión

Como resultado del análisis de los genomas representativos de la base de datos RefSeq del NCBI, se obtuvieron 5.223 genomas bacterianos en forma de cromosomas, una cantidad exponencialmente mayor que los 678 genomas bacterianos que se analizaron en la tesis doctoral de Pallejà Caro (2009). Esto es

debido al incremento en el número de genomas secuenciados depositados en las bases de datos actuales respecto a los de 2008.

De entre todos los genomas bacterianos obtenidos para este estudio, se identificaron diferentes códigos genéticos. Este dato se encuentra anotado en cada archivo GenBank descargado para cada genoma bajo el término "*transl_table*", que por lo general corresponde al id=1, correspondiente al código genético estándar (Elzanowski & Ostell, 2019). El código genético es el que determina qué aminoácidos o codones de STOP se generarán a partir de los diferentes codones posibles (Lamolle et al., 2023), es decir, qué reglas utiliza para realizar la traducción a proteína (Elzanowski & Ostell, 2019).

No obstante, no todos los organismos están anotados utilizando el código genético estándar, existen diferentes variaciones que difieren según la taxonomía del organismo analizado (Elzanowski & Ostell, 2019).

Tras el análisis se encontró que, de todos los genomas bacterianos recopilados, un total de 5.104 genomas presentan un código genético 11 (transl_table=11). Además, se identificaron 119 genomas que presentan un código genético distinto: 118 que presentan un código genético 4 (transl_table=4) y un único genoma que presenta un código genético 25 (transl_table=25). Este último genoma, debido a su singularidad y diferencia significativa con respecto a los otros patrones identificados, no se considerará en el análisis y resultados posteriores del estudio.

El código genético 11 es comúnmente utilizado por Bacterias, Arqueas, virus procariotas y proteínas de cloroplastos (Elzanowski & Ostell, 2019). En cambio, el código genético 4 es utilizado, dentro del dominio Bacteria, para *Entomoplasmatales* y *Mycoplasmatales*, además de en algunos tipos de hongos, metazoos y otros organismos eucariotas (Elzanowski & Ostell, 2019). Finalmente, el código genético 25, el cual no se considerará, se utiliza en dos grupos de bacterias no cultivadas que se encuentran en el medio ambiente marino y de agua dulce y en los intestinos y las cavidades orales de los mamíferos, entre otros (Elzanowski & Ostell, 2019).

8.1 | Cantidad de solapamientos según el tamaño del genoma

Mediante un diagrama de dispersión, se puede observar la posible relación entre el número de genes solapados y el tamaño del genoma.

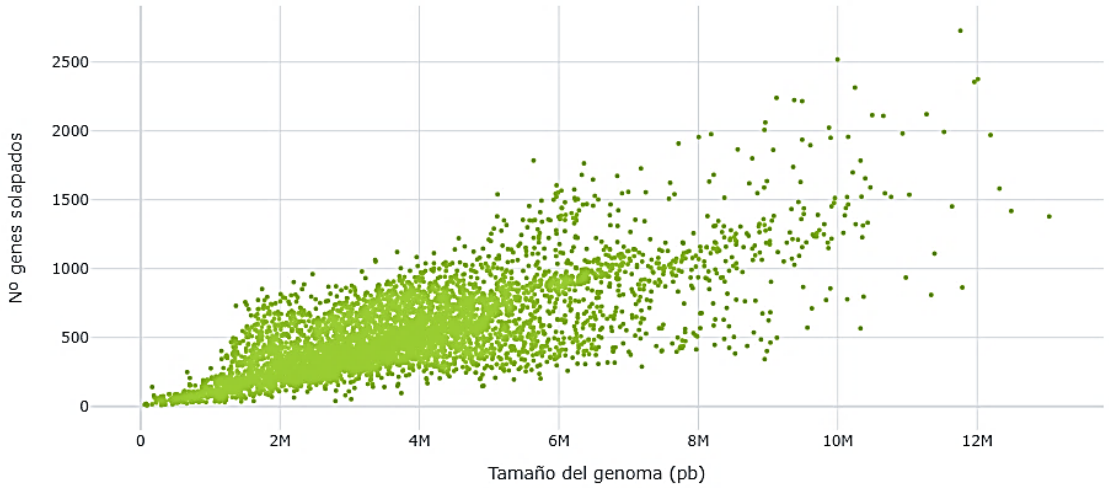

Figura 5 | Diagrama de dispersión que relaciona el tamaño del genoma con la cantidad de genes solapados, con código genético 11, presentes en dicho genoma. (n=5.104)

En este caso particular de los genomas con código genético 11, existe demasiada dispersión para poder considerarse una relación lineal, no obstante, se puede observar, de manera general, una tendencia a que, a medida que el tamaño del genoma aumenta, tiende a haber una mayor cantidad de genes que se solapan entre sí.

Por ello, esta observación sugiere que existe una posible relación entre el tamaño del genoma y la presencia de genes solapados, destacando que, aunque la relación no sea completamente lineal, se puede apreciar una asociación general entre estas dos variables.

Al analizar el diagrama de la Figura S1, correspondiente a la misma relación entre los genomas con código genético 4 (118 genomas), que tienen un tamaño mucho menor (máximo 2M pb) que los genomas con código genético 11, se observa cómo, a pesar de presentar igualmente una dispersión muy amplia, se aprecia igualmente la tendencia a aumentar la cantidad de genomas solapados cuando el tamaño del genoma aumenta.

8.2 | Tipos de solapamientos

Resultado del análisis de cada uno de los 2.722.696 de solapamientos con código genético 11, se obtiene la siguiente distribución de tipos de solapamiento que se muestra en la Figura 6.

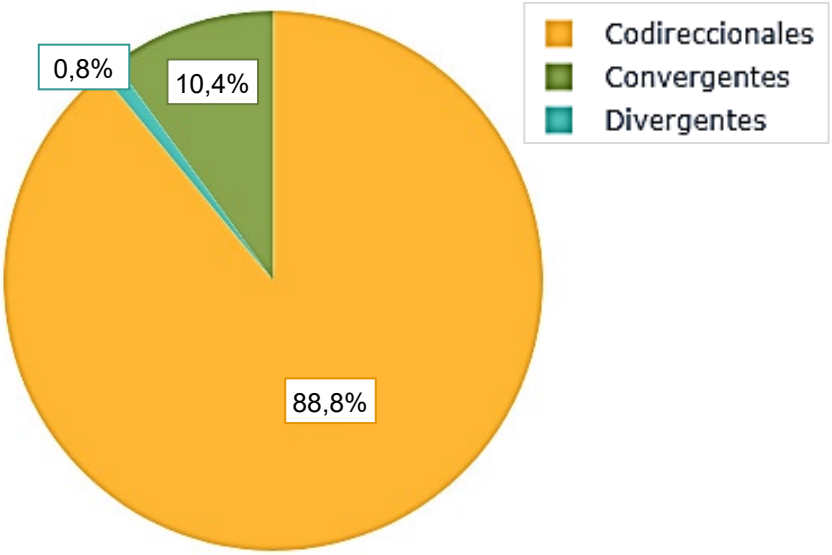

Figura 6 | Gráfico circular representando las proporciones de tipos de solapamientos presentes en los 2.722.696 solapamientos analizados con código genético 11. (n=2.722.696)

Tal como se esperaba, los solapamientos de tipo codireccional representan la inmensa mayoría (aproximadamente un 89%), mientras que los solapamientos divergentes no llegan prácticamente al 1% del total de solapamientos analizado. Estos datos difieren ligeramente de las estimaciones hechas por Wright et al. (2022), puesto que no tenemos en consideración los pares de genes anidados (*nested*) y, a diferencia de ellos, en nuestro caso se ha decidido mantener los genes codificantes para proteínas hipotéticas. A su vez, los resultados obtenidos por Pallejà Caro (2009), aparte de analizar un número menor de genomas, no tenía en cuenta los genes codificantes para proteínas hipotéticas, dando lugar a la diferencia de resultados obtenidos.

Al analizar la distribución de los genomas con código genético 4 que se muestra en la Figura S2, del total de 16.341 solapamientos, se observa como los solapamientos codireccionales representan también la mayoría, pero con un porcentaje mucho

mayor (un 94,3%), los convergentes tienen menor porcentaje (un 5,3%) respecto al 10% de convergentes en los solapamientos de los genomas con código genético 11, y los divergentes se encuentran mucho menos presentes con apenas un 0,4%, correspondientes a 59 casos.

En ambos casos, la poca presencia de solapamientos divergentes puede ser debido a la presencia de secuencias críticas en el extremo 5' de los CDSs que imponen restricciones evolutivas adicionales a la hora de retener estos solapamientos (Wright et al., 2022).

8.3 | Longitud de los solapamientos

Mediante un gráfico de barras se puede observar la distribución de las longitudes de los solapamientos. En este caso, segregadas por tipo de solapamiento.

Como se podrá observar a continuación en cada tipo de solapamiento, a medida que aumenta el tamaño del solapamiento, menor es la frecuencia de número de solapamientos para esas longitudes. Existen evidencias de que algunos solapamientos mayores de 60 pb pueden no ser solapamientos reales, sino que son producto de errores de anotación (Pallejà et al., 2008).

8.3.1 | Solapamientos codireccionales

Como se puede observar en la Figura 7 y en la Tabla 1, referente a las frecuencias de longitud del solapamiento para los genomas con código genético 11, los solapamientos más frecuentes son los de 1 y, sobre todo, los de 4 nucleótidos, ambos en fase 2, con 390 mil y 1,3 millones de solapamientos con dicha longitud respectivamente.

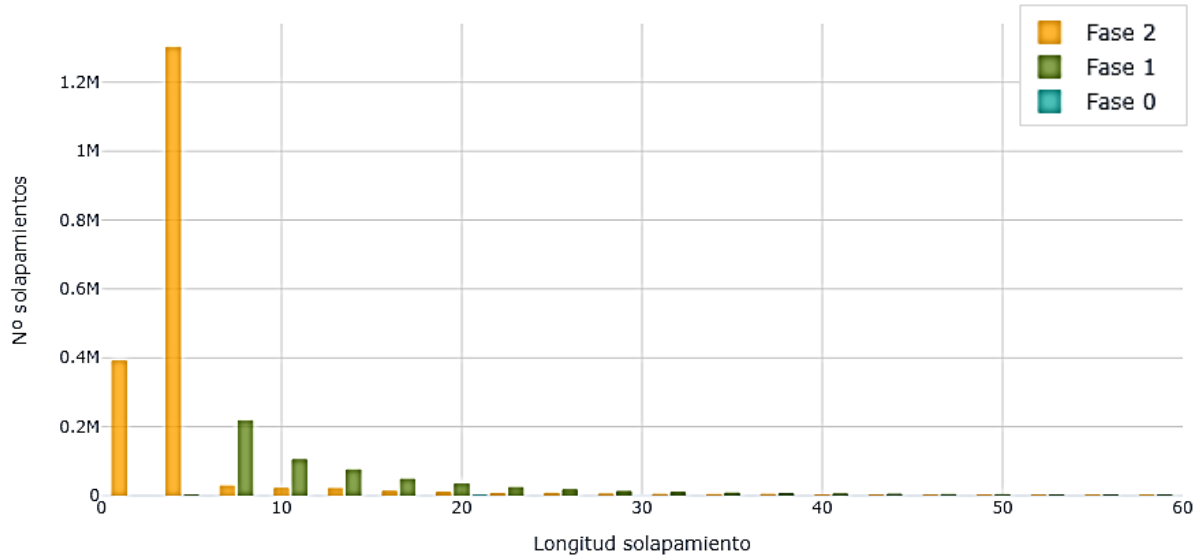

Figura 7 | Relación entre la longitud de los solapamientos y su reiteración en los solapamientos **codireccionales**, con código genético 11, analizados. (n=2.418.230)

Tabla 1 | Recuento del número de solapamientos **codireccionales** por cada longitud de solapamiento en los genomas con código genético 11. Las dos longitudes más frecuentes se encuentran resaltadas en naranja. Las longitudes correspondientes a la Fase 0 se encuentran en rojo.

Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos
1	390.850	44	3.901	83	592
4	1.301.290	46	1.652	85	437
5	558	47	2.966	86	558
7	27.851	49	1.707	88	463
8	217.330	50	2.676	89	497
10	21.147	52	1.316	91	431
11	104.200	53	2.229	92	440
13	20.361	55	1.145	94	402
14	74.717	56	1.935	95	417
16	12.851	58	898	97	440
17	47.573	59	1.643	98	360
19	9.609	61	799	100	379
20	34.207	62	1.356	101	389
21	1	64	784	103	324
22	6.329	65	1.178	104	312
23	23.411	66	1	106	323
25	6.681	67	759	107	302
26	17.004	68	1.242	109	294
28	4.427	70	667	110	281

29	12.227	71	1.027	112	280
31	3.551	72	1	113	323
32	9.578	73	635	114	1
34	3.018	74	838	115	350
35	7.302	76	551	116	318
37	3.469	77	765	118	286
38	6.015	78	1	119	257
40	2.120	79	553	140	1
41	4.776	80	676		
43	1.910	82	509		

Considerando como codones básicos: el codón de inicio ATG y los alternativos más frecuentes, GTG y TTG, junto con los codones de STOP: TAA, TAG y TGA; existen longitudes de solapamiento no permitidas.

En el caso de los solapamientos codireccionales, las longitudes de 2, 3 y 5 nucleótidos no están permitidas, puesto que:

- **Solapamientos de 2 nucleótidos** (Fase 1): Ninguno de los 3 codones de inicio presentan en sus 2 primeros nucleótidos igualdad con los 2 últimos nucleótidos de alguno de los 3 codones de STOP (Figura 8A).
- **Solapamientos de 3 nucleótidos** (Fase 0): Ninguno de los 3 codones de inicio presentan igualdad con alguno los 3 codones de STOP. Al encontrarse ambos genes en la misma pauta de lectura, da a entender que el segundo gen estará contenido en el primero y, por lo tanto, ningún solapamiento en fase 0 es posible (Figura 8B).
- **Solapamientos de 5 nucleótidos** (Fase 1): Ninguno de los 3 codones de inicio presentan igualdad en su último nucleótido con el primer nucleótido de alguno de los 3 codones de STOP (Figura 8C).

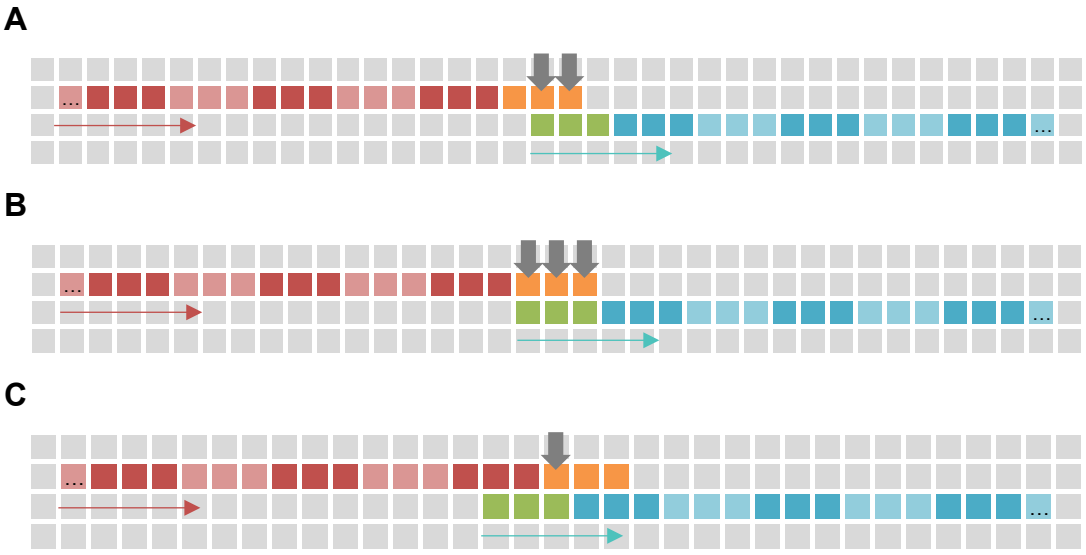

Figura 8 | Esquema del solapamiento teórico **codireccional** de: (A) 2 nucleótidos, (B) 3 nucleótidos y (C) 5 nucleótidos. Identificado de tal manera que el naranja corresponde al codón de STOP del gen 1, el verde al codón de inicio del gen 2 y las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente) en la misma cadena. Las flechas grises indican los nucleótidos solapados comunes en el codón de inicio y en el de STOP.

Sin embargo, si también se tienen en consideración los siguientes codones puntuales: CTG, ATC, ATT y ATA, de los cuales está descrito que actúan como codones de inicio en situaciones puntuales (Elzanowski & Ostell, 2019), los cuales se pueden observar en la Figura 9, los solapamientos no permitidos de 5 nucleótidos sí pueden ser posibles, puesto que el codón de inicio puntual ATT, que al presentar una timina como último nucleótido, permite que los codones de STOP puedan ser TAA, TAG y TGA. Esto lo demuestra la presencia de casos de solapamiento con longitudes no permitidas inicialmente, como la de 5 nucleótidos (558 casos).

En el caso de los solapamientos codireccionales en fase 0 que aparecen, tal como se pueden observar resaltados en rojo en la Tabla 1, no están permitidos, puesto que al encontrarse en fase 0, ambos genes “solapados” codifican para el mismo fragmento de ADN y ambos se verán afectados por el codón de STOP del otro, resultando en un gen dentro de otro. En estos casos, debe haber algún problema a la hora de interpretar los límites de los genes o en su anotación.

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu	TCA S Ser	TAA * Ter	TGA * Ter
TTG L Leu i	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu i	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile i	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile i	ACC T Thr	AAC N Asn	AGC S Ser
ATA I Ile i	ACA T Thr	AAA K Lys	AGA R Arg
ATG M Met i	ACG T Thr	AAG K Lys	AGG R Arg
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val i	GCG A Ala	GAG E Glu	GGG G Gly

Figura 9 | Equivalencias del **código genético 11**. Marcados en verde los codones de inicio más frecuentes y en naranja los puntuales. Marcados en rojo los codones de STOP. Figura adaptada de Elzanowski & Ostell (2019).

De los solapamientos codireccionales con longitud de 1 y de 4 nucleótidos, las secuencias más frecuentes son:

- "A" con 377.646 casos de los 390.850 totales con 1 nucleótido de longitud (Tabla 1). Esta adenina procede del último nucleótido de los codones de STOP TAA o TGA y del primer nucleótido del codón de inicio ATG. Puesto que estos codones pertenecen al grupo de los más frecuentes, explica el porqué es uno de los solapamientos más habituales.

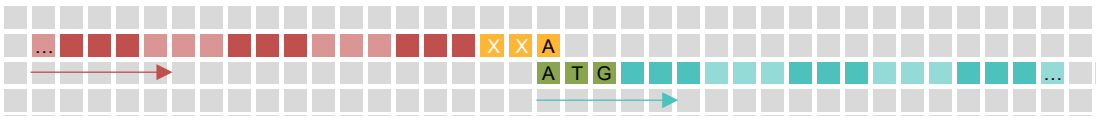

Figura 10 | Esquema del **solapamiento de 1 nucleótido codireccional** más frecuente. Identificado de tal manera que el naranja corresponde al codón de STOP del gen 1, el verde al codón de inicio del gen 2 y las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente) en la misma cadena.

- "ATGA" con 977.259 casos de los 1.301.290 totales con 4 nucleótidos de longitud (Tabla 1). Esta secuencia proviene del codón de STOP TGA (los 3 últimos nucleótidos de la secuencia) y del codón de inicio ATG (los 3 primeros

nucleótidos de la secuencia). Al igual que en el caso anterior, el codón de inicio ATG es el más frecuente, como también es de los más frecuentes el codón de STOP TGA, dando explicación a la alta frecuencia de casos con este solapamiento.

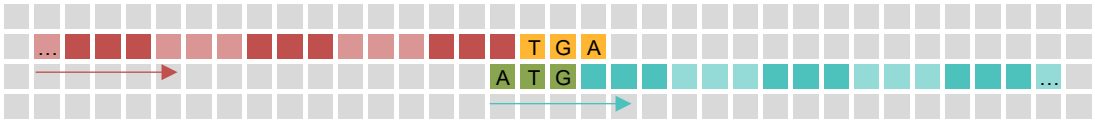

Figura 11 | Esquema del **solapamiento de 4 nucleótidos codireccional** más frecuente. Identificado de tal manera que el naranja corresponde al codón de STOP TGA del gen 1, el verde al codón de inicio ATG del gen 2 y las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente) en la misma cadena.

En relación con los genomas con código genético 4, tal como se puede observar en la Figura S3, los solapamientos más frecuentes son los de 1 nucleótido, mientras que de 4 nucleótidos hay muy poca presencia (únicamente 89 solapamientos). En este caso, no aparece ningún solapamiento no permitido en fase 0.

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu i	TCA S Ser	TAA * Ter	TGA W Trp
TTG L Leu i	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu i	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile i	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile i	ACC T Thr	AAC N Asn	AGC S Ser
ATA I Ile i	ACA T Thr	AAA K Lys	AGA R Arg
ATG M Met i	ACG T Thr	AAG K Lys	AGG R Arg
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val i	GCG A Ala	GAG E Glu	GGG G Gly

Figura 12 | Equivalencias del **código genético 4**. Marcados en verde los codones de inicio más frecuentes y en naranja los puntuales. Marcados en rojo los codones de STOP. Figura adaptada de Elzanowski & Ostell (2019).

En este caso, de los solapamientos codireccionales con longitud de 1 y de 4 nucleótidos, las secuencias más frecuentes son:

- "A" con 5.197 casos de los 5.408 totales con 1 nucleótido de longitud (Tabla S1). Esta adenina procede del último nucleótido del codón de STOP TAA y del primer nucleótido del codón de inicio ATG. Como se puede observar en la Figura 12, en el código genético 4, el codón TGA no es un codón de STOP, sino que codifica para el triptófano. Este caso sigue el mismo patrón mostrado en el esquema de la Figura 10.
- "TTAA" con 69 casos de los 89 totales con 4 nucleótidos de longitud (Tabla S1). Esta secuencia proviene del codón de STOP TAA (los 3 últimos nucleótidos de la secuencia) y del codón de inicio TTA (los 3 primeros nucleótidos de la secuencia). En este caso, hay muchos menos casos de solapamientos de 4 nucleótidos debido a que TGA no es un codón de STOP.

8.3.2 | Solapamientos convergentes

En el caso de los solapamientos convergentes, puesto que los genes que lo componen son transcritos de cadenas diferentes, las longitudes de solapamiento no permitidas son diferentes, ya que los codones del segundo gen deben ser complementarios a los de la cadena principal. Además, al encontrarse ambos genes en cadenas diferentes, todas las fases son posibles.

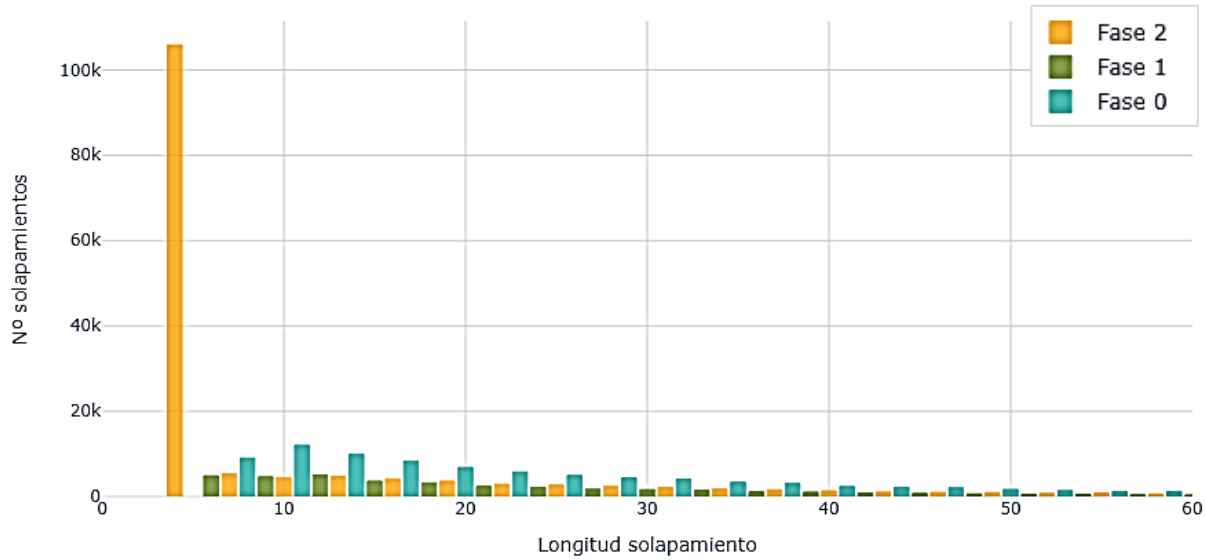

Figura 13 | Relación entre la longitud de los solapamientos y su reiteración en los solapamientos **convergentes**, con código genético 11, analizados. (n=283.752)

Tabla 2 | Recuento del número de solapamientos **convergentes** por cada longitud de solapamiento en los genomas con código genético 11. La longitud más frecuente se encuentra resaltada en naranja.

Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos
4	105.851	44	2.127	83	432
6	4.831	45	727	84	228
7	5.275	46	947	85	270
8	9.003	47	2.022	86	327
9	4.602	48	609	87	167
10	4.452	49	919	88	222
11	12.002	50	1.643	89	315
12	5.011	51	522	90	132
13	4.737	52	716	91	182
14	9.955	53	1.362	92	286
15	3.587	54	498	93	144
16	4.134	55	798	94	193
17	8.301	56	1.199	95	237
18	3.191	57	400	96	104
19	3.597	58	567	97	237
20	6.797	59	1.164	98	265
21	2.384	60	352	99	118
22	2.876	61	608	100	177
23	5.738	62	848	101	203
24	2.160	63	306	102	117
25	2.672	64	540	103	157
26	4.928	65	782	104	178
27	1.705	66	273	105	92
28	2.345	67	423	106	135
29	4.353	68	758	107	157
30	1.612	69	262	108	95
31	2.143	70	385	109	104
32	4.061	71	702	110	175
33	1.489	72	296	111	71
34	1.727	73	329	112	95
35	3.339	74	579	113	225
36	1.101	75	258	114	75
37	1.577	76	270	115	89
38	3.025	77	565	116	100
39	1.014	78	168	117	123
40	1.318	79	310	118	114
41	2.334	80	466	119	101
42	861	81	193	120	64

43	1.010	82	254	164	1

Como se puede observar en la Figura 13 y en la Tabla 2, la longitud de solapamiento más frecuente es la de 4 nucleótidos, con más de 105 mil solapamientos.

De esos 105.851 solapamientos convergentes de 4 nucleótidos de longitud (Tabla 2), la mayoría (41.295 solapamientos) presenta la secuencia solapada “CTAG” (de la cadena +1). Esta está formada por el codón de STOP TAG, de los más frecuentes, en ambos genes, pero teniendo en cuenta que el segundo gen se encuentra en la cadena complementaria.

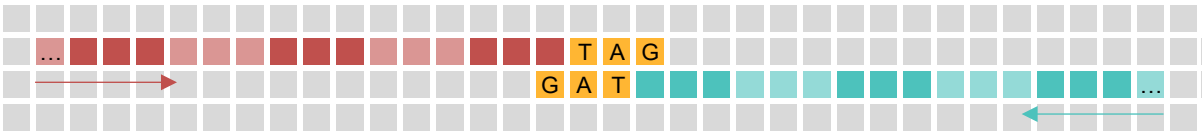

Figura 14 | Esquema del **solapamiento de 4 nucleótidos convergente** más frecuente. Identificado de tal manera que el naranja corresponde al codón de STOP de ambos genes y las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente) de diferentes cadenas.

Considerando los mismos codones de inicio y de STOP básicos del apartado de codireccionales, en el caso de los solapamientos convergentes, las longitudes de 1, 2, 3 y 5 nucleótidos no están permitidas, puesto que:

- **Solapamientos de 1 nucleótido** (Fase 2): El nucleótido complementario del último nucleótido de todos los codones de STOP no puede ser el último nucleótido de un codón de STOP (Figura 15A).
- **Solapamientos de 2 nucleótidos** (Fase 1): Ninguno de los nucleótidos complementarios de los últimos 2 nucleótidos de los codones de STOP coinciden con los 2 últimos nucleótidos de un codón de STOP (Figura 15B).
- **Solapamientos de 3 nucleótidos** (Fase 0): El codón complementario reverso de todos los codones de STOP no es un codón de STOP (Figura 15C).
- **Solapamientos de 5 nucleótidos** (Fase 1): El nucleótido complementario del primer nucleótido de todos los codones de STOP no se encuentra nunca como el primer nucleótido de un codón de STOP (Figura 15D).

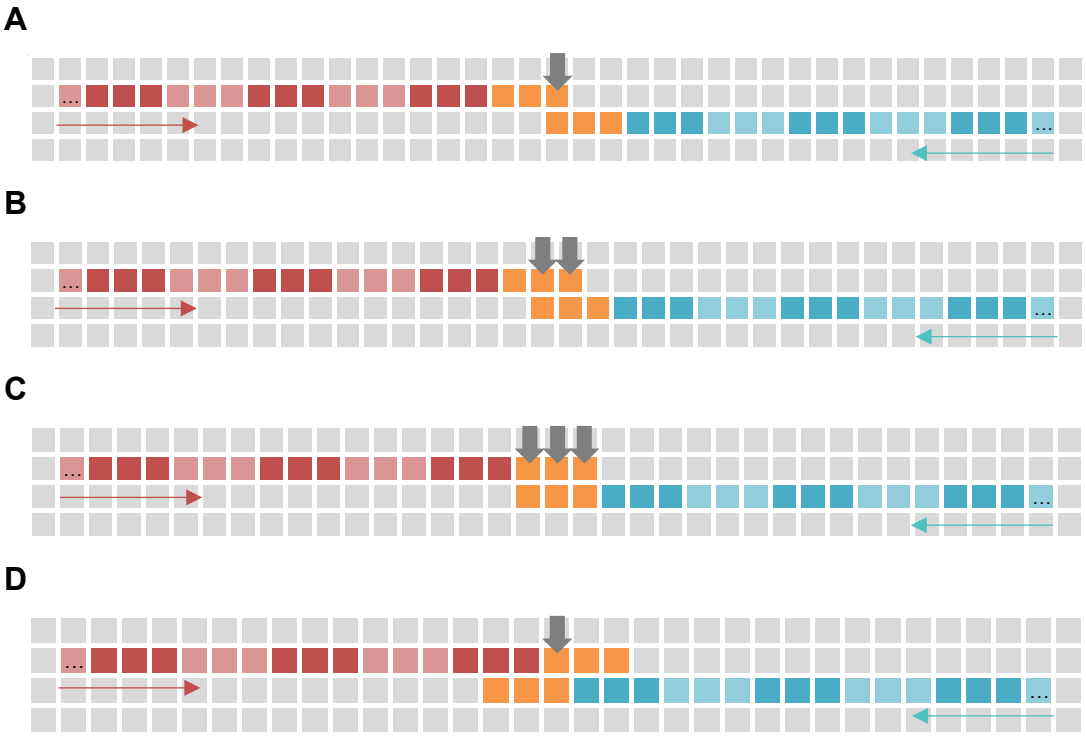

Figura 15 | Esquema del solapamiento teórico **convergente** de: (A) 1 nucleótido, (B) 2 nucleótidos, (C) 3 nucleótidos y (D) 5 nucleótidos. Identificado de tal manera que el naranja corresponde al codón de STOP de ambos genes y las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente) de diferentes cadenas. Las flechas grises indican los nucleótidos solapados de los codones STOP.

Puesto que no tiene en consideración los codones de inicio para el solapamiento, estas posiciones no permitidas siguen sin ser posibles aún considerando los codones puntuales.

En el caso de los solapamientos convergentes en los genomas con código genético 4, representados en la Figura S4, la longitud más frecuente sigue siendo la de 4 nucleótidos, en este caso con 340 casos del total de 862 solapamientos convergentes presentes.

De estos, la secuencia más frecuente, con 194 casos de los 340 (Tabla S2), es “TTAA”. Esta presenta el codón de STOP TAA en ambos genes (en cadenas diferentes). La segunda más frecuente es “TTAG”, con 70 casos. En este caso, el gen 1 presenta el codón de STOP TAG y el gen 2 el codón de STOP TAA.

Además, se puede observar cómo los solapamientos en fase 2, a excepción del de 4 nucleótidos de longitud, es muy poco frecuente.

8.3.3 | Solapamientos divergentes

En este caso, y al igual que en el caso de los solapamientos convergentes, al tratarse de genes localizados en cadenas diferentes, todas las fases son posibles.

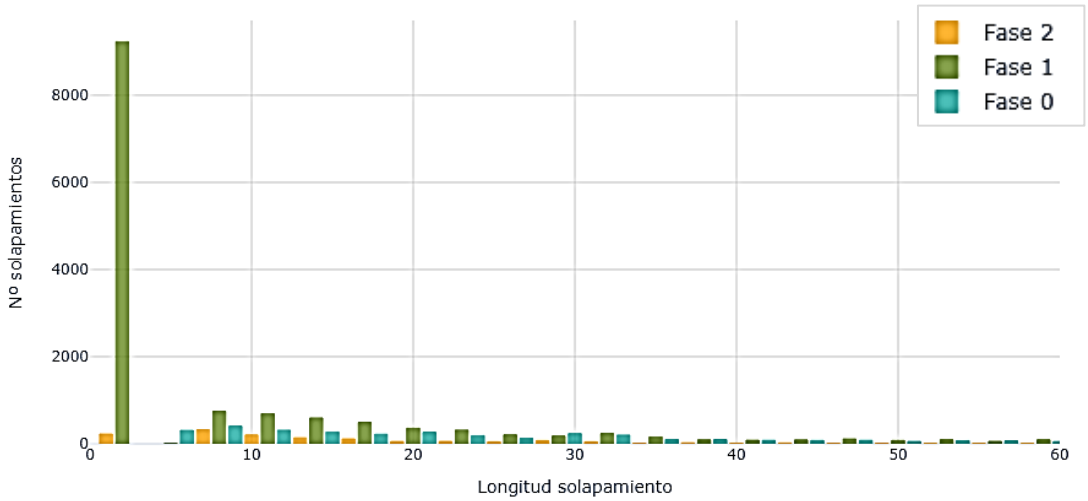

Figura 16 | Relación entre la longitud de los solapamientos y su reiteración en los solapamientos **divergentes**, con código genético 11, analizados. (n=20.714)

Tabla 3 | Recuento del número de solapamientos **divergentes** por cada longitud de solapamiento en los genomas con código genético 11. La longitud más frecuente se encuentra resaltada en verde.

Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos
1	226	43	21	83	45
2	9.231	44	91	84	43
5	7	45	78	85	12
6	309	46	11	86	40
7	324	47	113	87	51
8	747	48	81	88	3
9	407	49	14	89	42
10	205	50	78	90	47
11	693	51	55	91	5
12	312	52	7	92	38
13	139	53	95	93	51
14	594	54	75	94	6
15	273	55	9	95	33
16	114	56	57	96	49

17	493	57	70	97	4
18	221	58	7	98	35
19	54	59	94	99	48
20	359	60	52	100	2
21	267	61	11	101	38
22	56	62	126	102	54
23	315	63	67	103	6
24	189	64	10	104	33
25	48	65	75	105	52
26	207	66	69	106	3
27	131	67	5	107	37
28	77	68	84	108	40
29	180	69	57	109	1
30	235	70	6	110	33
31	48	71	58	111	43
32	237	72	67	112	4
33	200	73	7	113	40
34	15	74	50	114	56
35	160	75	71	115	3
36	101	76	5	116	37
37	24	77	30	117	36
38	92	78	50	118	3
39	98	79	9	119	45
40	17	80	49	120	41
41	84	81	66		
42	81	82	5		

Tal como se aprecia en la Figura 16 y en la Tabla 3, la longitud de solapamiento más frecuente, con diferencia, es la de 2 nucleótidos, con más de 9.200 solapamientos con esta longitud.

Figura 17 | Esquema del **solapamiento de 2 nucleótidos divergente** más frecuente. Identificado de tal manera que el verde corresponde los codones de inicio ATX de ambos genes y las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente) de diferentes cadenas.

En este caso, la totalidad de solapamientos con una longitud de 2 nucleótidos presenta la secuencia “AT” (de la cadena +1, correspondiente al gen 2) (9.231 solapamientos) (Tabla 3). Puede observarse un esquema de este solapamiento en la Figura 17 Que este solapamiento sea tan frecuente tiene sentido, puesto que la totalidad de codones que presentan esta secuencia AT en las dos primeras posiciones del codón, pueden codificar para codones de inicio (ATT, ATC, ATA, ATG), tanto en el código genético 11 como en el código genético 4.

Considerando los mismos codones de inicio y de STOP básicos del apartado de codireccionales, en el caso de los solapamientos divergentes, las longitudes de 3, 4 y 5 nucleótidos no están permitidas, puesto que:

- **Solapamientos de 3 nucleótidos** (Fase 0): El codón complementario reverso de todos los codones de inicio no es nunca un codón de inicio (Figura 18A).
- **Solapamientos de 4 nucleótidos** (Fase 2): Los nucleótidos complementarios de los primeros 2 nucleótidos de los codones de inicio no pueden ser los 2 primeros nucleótidos de un codón de inicio (Figura 18B).
- **Solapamientos de 5 nucleótidos** (Fase 1): El nucleótido complementario del último nucleótido de un codón de inicio no puede ser el último nucleótido de un codón de inicio (Figura 18C).

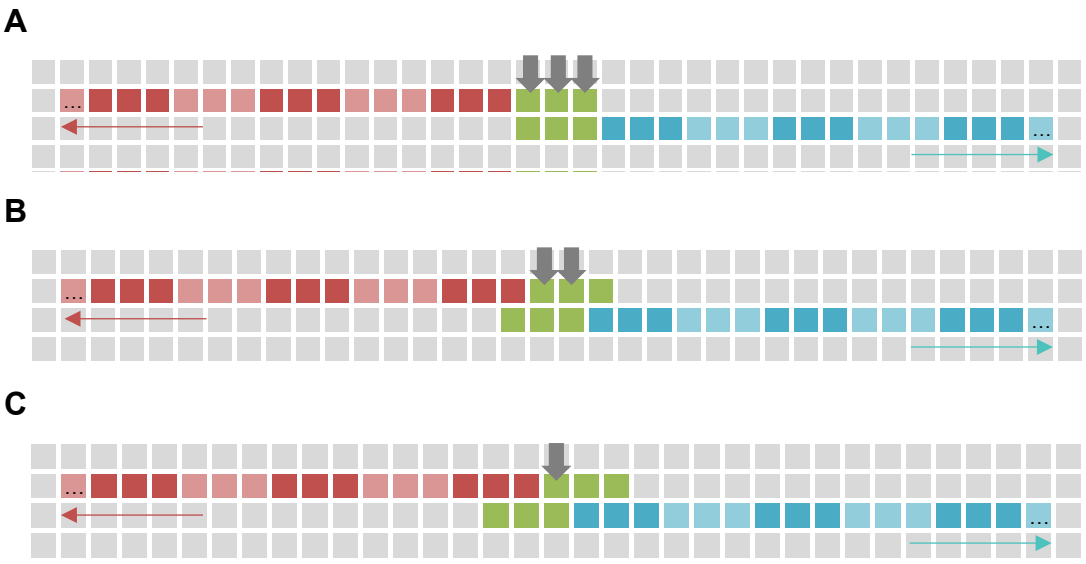

Figura 18 | Esquema del solapamiento teórico **divergente** de: (A) 3 nucleótidos, (B) 4 nucleótidos y (C) 5 nucleótidos. Identificado de tal manera que el verde corresponde al codón de inicio de ambos genes y las zonas de rojo y azul a los genes solapados (1 y 2 respectivamente) de diferentes cadenas. Las flechas grises indican los nucleótidos solapados de los codones de inicio.

En este caso, se observa la presencia de 7 solapamientos con longitud de 5 nucleótidos, lo que puede ser explicado si también se tiene en consideración que los codones CTG, ATC, ATT y ATA, los cuales se pueden observar en la anterior Figura 9 marcados en naranja, en algunos casos pueden funcionar como codones de inicio, al igual que pasa en el caso de los solapamientos codireccionales.

Al tomar en consideración esos codones puntuales, las posiciones de 4 nucleótidos solapados y de 5 nucleótidos solapados sí están permitidas, puesto que: el codón de inicio puntual ATA, al tener TA como sus dos últimos nucleótidos, permite que el codón de inicio de la cadena complementaria pueda ser ATA también, permitiendo la formación de un solapamiento de 4 nucleótidos; y todos los codones de inicio pasan a presentar un nucleótido en su ultima posición que es complementario al ultimo nucleótido de algún otro codón de inicio, permitiendo los solapamientos de 5 nucleótidos.

En el caso de los solapamientos divergentes del código genético 4, tal y como se puede observar en la Figura S5, hay muy pocos casos de solapamiento de este tipo (59), además de que ninguna longitud destaca muy por encima del resto, más que la

de 1 nucleótido (con 6 casos) (Tabla S3), y las de 11, 14 y 17 nucleótidos con 5 casos cada una.

9 | Conclusiones

A lo largo de este estudio, se han analizado una gran cantidad de genomas bacterianos, en total 5.223 genomas, de los cuales se han obtenido 2.739.037 solapamientos (2.722.696 con código genético 11 y 16.341 con código genético 4).

En todos los casos, se puede observar cómo, a medida que aumenta la longitud de los solapamientos, menos casos existen.

No parece haber diferencias muy notorias entre los resultados obtenidos con los genomas con código genético 11 y los genomas con código genético 4 en lo que respecta a frecuencias: Los solapamientos con código genético 11 presentan una proporción en sus solapamientos: 88,8% codireccional, 10,4% convergente y 0,8% divergente; Mientras que en el código genético 4: 94,3% codireccional, 5,3% convergente y 0,4% divergente.

Para los codireccionales, los solapamientos más frecuentes han sido los de 1 y 4 nucleótidos, mientras que los de 2 y 3 nucleótidos y los que se encuentran en Fase 0 no están permitidos. En el caso de los solapamientos convergentes, los más frecuentes han sido los de 4 nucleótidos, mientras que los de 1, 2, 3 y 5 nucleótidos no están permitidos. Y en el caso de los divergentes, la mayor frecuencia la han presentado los solapamientos de 2 nucleótidos, mientras que los de 3 nucleótidos no están permitidos.

No obstante, sí que hay diferencias significativas entre los códigos genéticos 11 y 4 en la escala, tanto de cantidad de genomas, como de tamaños de estos, esto último puesto que ningún genoma con código genético 4 supera los 2M de pb, mientras que en algún caso de los de código genético 11 llegan incluso a medir 13M pb.

Otra diferencia importante entre los solapamientos de ambos códigos genéticos es la drástica reducción en la cantidad de solapamientos de 4 nucleótidos en los solapamientos codireccionales con código genético 4, debido a que el codón TGA no codifica por un codón de STOP en este código genético.

A excepción de los casos de solapamientos en fase 0 presentes en los resultados del análisis de los solapamientos codireccionales, los resultados obtenidos presentan ligeras diferencias respecto a las previsiones hechas por la literatura y los antiguos resultados de Pallejà Caro (2009), por lo tanto, hay que tener en consideración los datos de partida y bajo qué filtros se analizan, ya que cada estudio tomó en consideración diferentes aspectos para decidir cómo realizar su análisis.

10 | Bibliografía

Bartonek, L., Braun, D., & Zagrovic, B. (2020). Frameshifting preserves key physicochemical properties of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(11), 5907-5912. https://doi.org/10.1073/PNAS.1911203117

Blazejewski, T., Ho, H. I., & Wang, H. H. (2019). Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science (New York, N.Y.)*, *365*(6453), 595-598. https://doi.org/10.1126/SCIENCE.AAV5477

Chapman, B., & Chang, J. (2000). Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter*, *20*(2), 15-19. https://doi.org/10.1145/360262.360268

Elzanowski, A., & Ostell, J. (2019, enero 7). *The Genetic Codes*. National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Graf, F., Zehentner, B., Fellner, L., Scherer, S., & Neuhaus, K. (2023). Three Novel Antisense Overlapping Genes in E. coli O157:H7 EDL933. *Microbiology spectrum*, *11*(1). https://doi.org/10.1128/SPECTRUM.02351-22

Huber, M., Faure, G., Laass, S., Kolbe, E., Seitz, K., Wehrheim, C., Wolf, Y. I., Koonin, E. V., & Soppa, J. (2019). Translational coupling via termination-reinitiation in archaea and bacteria. *Nature communications*, *10*(1). https://doi.org/10.1038/S41467-019-11999-9

Huvet, M., & Stumpf, M. P. H. (2014). Overlapping genes: a window on gene evolvability. *BMC genomics*, *15*(1). https://doi.org/10.1186/1471-2164-15-721

Kreitmeier, M., Ardern, Z., Abele, M., Ludwig, C., Scherer, S., & Neuhaus, K. (2022). Spotlight on alternative frame coding: Two long overlapping genes in Pseudomonas aeruginosa are translated and under purifying selection. *iScience*, *25*(2). https://doi.org/10.1016/J.ISCI.2022.103844

Lamolle, G., Simón, D., Iriarte, A., & Musto, H. (2023). Main Factors Shaping Amino Acid Usage Across Evolution. *Journal of molecular evolution*. https://doi.org/10.1007/S00239-023-10120-5

Luo, Y., Fu, C., Zhang, D. Y., & Lin, K. (2006). Overlapping genes as rare genomic markers: the phylogeny of gamma-Proteobacteria as a case study. *Trends in genetics : TIG*, *22*(11), 593-596. https://doi.org/10.1016/J.TIG.2006.08.011

Pallejà, A., Harrington, E. D., & Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC genomics*, *9*. https://doi.org/10.1186/1471-2164-9-335

Pallejà Caro, A. (2009). *Computational insights into intergenic regions and overlapping genes among prokaryote genomes*. Universidad Rovira i Virgili.

Wichmann, S., Scherer, S., & Ardern, Z. (2021). Biological factors in the synthetic construction of overlapping genes. *BMC genomics*, *22*(1). https://doi.org/10.1186/S12864-021-08181-1

Wright, B. W., Molloy, M. P., & Jaschke, P. R. (2022). Overlapping genes in natural and engineered genomes. *Nature reviews. Genetics*, *23*(3), 154-168. https://doi.org/10.1038/S41576-021-00417-W

11 | Autoevaluación

La realización de este estudio me ha permitido ampliar mis conocimientos en el lenguaje Python y diversas librerías de gran utilidad en el campo de la bioinformática. Inicialmente partía con conocimientos básicos sobre programación con Python, fruto de la asignatura de Bioinformática impartida en el grado de Biotecnología, no obstante, esos conocimientos eran insuficientes para realizar este estudio en su totalidad, por lo que era necesario actualizarse y enriquecerse en el tema.

A su vez, la utilización de principios de genética y biotecnología me ha permitido revisar y refrescar conceptos adquiridos de diferentes asignaturas cursadas durante mi carrera universitaria.

Este trabajo de final de grado y la realización del proyecto no habrían sido posibles sin la colaboración del Dr. Santiago Garcia-Vallvé, profesor de algunas asignaturas del grado y mi tutor profesional en la realización de mis prácticas curriculares y extracurriculares en la Universidad Rovira i Virgili, quien me propuso el tema inicialmente y me orientó a lo largo de todo el proceso. Su experiencia y conocimientos en el campo de la genética y la bioinformática han sido fundamentales para el exitoso desarrollo de este proyecto.

Además, quiero agradecer a mi tutor académico, el Dr. Gerard Pujadas Anguiano, por su asesoramiento en la realización de este trabajo de final de grado.

12 | Anexos

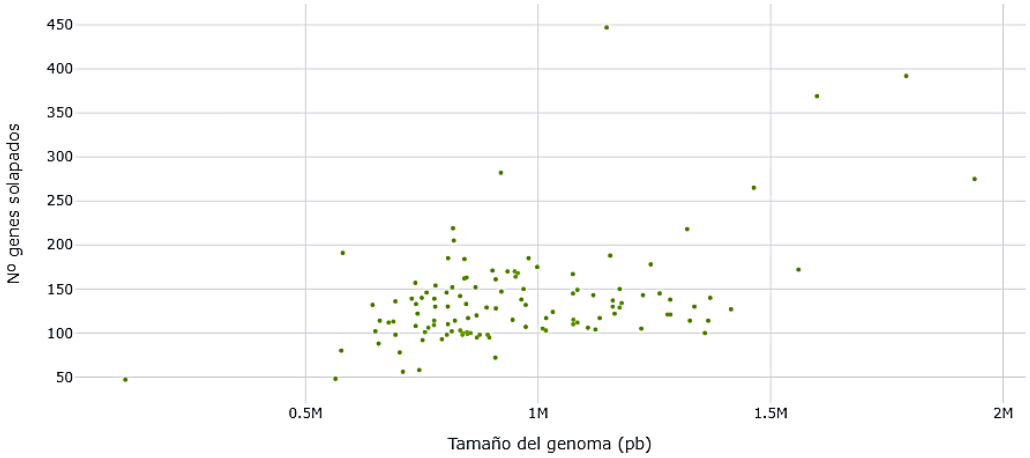

Figura S1 | Diagrama de dispersión que relaciona la medida del genoma con la cantidad de genes solapados, con código genético 4, presentes en dicho genoma. (n= 118)

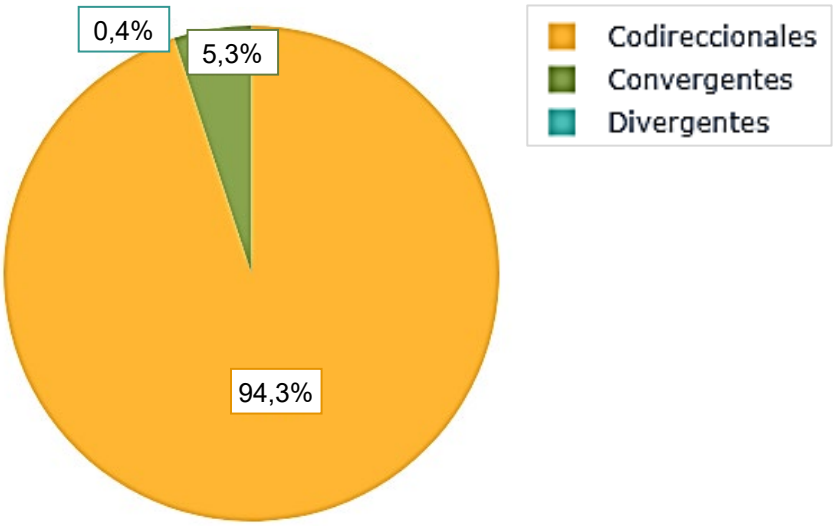

Figura S2 | Gráfico circular representando las proporciones de tipos de solapamientos presen-tes en los solapamientos analizados con código genético 4. (n=16.341)

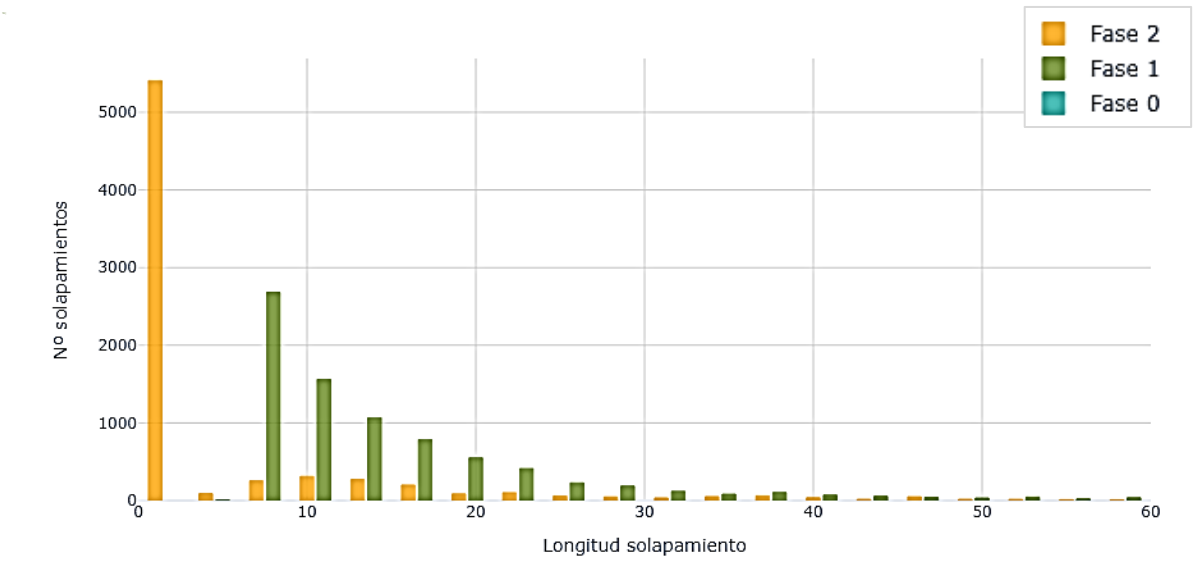

Figura S3 | Relación entre la longitud de los solapamientos y su reiteración en los solapamientos **codireccionales**, con código genético 4, analizados. (n=15.420)

Tabla S1 | Recuento del número de solapamientos **codireccionales** por cada longitud de solapamiento en los genomas con código genético 4.

Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos	Longitud solapamiento	Nº solapamientos
1	5.408	40	37	77	9
4	89	41	71	79	8
5	5	43	16	80	4
7	254	44	56	82	6
8	2.681	46	49	83	10
10	311	47	43	85	1
11	1.561	49	16	86	7
13	271	50	33	88	4
14	1.067	52	15	91	1
16	200	53	45	92	10
17	786	55	11	94	3
19	86	56	21	95	2
20	550	58	10	97	1
22	101	59	37	98	2
23	412	61	21	100	2
25	60	62	18	101	2
26	228	64	8	103	1
28	47	65	13	104	2
29	187	67	7	107	2

31	34	68	8	109	1
32	119	70	8	112	2
34	51	71	5	113	3
35	83	73	16	116	3
37	61	74	6	119	3
38	104	76	5		

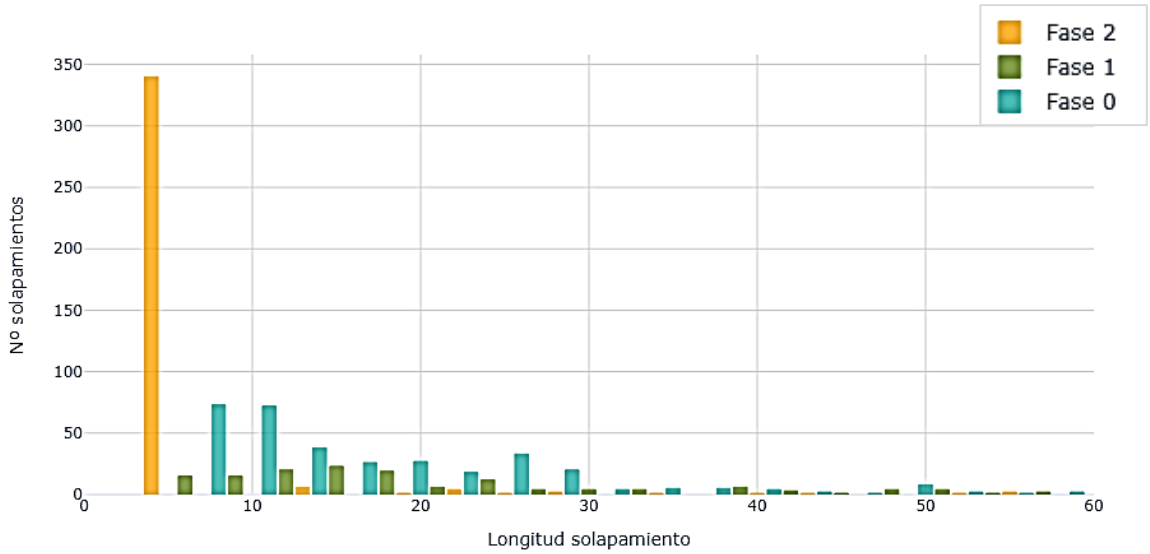

Figura S4 | Relación entre la longitud de los solapamientos y su reiteración en los solapamientos **convergentes**, con código genético 4, analizados. (n=862)

Tabla S2 | Recuento del número de solapamientos **convergentes** por cada longitud de solapamiento en los genomas con código genético 4.

Longitud sola-pamiento	Nº solapa-mientos	Longitud sola-pamiento	Nº solapa-mientos	Longitud sola-pamiento	Nº solapa-mientos
4	340	27	4	51	4
6	15	28	2	52	1
8	73	29	20	53	2
9	15	30	4	54	1
11	72	32	4	55	2
12	20	33	4	56	1
13	6	34	1	57	2
14	38	35	5	59	2
15	23	38	5	62	1
17	26	39	6	65	2
18	19	40	1	66	1
19	1	41	4	69	1
20	27	42	3	70	1

21	6	43	1	71	2
22	4	44	2	81	1
23	18	45	1	83	2
24	12	47	1	86	3
25	1	48	4	101	3
26	33	50	8	115	1

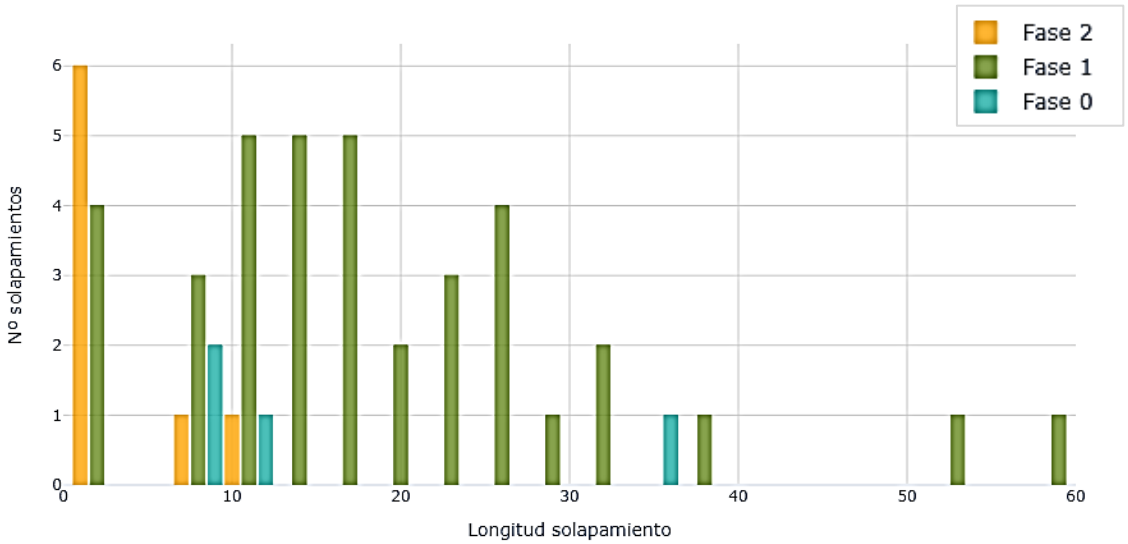

Figura S5 | Relación entre la longitud de los solapamientos y su reiteración en los solapamientos **divergentes**, con código genético 4, analizados. (n=59)

Tabla S3 | cuento del número de solapamientos **divergentes** por cada longitud de solapamiento en los genomas con código genético 4.

Longitud sola-pamiento	Nº solapa-mientos	Longitud sola-pamiento	Nº solapa-mientos	Longitud sola-pamiento	Nº solapa-mientos
1	6	20	2	71	1
2	4	23	3	74	1
7	1	26	4	86	1
8	3	29	1	89	1
9	2	32	2	92	1
10	1	36	1	110	1
11	5	38	1	113	1
12	1	53	1	116	1
14	5	59	1		
17	5	62	2		