ARCHIVOS FASTQ:
Estructura:

- La primera línea siempre empieza con @, y contienen información sobre la descripción de la secuencia y una identificación única. Una cabecera o ID única.
- La segunda línea contiene la secuencia sin procesar, la lectura de la secuencia.
- En general, la tercera línea solo contiene el signo "+" o el signo "+" y una repetición de la identificación de la secuencia.
- La cuarta línea contiene los valores de calidad y debe contener el mismo número de caracteres que la línea de secuencia. Y se codifican en código o formato ASCII, dependiendo del método de secuenciación.

Análisis de datos y de secuencia: Los datos de secuenciación se recopilan y se analizan para obtener la secuencia completa del ADN mediante una serie de programas bioinformáticos:
Limpieza y calidad de lecturas: Las secuencias de ADN generadas por la secuenciación son fragmentos cortos llamados "lecturas" o "reads". El primer paso será comprobar su calidad y "limpiarlas".
Ensamblaje de lecturas o mapeado de lecturas en función de la estrategia de secuenciación: El ensamblaje combina estas lecturas para reconstruir la secuencia completa del genoma. Se utilizan algoritmos para superponer y ensamblar las lecturas en una secuencia coherente. El mapeado de lecturas utiliza un genoma de referencia conocido para situar las lecturas sobre este genoma.

DOS ESTRATEGIAS DE SECUENCIACIÓN:

El **mapeado utilizando un genoma de referencia** es un enfoque en el que las secuencias de ADN de una muestra se comparan con un genoma de referencia previamente conocido. En este enfoque, se utilizan algoritmos para alinear las secuencias de ADN de la muestra con las secuencias correspondientes en el genoma de referencia. Este enfoque es comúnmente utilizado para identificar SNPs y otras variantes genéticas comunes en la población.
El éxito de la técnica dependerá en gran medida de la calidad del genoma de referencia. En humanos y organismos modelo tenemos genomas de referencia de gran calidad.

El **mapeado de novo** es una técnica que se utiliza para identificar y mapear secuencias de ADN previamente desconocidas en un genoma, sin utilizar un genoma de referencia previo. En este enfoque, se utilizan algoritmos para ensamblar secuencias de ADN a partir de lecturas individuales de ADN sin un genoma de referencia previo. Este enfoque es útil para identificar variantes genéticas que no están presentes en el genoma de referencia utilizado en el mapeado de lecturas cortas, como variantes estructurales complejas y mutaciones raras asociadas con enfermedades genéticas. Además, el mapeado de novo también puede utilizarse para ensamblar el genoma completo de una especie desconocida.
El mapeado utilizando un genoma de referencia es más rápido y menos computacionalmente intensivo, pero puede perder información sobre variantes genéticas que no están presentes en el genoma de referencia. Por otro lado, el mapeado de novo es más preciso para la identificación de variantes genéticas no comunes y

desconocidas, pero es más computacionalmente intensivo y requiere mayores cantidades de datos de secuenciación de alta calidad.

Un genoma de referencia es una secuencia de ADN que se utiliza como punto de comparación para analizar y comprender los genomas de individuos de la misma especie. Un genoma de referencia se crea mediante la secuenciación del ADN de un número limitado de individuos de una población y la posterior combinación de esas secuencias para crear una representación del genoma de la especie en cuestión.

Un buen genoma de referencia es aquel que es lo más completo y preciso posible, y que representa adecuadamente la variabilidad genética de la especie en cuestión. Un genoma de referencia de alta calidad debe tener una alta cobertura, lo que significa que se han secuenciado y ensamblado la mayor parte del genoma. También debe tener una baja tasa de errores, lo que significa que la secuencia del genoma es precisa y está libre de artefactos de secuenciación.
Además, un buen genoma de referencia debe ser representativo de la variabilidad genética de la especie. Esto significa que se deben incluir secuencias de múltiples individuos de diferentes poblaciones para abarcar la variabilidad de la especie. Un genoma de referencia que representa adecuadamente la variabilidad genética puede ser utilizado para estudiar la diversidad genética, las adaptaciones evolutivas y la estructura poblacional de la especie.
Por último, es importante que el genoma de referencia esté correctamente anotado con información funcional, como la ubicación y función de los genes y otras regiones del ADN. La anotación adecuada del genoma de referencia permite la identificación y el estudio de genes específicos y otras regiones funcionales del ADN.
En resumen, un buen genoma de referencia es aquel que es lo más completo, preciso y representativo posible de la especie en cuestión, y que está correctamente anotado con información funcional.

Foros de bioinformática principales: seqanswers y biostars.

ARCHIVO FASTA
Que se puede encontrar como .fasta, .fsa o .faa.

El archivo FASTA contiene la secuencia de bases de nucleótidos del genoma de referencia.
Cada línea del archivo representa una región del genoma, con una etiqueta “>” seguida de la descripción y la secuencia de bases.

También será necesario un archivo .gtf o .gff donde se encuentra la información anotada o para entendernos, los metadatos del genoma (quien es quien). El archivo GTF (General Transfer Format) es un archivo tabulado que describe la ubicación de diferentes características en el genoma, como genes, transcritos, exones y secuencias codificantes. Proporciona información sobre la estructura del genoma y la anotación de genes. Existen otros formatos de anotación genómica, como GFF (General Feature Format)

Finalmente, la anotación del genoma es un proceso crítico en la investigación genómica, ya que permite identificar y comprender las regiones funcionales del ADN y los genes que están presentes en el genoma. La anotación del genoma implica asignar funciones biológicas y características a las diferentes regiones del genoma, y proporcionar información sobre las secuencias de ADN que son importantes para la regulación de la expresión génica y otros procesos celulares.

La anotación del genoma es importante por varias razones:
1. Identificación de genes: La anotación del genoma permite identificar los genes que están presentes en el genoma, y proporciona información sobre su ubicación y función. Esto es fundamental para comprender cómo los genes interactúan y regulan los procesos celulares y fisiológicos.
2. Interpretación de variaciones genéticas: La anotación del genoma permite identificar las variaciones genéticas que pueden estar asociadas con enfermedades o rasgos específicos. Esto es importante para la investigación médica y para la identificación de objetivos terapéuticos.
3. Estudio de la evolución: La anotación del genoma permite comparar el genoma de diferentes especies y comprender cómo ha evolucionado el genoma a lo largo del tiempo. Esto puede proporcionar información sobre la historia evolutiva de una especie y sobre cómo se han desarrollado características biológicas específicas.
4. Mejora de la calidad del genoma: La anotación del genoma puede ayudar a identificar y corregir errores en la secuencia del genoma, lo que puede mejorar la calidad del genoma de referencia utilizado en la investigación genómica.

METATAXONÓMICA

La metataxonómica se centra en el análisis y la clasificación de las comunidades microbianas presentes en muestras ambientales. En la metataxonómica no secuencia el ADN de una muestra compleja de organismos en su totalidad. Se secuencia una zona concreta del ADN, mediante la amplificación de un fragmento de ADN, generalmente, genes específicos altamente conservados.

Junto a las regiones altamente conservadas, el rRNA 16S tiene nueve regiones hipervariables que permiten la clasificación por género e incluso a nivel de especie. La metataxonómica 16S se basa en la amplificación por PCR de regiones variables en el rRNA 16S utilizando cebadores específicos, seguida de la secuenciación de los amplicones. En este tipo de secuenciación se valora más una alta profundidad de secuenciación que una mayor longitud de secuencia.
A partir de la metagenómica estudiamos el conjunto de todos los genomas provenientes de una muestra compleja, mientras que mediante la metataxonómica solo vamos a amplificar el gen de ARN ribosómico en cuestión, teniendo solo la información de la taxonomía de la comunidad analizada.

Para eucariotas como hongos y levaduras, la región de interés son los ITS (Internal Transcribed Spacer), también del RNAr.

Shotgun seq vs amplicon seq

La secuenciación de amplicones tiene una serie de ventajas y desventajas. Entre las principales ventajas destaca el ser una metodología sencilla y muy barata, además de que el análisis de los resultados es "sencillo" ya que está muy bien estudiado y las bases de datos hace años que tiene un gran porcentaje de la información sobre especies más frecuentes en todos los medios (se actualizan con frecuencia, pero añadiendo muy poca información nueva entre versiones). Como contraparte, solo estamos analizando un único gen del conjunto de genes que tenemos en nuestra comunidad bacteriana. La resolución a niveles taxonómicos bajos (género o especie) es cuestionable, ya que la variabilidad entre distintas especies en la mayoría de ocasiones es tan baja que podemos obtener resultados poco significativos.

Para la DNAseq se requiere una cantidad/calidad mínima de masa, que depende del protocolo de secuenciación, pero debe estar entre 100ng < DNA < 1 microgramo.

TRANSCRIPTÓMICA

Existen dos técnicas principalmente, microarrays o RNAseq (secuenciación de RNA). Independientemente de cuál se use, siempre se parte de dos muestras, control y problema. Siempre se compara como ha variado nuestro transcriptoma con el control.
Los microarrays han quedado relegados al análisis de RNA no codificantes, mientras que el RNAm se analiza a día de hoy siguiendo técnicas y métodos de RNAseq.

RNAseq
El RNAseq o secuenciación de ARN mensajero es una técnica actual desarrollada en bioinformática para evaluar la presencia y cantidad de determinados ARN mensajeros en distintas muestras. Presenta una serie de ventajas sobre el análisis de microarray, como una mayor escala o identificación de polimorfismos de nucleótidos (tenemos acceso a la secuencia de ARN).

Ventajas en contraste con los microarrays:

- Permite análisis de todo el transcriptoma: visión global de la expresión génica
- No requiere conocimiento previo de las secuencias. No depende de sondas diseñadas previamente.
- Mayor sensibilidad: puede detectar transcripciones de baja abundancia y diferenciar isoformas.
- Identificación de variantes: puede identificar variantes genéticas y mutaciones.
- No sufre problemas de hibridación cruzada: la secuenciación de RNA no está limitada por problemas de hibridación no específica.

El alineamiento de las secuencias sobre el genoma de referencia se analiza, midiendo la cantidad de secuencias que caen sobre cada posición del genoma. Obteniéndose un valor denominado **RPKMs (Reads Per Kilobase Million)** que es la medida utilizada para cuantificar la expresión génica relativa. Los RPKMs representan la abundancia relativa de un gen específico en una muestra de ARN. Se calculan normalizando la cantidad de lecturas (reads) mapeadas a un gen por su longitud y el número total de lecturas en la muestra. Permiten comparar la expresión de diferentes genes dentro de una muestra.

Cálculo de RPKMs: El cálculo se realiza en tres pasos:
Se cuentan las lecturas mapeadas a un gen específico.
Se divide el conteo de lecturas por la longitud del gen en kilobases (KB).
Finalmente, se normaliza dividiendo por el número total de lecturas en millones.

Estos valores permitirán comparar la expresión de los genes en la muestra problema sobre (dividido por) la expresión de los genes en la muestra control, a este nuevo valor se le denomina **tasa de cambio o "fold change**". Generalmente los datos se transforman y se representa de forma logarítmica como Log2 fold change.

Los valores recogidos por RNAseq deben ser lo más precisos posibles, ya que la mínima variación en el cálculo del fold-change en función del gen analizado puede dar resultados significativamente muy diferentes entre análisis del mismo tipo.
La señal de expresión está limitada por la profundidad de secuenciación y la longitud del mensajero. Por probabilidad, a mayor profundidad conseguiremos representación de genes poco expresados, mientras que a mayor longitud de la secuencia analizada tendremos más problemas en la aparición de sesgos respecto a la representación de estas secuencias. De esta forma llegamos a la pregunta, ¿cuál es la profundidad de secuenciación mínima requerida para RNASeq?
Esta pregunta depende de qué queremos estudiar y la cantidad de mensajeros que queramos analizar. Un rango oscila entre los 5 a 20 millones de secuencias por muestra.

BASES DE DATOS

Las bases de datos primarias almacenan lo generado mediante las diversas técnicas, mientras que las secundarias almacenan la información procesada desde las bases de datos primarias. Por su parte, las especializadas reúnen información sobre, por ejemplo, organismos modelos concretos.

Principales bases de datos primarias: GenBank (NCBI), ENA (EBI) y DDBJ "Data Bank of Japan". La principal base de datos de proteínas es, a día de hoy, Uniprot.

ArrayExpress o Expression Atlas, son dos bases de datos. Dos bases de datos independientes, pero altamente conectadas. Mientras que ArrayExpress contiene Experimentos, o archivos en bruto, Expression Atlas contiene los resultados de análisis de ArrayExpress. Son dos bases de datos que se encuentran entre los recursos de datos del EBI (Instituto Europeo de Bioinformática) que pertenece al EMBL,

La base de datos ENA es la base de nucleótidos europea. Similar al NCBI. Presenta una particularidad, y es que almacena secuencias EST procedentes de transcriptomas.

ENSEMBL (del EMBL): bases de datos de genomas completos de organismos vertebrados a través de la cual es posible realizar investigación en genómica comparativa, evolución, variación de secuencia y regulación transcripcional. Ensembl anota genes, permite efectuar alineaciones múltiples, predecir la función reguladora de un gen y recopilar

datos de enfermedades. Incluye programas bioinformáticos como BLAST/BLAT, BioMart, Variant Effect Predictor (VEP).

El NCBI de los EEUU ofrece un nº variado de recursos científicos:
GenBank (almacena y constantemente actualiza información referente a secuencias genómicas)
PubMed (contiene resúmenes y citas de artículos científicos referentes a biomedicina, biotecnología, bioquímica, genética y genómica)
OMIM (contiene una recopilación de enfermedades genéticas humanas)
dbSNP (base de datos de polimorfismos de nucleótidos simples)
RefSeq (incluye la anotación precisa de genes conocidos y permite la predicción de nuevos genes basados en la evidencia de transcripción disponible)
BLAST (alineamiento múltiple de secuencias).

UniProt es la principal base de datos de proteínas la cual aúna información de otras bases de datos. Es una combinación de distintos institutos, como el EBI, el instituto suizo y la universidad de Georgetown. Esta base de datos aúna información de Swiss-Prot y TrEMBL. Estas dos bases de datos se diferencian principalmente porque Swiss-Prot revisa manualmente toda la información anotada en la misma, mientras que en TrEMBL se analiza de forma automática.

Entre las **bases de datos secundarias** encontramos aquellas que derivan información desde la secuencia de aminoácidos. Prosite es una de ellas, y guarda información de **motivos proteicos**, los cuales se relacionan con la función de las proteínas.

La base de datos KEGG contiene información sobre rutas metabólicas y las macromoléculas que están involucradas. De esta forma podemos predecir, en función de la información existente en la base de datos, cuáles son las funciones relacionadas de nuestras proteínas

PDB (Protein data bank). Reúne información relacionada con estructura de proteínas, DNA o RNA mediante cristalografía de rayos X o RMN. Determinación experimental de la cristalografía de macromoléculas. Generalmente hay proteínas, pero también hay DNA, RNA y complejos híbridos.

ALINEAMIENTO DE SECUENCIAS

Los alineamientos de secuencias sirven para conocer la similitud o diferencia entre las distintas secuencias ya sea de nucleótidos o aminoácidos. Estas diferencias son debidas a la acumulación de mutaciones en el ADN a lo largo del tiempo, y es la causa de que las secuencias de un mismo gen en dos especies distintas no sean idénticas.
Cuanto más tiempo pase desde el último antecesor común más diferentes serán las secuencias.

Los alineamientos de secuencias en bioinformática son útiles para:
- encontrar parecidos en nuevas secuencias con secuencias de las que ya se conoce su función, identificar similitudes entre secuencias y poder cuantificarlas.

- entender las dinámicas poblacionales y las relaciones evolutivas entre genes y especies, producir árboles filogenéticos.
- Encontrar dominios funcionales mediante la identificación de regiones importantes en las secuencias.

Permiten conocer como dos especies han divergido en el tiempo.
Ayudan a entender que regiones son más importantes al identificar regiones con alta similitud.
Desde que somos capaces de conocer secuencias de nucleótidos o aminoácidos, su comparación entre especies o individuos han ayudado a entender su función en el organismo.

La **homología** es un concepto abstracto que deriva de la identidad de secuencia y la función proteica, prediciendo un ancestro común. Dos secuencias podrán ser o no homologas en función de si derivan de un ancestro común, independientemente de su similitud. En teoría, podríamos tener dos secuencias con alta similitud y que no fuesen homólogas (aunque no es lo habitual). A diferencia de la **similitud**, la homología no es un término cuantitativo, dos secuencias son homólogas o no lo son.

Genes:

- Parálogos: derivan de un ancestro común, misma proteína, pero tienen diferente función.
- Ortólogos: derivan de un ancestro común y tienen la misma función. La especiación da lugar a la misma proteína en diferentes organismos, preservando la misma función.
- Análogos: no derivan de un ancestro común y realizan la misma función.

Similitud de secuencia: Considera todas las coincidencias y sustituciones.
Indica el grado de coincidencia entre dos pares de secuencias (suele expresarse en porcentajes). Incluye tanto las coincidencias exactas como las sustituciones (cuando una letra se reemplaza por otra similar).

Identidad de secuencia: Solo considera las coincidencias exactas. Indica la coincidencia total entre los pares de secuencias (similitud del 100%). La **identidad** es un **valor** que viene dado según el método de alineamiento empleado.

Una alta identidad de secuencias sugiere que el ancestro común es más reciente, mientras que una baja identidad sugiere una mayor divergencia.

Cuando comparamos secuencias lo hacemos con el objetivo principal de caracterizar a secuencia problema, asignarle un organismo y función. Para esto nos basaremos principalmente en el valor de similitud e identidad entre dos secuencias. Este valor se puede utilizar para inferir homología, pero no es un valor directo de la misma. Si bien dos secuencias que se parecen por encima de un umbral de identidad pueden tener la misma función, no existe un valor exacto que lo indique, solo podremos decir que guardan cierta homología.

A pesar de lo que estamos indicando, en la práctica, cuando alinean secuencias largas, estas solo ocurren por evolución. En cambio, los alineamientos cortos o pequeños si pueden deberse a la convergencia evolutiva.

Puntuación de alineamientos

- Número de letras que coinciden
- Porcentaje de identidad, número de coincidencias cada cien posiciones.
- Porcentaje de similitud, tiene en cuenta la similitud fisicoquímica de los diferentes aminoácidos.

Se suelen incluir dos penalizaciones para los *gaps*, una para abrir el *gap* y otra para extenderlo. Este último suele ser menos costoso. De entre todos los alineamientos posibles el óptimo es el que presenta una máxima puntuación para el sistema de puntuación dado.

Algunas de las variables más comunes que se tienen en cuenta en estos sistemas de puntuación son, el número de letras que coinciden (puntuación positiva al coincidir y negativa al no coincidir), el porcentaje de identidad, número de coincidencias cada 100 posiciones o el porcentaje de similitud (al alinear aminoácidos, tiene en cuenta sus características físico-químicas).

Existen dos tipos de alineamientos, a nivel de nucleótido y a nivel de aminoácidos. La comparación de secuencia a nivel de nucleótido es útil para estudiar filogenia y buscar diferencias singulares como la aparición de SNPs. Por otro lado, el alineamiento de secuencias de aminoácidos es útil para el estudio de homología y función de las proteínas, ya que a mayor identidad en la secuencia proteica mayor posibilidad de que se comparta la función.

Pairwise o por parejas

El alineamiento más sencillo es el alineamiento pairwise o por parejas, donde se comparan secuencias de dos en dos.
Comparación de dos secuencias mediante búsqueda de patrones de caracteres comunes.

ALINEAMIENTO GLOBAL: el objetivo es alinear toda la extensión de dos o más secuencias, maximizando la similitud a lo largo de toda su longitud.

- Secuencias de longitud similar
- Se busca estudiar la similitud base a base. Busca regiones de similitud a lo largo de toda la secuencia.
- Penaliza las inserciones y deleciones (gaps) para mantener la alineación global.
- Adecuado para secuencias con alta similitud y longitud similar (como secuencias filogenéticamente cercanas).

Usos:

- Comparación de secuencias homólogas (con un ancestro común) para estudiar la evolución y función.
- Identificación de mutaciones y variaciones genética.
- Análisis filogenético para construir árboles evolutivos.

Algoritmo: Needleman-Wunsch
Ej: ClustalOmega

ALINEAMIENTO LOCAL: el objetivo es identificar regiones de alta similitud dentro de secuencias que pueden ser de diferente longitud o tener regiones no relacionadas.

- Ambas secuencias no tienen por qué tener el mismo tamaño
- Se buscan regiones de alta identidad para caracterizar la secuencia.
- Se enfoca en encontrar las regiones más similares, sin necesidad de alinear toda la secuencia.
- Permite gaps al inicio y final de las secuencias.
- Útil para secuencias con baja similitud global o dominios conservados.

Usos:

- Detección de motivos y dominios conservados en proteínas
- Búsqueda de secuencias similares en bases de datos
- Predicción de la función de proteínas basándose en la similitud con proteínas conocidas

Algoritmos comunes: Smith-Waterman, BLAST.
Ej: BLAST, Bowtie, STAR

Métodos y algoritmos de alineamiento

- Matriz de puntos
- Programación dinámica
- Métodos heurísticos

MATRICES

Matriz de puntos.
Matriz de sustitución.
Una es un método de visionar los alineamientos y la otra es un método de cuantificación de las sustituciones.

Matrices PAM Dayhoff (Point Accepted Mutation). Se calcula observando las diferencias en proteínas estrechamente relacionadas. Al tratarse de homólogas muy próximas, no se espera que las mutaciones observadas cambien significativamente las funciones comunes de las proteínas. Así pues, se considera que las sustituciones observadas (por mutaciones puntuales) son aceptadas por la selección natural.
Al tratarse de homólogas muy próximas, no se espera que las mutaciones observadas cambien significativamente las funciones comunes de las proteínas. Así pues, se considera que las sustituciones observadas (por mutaciones puntuales) son aceptadas por la selección natural.
Normalmente se utilizan las matrices PAM30 y PAM70. Una matriz para secuencias relacionadas a mayor distancia puede calcularse a partir de una matriz para secuencias estrechamente relacionadas llevando la segunda matriz a una potencia. Así se calcula la matriz PAM250. Siendo este su principal punto débil.
El nº después de PAM es el número de veces que la matriz se ha multiplicado por sí misma. Número más grande = mejor para relaciones más distantes.

Principales características:

- Está basada en el código genético y propiedades fisicoquímicas de los aminoácidos
- Se basa en el alineamiento de secuencias muy relacionadas. Se observan las mutaciones entre ambas y se calcula una probabilidad. Los valores positivos son las mutaciones más frecuentes.
- Se genera la matriz PAM1. Una unidad PAM se define como el número de sustituciones si el 1% de posiciones de aminoácidos han cambiado.
- Las matrices PAMX derivan de la PAM1. Una PAM50 se origina multiplicando la probabilidad de cada aminoácido por 50.
- Al realizar un alineamiento de proteínas, la probabilidad de sustitución entre aminoácidos dependerá de la matriz PAM que usemos, en función de la similitud de las secuencias.

Matrices BLOSUM (Block Substitution Matrix). Basado en alineamientos locales de bloques, que son regiones cortas altamente homólogas entre pares de proteínas relacionadas que no pueden contener gaps.

- Las diferentes matrices BLOSUM tienen límites específicos para las identidades de aminoácidos.
- Se calcula la probabilidad para cada sustitución, pero en lugar de tomar el logaritmo en base 10 y multiplicar el resultado por 10 como en PAM, BLOSUM toma el logaritmo en base 2 y lo multiplica por 2.

Los números más grandes implican una distancia evolutiva más cercana, por lo que BLOSUM80 es mejor para especies estrechamente relacionadas que BLOSUM45.

Se utilizaría una matriz BLOSUM con un número más alto para alinear dos secuencias estrechamente relacionadas y un número más bajo para secuencias más divergentes (al contrario que en las PAM).
La matriz BLOSUM62 es muy buena detectando similitudes en secuencias distantes, y ésta es la matriz utilizada por defecto en las aplicaciones de alineamiento más recientes, como BLAST.

Características:

- Criterio empírico de sustitución de aminoácidos entre proteínas relacionadas, basada en bloque de alineamientos perfectos de al menos 60 aminoácidos.
- Más utilizada que PAM, mejora la fiabilidad de alineamientos locales.
- En este caso se crea una matriz en cada caso. Para la BLOSUM62, se crea una a partir de secuencias con una identidad del 62%. Al contrario que en PAM, cuanto más alto valor de la matriz, más similitud existen entre las secuencias.

El sistema de matrices BLOSUM se basa en un sistema empírico de proteínas relacionadas. Las matrices están creadas a todos los niveles de identidad, con bloques de proteínas que se relacionan desde un 1% hasta el 100%, en función de su similitud. Como se puede observar el sistema de puntuación del sistema escala de forma inversa a PAM, ya que, a mayor número, mayor es la identidad. La matriz BLOSUM62 es la más

usada por defecto. Este sistema es muy útil para valorar alineamientos locales, ya que tiene en cuenta alineamientos en bloques de 60 aminoácidos.

Las matrices PAM y BLOSUM son herramientas utilizadas en bioinformática para comparar secuencias de proteínas y entender cómo han evolucionado a lo largo del tiempo.
La diferencia clave entre ellas es la forma en que consideran la evolución:
Matrices PAM (Matrices de Puntuación de Afinidad de Puntos Mutantes): Estas matrices se basan en un modelo evolutivo explícito, lo que significa que consideran las sustituciones de aminoácidos en función de un árbol filogenético que muestra cómo se cree que las especies han evolucionado con el tiempo. Es como si observaran la historia evolutiva directamente para hacer sus cálculos.
Matrices BLOSUM (Matrices de Bloques de Aminoácidos Conservados): En cambio, las matrices BLOSUM se basan en un modelo de evolución implícito, lo que significa que no consideran un árbol filogenético específico. En lugar de eso, se centran en la similitud de bloques de aminoácidos conservados en secuencias de proteínas. Estas matrices se calculan observando cuánto se conservan ciertos bloques de aminoácidos en proteínas relacionadas.
En resumen, PAM se basa en un modelo que sigue la evolución de especies específicas a lo largo del tiempo, mientras que BLOSUM se enfoca en patrones de aminoácidos conservados en proteínas relacionadas sin preocuparse por la historia evolutiva exacta.

El valor predeterminado de la herramienta BLAST es una penalización de -11 por abrir gap y -1 por cada base adicional, se resume como: gap (11/1).

PROGRAMACIÓN DINÁMICA

La técnica de programación dinámica puede aplicarse para producir alineamientos globales mediante el algoritmo Needleman-Wunsch, y alineamientos locales mediante el algoritmo Smith-Waterman.

Generalmente, los alineamientos de proteínas utilizan una matriz de sustitución (PAM o BLOSUM) para asignar puntuaciones a las coincidencias o discordancias de aminoácidos y una penalización por hueco para hacer coincidir un aminoácido de una secuencia con un hueco de la otra y de esta forma rellenar la matriz de alineamiento que se ha formado en la técnica de programación dinámica. Los alineamientos de ADN y ARN pueden utilizar una matriz de puntuación más simple, pero en la práctica a menudo simplemente se asigna una puntuación positiva a las coincidencias, una puntuación negativa a los cambios y una penalización negativa a los huecos. En la programación dinámica estándar, la puntuación de cada posición de un aminoácido es independiente de la identidad de sus vecinos, por ejemplo, es posible tener en cuenta dichos efectos modificando el algoritmo, complicándolo y de esta forma teniendo en cuenta más parámetros.

De esta forma, la programación dinámica se evalúa de forma diferente en función del tipo de alineamiento que estemos llevando a cabo. Para un alineamiento local se utiliza

el algoritmo de Smith-Waterman, el cual, es muy útil para secuencias divergentes o de distinta longitud. Por su parte, el alineamiento global usa el algoritmo de Needleman-Wunsch.

Años de desarrollo de los algoritmos de alineamiento de secuencias:
Needleman-Wunsch: Desarrollado en 1968 por Saul Needleman y Christian Wunsch. Es un algoritmo de programación dinámica que busca el alineamiento global que maximiza la puntuación global, considerando tanto la similitud entre las secuencias como la penalización por gaps.
Smith-Waterman: Desarrollado en 1976 por Temple F. Smith y Michael S. Waterman. Es un algoritmo de programación dinámica que busca el alineamiento local que maximiza la puntuación local, considerando tanto la similitud entre las secuencias como la penalización por gaps.
ClustalW: Desarrollado en 1994 por Desmond Higgins y Aidan Clustal. Es un algoritmo de alineamiento múltiple que utiliza una técnica de "divide y vencerás" para mejorar la eficiencia del alineamiento de un conjunto de secuencias.

MÉTODOS HEURÍSTICOS
Al ser una gran cantidad de alineamientos la capacidad de cómputo requerida aumenta exponencialmente, siendo inviable comparar las secuencias utilizando los métodos de matrices de punto o programación dinámica.
Por esto, se desarrollan en este punto, métodos heurísticos capaces de realizar el proceso hasta 100 veces más rápido. A diferencia de los anteriores métodos, estos tienen un enfoque que se basa en reglas prácticas y estrategias aproximadas para encontrar soluciones que pueden no ser necesariamente las mejores o las óptimas, pero que son lo suficientemente buenas en un tiempo razonable. Los métodos heurísticos sacrifican la exhaustividad en favor de la eficiencia, lo que significa que pueden encontrar soluciones aceptables más rápidamente, especialmente en casos en los que la búsqueda exhaustiva de todas las posibles combinaciones es computacionalmente costosa o impracticable. Estos métodos heurísticos son particularmente útiles cuando se trabaja con secuencias largas o bases de datos grandes, donde los algoritmos de programación dinámica, que buscan soluciones óptimas, pueden ser computacionalmente prohibitivos.

Los métodos heurísticos ofrecen un equilibrio entre la calidad de la solución y el tiempo de procesamiento, lo que los hace ampliamente utilizados en la bioinformática y el análisis de secuencias biológicas. Ejemplos de estos métodos son los métodos FASTA y BLAST.

BLAST (Basic Local Alignment Search Tool): BLAST es una herramienta bioinformática gratuita y de código abierto desarrollada por el National Center for Biotechnology Information (NCBI) de los Estados Unidos. Permite comparar una secuencia de consulta (query) con una gran base de datos de secuencias (subject) para identificar regiones de similitud local. Esta comparación se basa en un algoritmo heurístico que evalúa la similitud entre las secuencias mediante matrices de puntuación y modelos estadísticos.

BLAST es uno de los métodos heurísticos más ampliamente utilizados para la búsqueda y alineamiento de secuencias en bases de datos biológicas. Opera realizando búsquedas locales, lo que significa que busca regiones similares en lugar de alinear secuencias completas. BLAST utiliza una estrategia de búsqueda basada en "palabras" (k-mers) y extensiones, lo que lo hace rápido y eficiente para buscar similitudes en grandes conjuntos de datos.

El e-value es el valor más importante a tener en cuenta al analizar un alineamiento mediante Blast. El e-value calcula la probabilidad de obtener un falso positivo en nuestro alineamiento y depende de forma directa del tamaño de la base de datos utilizada. Surge de la problemática de encontrar por azar nuestra secuencia en una base de datos al aumentar el tamaño de las bases de datos de secuencias de forma exponencial en el transcurso del tiempo.

El valor **Bitscore**, muy similar al e-value. En este caso se valora la identidad y similitud de dos secuencias, pero sin tener en cuenta el tamaño de la base de datos. Este valor se instauró porque se vio un posible problema en la disminución de la sensibilidad del sistema al aumentar el tamaño de la base de datos, donde dos secuencias similares dejaran de ser tan similares al aumentar la información en ellas.

FASTA: Similar a BLAST, el algoritmo FASTA también es una herramienta heurística que se utiliza para buscar secuencias similares en bases de datos biológicas. Utiliza un enfoque de búsqueda local y utiliza estrategias de palabras clave y extensiones para encontrar alineamientos.

La principal diferencia entre BLAST y FASTA es que BLAST se dedica sobre todo a encontrar alineamientos de secuencias óptimas a nivel local sin gaps (o lo menos posible), mientras que FASTA se dedica a encontrar similitudes entre secuencias menos parecidas.

ALINEAMIENTOS MÚLTIPLES

Es útil para alineamiento de secuencias de nucleótidos y aminoácidos, siendo especialmente relevante este último para estudios de filogenia.
La principal utilidad de estos alineamientos es el estudio evolutivo de las secuencias, a través del alineamiento de secuencias homólogas y conservadas entre distintas especies localizando los dominios importantes de función en proteínas.
En este método se siguen utilizando las matrices de puntuación mediante el sistema de "suma por pares", que calcula el valor entre todas las combinaciones de bases alineadas.

Algunas de las aplicaciones más comunes incluyen:

1. Filogenia Molecular: Los alineamientos múltiples se utilizan para construir árboles filogenéticos que representan las relaciones evolutivas entre diferentes especies. Al comparar las diferencias y similitudes en las secuencias genéticas, se puede inferir la historia evolutiva y la diversidad de las especies.
2. Análisis de Homología: Los alineamientos múltiples permiten identificar genes homólogos, es decir, genes relacionados por descendencia evolutiva común. Esto

es esencial para la predicción de funciones de genes desconocidos y la comprensión de las similitudes y diferencias genéticas entre especies.

3. Análisis de Regiones Conservadas: Los alineamientos múltiples ayudan a identificar regiones genómicas o proteicas conservadas en diferentes especies. Estas regiones conservadas suelen estar relacionadas con funciones biológicas importantes y se utilizan en la anotación de genes y la identificación de elementos reguladores.
4. Identificación de Motivos y Dominios: Los alineamientos múltiples se utilizan para identificar motivos o dominios conservados en secuencias proteicas. Esto es útil para predecir funciones de proteínas, como la unión a ligandos o la interacción con otras proteínas.
5. Análisis de Evolución Molecular: Al comparar alineamientos múltiples de secuencias de genes o proteínas, es posible estudiar cómo han evolucionado a lo largo del tiempo. Esto proporciona información sobre los eventos de duplicación génica, mutaciones y adaptaciones a diferentes entornos.
6. Análisis de Variabilidad Genética: Los alineamientos múltiples se utilizan para identificar polimorfismos de nucleótidos únicos (SNPs) y otros tipos de variabilidad genética en poblaciones. Esto es esencial en estudios de genética de poblaciones y asociación de variantes con enfermedades.
7. Diseño de Cebadores y Sondas: Los alineamientos múltiples se utilizan para diseñar cebadores y sondas específicas para la amplificación o detección de secuencias genéticas de interés en técnicas como la PCR o la hibridación in situ.
8. Análisis de Expresión Génica: Al comparar secuencias de genes y sus regiones reguladoras en múltiples especies, es posible estudiar la expresión génica y las diferencias en la regulación génica entre diferentes organismos.

Alineamiento Múltiple Progresivo:

Enfoque: El método progresivo construye el alineamiento paso a paso, comenzando con la alineación de dos secuencias y luego agregando secuencias adicionales una a una. Se basa en la jerarquía de alineamientos.

Estrategia: En cada paso, las secuencias ya alineadas se tratan como una sola secuencia "consenso" y se alinean con la siguiente secuencia de entrada. Las dos secuencias más similares son las guías. Este proceso se repite hasta que se alinean todas las secuencias.

Posibles problemas: las dos primeras secuencias generen un mal alineamiento.

Ejemplo: Un ejemplo de alineamiento progresivo es el método ClustalW, MAFFT, MUSCLE.

Pasos:

Selección de pares de secuencias: Se seleccionan dos secuencias del conjunto y se alinean utilizando un método de alineamiento de pares, como la programación dinámica o heurísticas.

Construcción de un árbol guía: Se construye un árbol filogenético que representa las relaciones de similitud entre las secuencias. Este árbol guía sirve como referencia para el alineamiento posterior.

Alineación progresiva: Se alinean secuencias adicionales al alineamiento inicial en función del árbol guía. Se inserta cada nueva secuencia en la posición más probable, considerando su similitud con las secuencias ya alineadas.

Refinamiento del alineamiento: Se realizan ajustes al alineamiento para optimizar la puntuación global y minimizar los gaps (huecos).

Alineamiento Múltiple Iterativo:
A diferencia del Alineamiento Múltiple Progresivo, que construye el alineamiento de forma gradual a partir de pares de secuencias, el AMI utiliza un enfoque iterativo que refina un alineamiento inicial de manera global.
Enfoque: El método iterativo comienza con un alineamiento inicial, que puede ser generado por un método progresivo o por un método heurístico. Luego, se mejora iterativamente el alineamiento, reajustando las secuencias para optimizar la puntuación del alineamiento.
Estrategia: Después del primer alineamiento, las secuencias se reeligen en función de sus similitudes y se vuelven a alinear en cada iteración. Este proceso se repite hasta que la calidad del alineamiento converge.
Ejemplo: El algoritmo MAFFT es un ejemplo de un método de alineamiento múltiple iterativo.
Mediante este método se generan inicialmente alineamientos entre secuencias aleatorias del sistema y se valora el bloque final resultado. Se realiza el sistema de nuevo y se va guardando la información sobre los mejores alineamientos que se generaron en el paso anterior. Las interacciones que sufre el sistema le permiten aprender cual son los mejores alineamientos. Este sistema se repetirá hasta llegar a un mínimo, una solución estable, existiendo el potencial problema de que nunca se llegue al mismo. Uno de los ejemplos más relevantes es el de ClustalOmega, una evolución de ClustalW.

Pasos:
Obtención de un alineamiento inicial: Se utiliza un método de alineamiento de pares o un alineamiento aleatorio como punto de partida.
Evaluación del alineamiento: Se calcula una puntuación que refleja la calidad del alineamiento, como la suma de los gaps o la divergencia entre las secuencias.
Optimización del alineamiento: Se realizan cambios locales en el alineamiento para mejorar la puntuación, como mover segmentos de secuencias o insertar gaps.
Repetición de los pasos 2 y 3: Se repiten la evaluación y optimización del alineamiento hasta que no se observen mejoras significativas en la puntuación.

Alineamiento Múltiple Basado en Bloques:
El último criterio de alineamiento se denomina basado en bloques. En este sistema se originan alineamientos locales entre las secuencias, buscando bloques conservados y evaluándolos según el sistema de matrices de puntos y suma por pares. Este sistema apenas se usa porque realmente está basado en alineamientos locales. A diferencia del alineamiento múltiple progresivo (AMP) y el alineamiento múltiple iterativo (AMI), que alinean las secuencias en su totalidad, el AMB se enfoca en identificar y alinear regiones conservadas entre las secuencias, también llamadas bloques.
Enfoque: Este enfoque divide las secuencias de entrada en bloques antes de realizar el alineamiento. Cada bloque contiene secuencias que se cree que tienen una relación evolutiva cercana.
Estrategia: Cada bloque se alinea por separado, y luego los bloques alineados se ensamblan en un alineamiento múltiple completo. Esta estrategia se utiliza para lidiar

con secuencias altamente divergentes o con estructura de dominio compleja.
Ejemplo: El programa GBlocks es un ejemplo de un método de alineamiento múltiple basado en bloques.

Enfoque de construcción: En el progresivo, se construye el alineamiento de manera incremental. En el iterativo, se parte de un alineamiento inicial y se refina iterativamente. En el basado en bloques, se dividen las secuencias en bloques y se alinean por separado antes de ensamblar el alineamiento final.
Estrategia: El progresivo se basa en la jerarquía de alineamientos, el iterativo busca la convergencia y mejora de la calidad, y el basado en bloques se centra en segmentos conservados.
Aplicación: El enfoque progresivo es útil para secuencias estrechamente relacionadas. El iterativo es efectivo para refinar alineamientos iniciales. El basado en bloques es útil cuando se tienen secuencias altamente divergentes o con estructura de dominio compleja.
La elección entre estos enfoques depende de la naturaleza de las secuencias y el objetivo del análisis de alineamiento múltiple. Cada uno tiene sus propias ventajas y desventajas.

A la hora de elegir el mejor método de secuenciación dependerá del tipo de estudio que vayamos a realizar. Si secuenciamos un solo genoma y queremos estudiar variabilidad a nivel de nucleótido único la mejor alternativa puede ser Sanger o Illumina, si queremos estudiar un metagenóma de un entorno complejo o amplicones 16S usaremos Illumina MiSeq. Por el contrario, si queremos secuenciar un genoma de alta concentración de regiones repetitivas, que presentarán dificultades en su ensamble, lo mejor será usar secuenciadores de tercera generación.

HERRAMIENTA FASTQ-DUMP o FASTERQ-DUMP
Para descargar los runes de un Bioproject
Para descargar todos a la vez se puede usar el comando: **xargs -n1 fastq-dump <** SRR_Acc_List.txt
Y el nombre de la lista donde los tengas descargados: en este caso SRR_Acc…

BÚSQUEDA DE SECUENCIA PROBLEMA

Hay que saber el nombre de la base de datos que le hemos puesto, donde están todos los archivos de secuencias que descargamos.
Con el **parámetro -db (**database) el argumento que hay que introducir es **el nombre de la base de datos**, no de los archivos.
Ej:
blastn -query secuencia_problema.fasta -db ViralDB -out resultado_blast_secuencia_problema.txt -outfmt 1

CLUSTALO
Es una de las herramientas de alineamiento que más se usan.

clustalo -i prot_clustalOmega.txt -o alinea_seqs --distmat-out=salidaDISMAT --percent-id –full

con el distmat (distan matrix) es para que te cree una matriz, y lo último es que lo muestre en porcentaje de identidad (y que vaya a full)

MUSCLE

ABIVIEW
Usarlo dentro del environment que tenemos el 03MBIF_v4

Para hacer una lectura de la secuencia fastq: **grep "@**
En este caso grep "@MISEQ" -c Metagenoma_R1.fastq

Cuando te llegan unas secuencias directamente del secuenciador primero se comprueba que la secuencia este bien, y el número de lecturas que tiene. Luego se hace un Blast para saber de qué organismo se trata.

IGV

Podemos meter lecturas contra un genoma de referencia.

CONTAR LECTURAS
grep @ el nombre del secuenciador -c y el nombre del archivo.
Ej: grep @MISEQ -c Viroma.fastq

BLAST
En blast, para secuencias que no son tan tan similares abajo hay que usar la opción que pone: more dissimilar sequences (discontiguous megablast)

HERRAMIENTA FASTQC
fastqc del archivo que se quiere

PRINSEC-LITE

Lo hemos usado para recortar la parte de la secuenciación que no era buena, y quedarnos con la mejor parte.
Ej: **prinseq-lite.pl** -fastq Viroma.fastq -**min_qual_mean 30**

Hay que ver donde está la subpoblación de adaptadores, que es la que fastidia un poco el análisis.

PROCESAMIENTO Y CALIDAD

Herramientas de eliminación de adaptadores: cutadapt, trommoatic, fastp