

## TP5 – Alineamiento y BLAST

👉 **PARA PENSAR:** ¿Qué tipo de información se puede extraer de la comparación de secuencias?  
¿Cómo esperas que se vea en una comparación? 😞

Podemos analizar semejanzas entre dos especies.

Esperaría ver una comparación letra por letra tratando de que coincidan la mayor cantidad posible.

👉 **PARA PENSAR:** ¿Por qué crees que es mejor evaluar las relaciones evolutivas lejanas comparando proteínas? 😞

Porque las proteínas se ven menos modificadas con el tiempo de lo que puede ser una secuencia de ácidos nucleicos, que pudo sufrir muchas alteraciones por distintos factores.

👉 **RETO I:** Intentemos, entonces alinear estas dos palabras, para comprender mejor el problema.  
Alinea en la siguiente table de comparaciones las palabras "BANANA" y "MANZANA".

¡Tomá nota de tus observaciones y de las conclusiones que se desprendan de estas observaciones!

☒ **PREGUNTAS DISPARADORAS:** ¿Existe una única forma de alinearlas? ¿Es alguno de los posibles alineamientos mejor que otro? Si así fuera ¿Por qué?

B	A	N	A	N	A	-	-
M	A	N	Z	A	N	A	-
✗	✓	✓	✗	✗	✗	✓	✓


☒ **PREGUNTAS DISPARADORAS:** ¿Qué representan esos guiones?

El mejor alineamiento que se logro fue el siguiente:


B	A	N	-	A	N	A	-
M	A	N	Z	A	N	A	-
✗	✓	✓	✓	✓	✓	✓	✓

No existe una única forma para alinearlas. Sin embargo, hay algún alineamiento que resulta mejor que el otro. Por ejemplo, en el alineamiento que se realizó, coinciden mayor número de letras respecto del que se mostró como ejemplo.









Los guiones son espacios en blanco que pueden representar la falta de un aminoácido de la secuencia.


 **RETO II:** En la siguiente tabla probá distintos alineamientos para las palabras "ANA" y "ANANA". Verás que en el margen superior izquierdo aparece un valor de identidad calculado para cada alineamiento que intentes.

Tomá nota de los valores de identidad observados y de las conclusiones que se desprendan de estas observaciones.

 **PREGUNTAS DISPARADORAS:** ¿Son todos los valores iguales? ¿Qué consideraciones deberían tenerse en cuenta a la hora de realizar el cálculo? ¿Se te ocurre, distintas formas de calcularlo? ¿Serán todas ellas igualmente válidas en Biología?

Penalidad  Identidad


A	N	A	-	-	
A	N	A	N	A	
					



Después de probar distintos alineamientos, notamos que no son todos iguales. No es lo mismo cuando se comparan las mismas letras que cuando no se hace.

Se debería tener en cuenta si se comparan dos letras iguales, diferentes, o una letra y un espacio en blanco. Se podría calcular todos los alineamientos posibles para ambas secuencias y quedarse con el que mayor puntaje tenga. A la hora de realizar el cálculo se deben tener en cuenta varias cosas. Por ejemplo, la penalidad por comparar una letra con un gap o la penalidad de comparar dos letras distintas.

No todas serán igualmente válidas, ya que cuanto menor sea el puntaje, seguramente sea mucho menos probable que sean la misma secuencia o tengan relación alguna.

 **RETO III:** En la siguiente tabla probá distintos alineamientos para las palabras "ANA" y "ANANA". Verás que en el margen superior izquierdo aparece un valor de identidad calculado para cada alineamiento que intentes y un botón para cambiar la penalidad que se le otorga a dicho para el cálculo de *identidad*.

Probá varias combinaciones, tomá nota de los valores de identidad observados y de las conclusiones que se desprendan de estas observaciones.

☒ **PREGUNTAS DISPARADORAS:** ¿Cómo se relacionan los valores de identidad obtenidos con las penalizaciones que se imponen al gap? ¿Qué implicancias crees que tiene una mayor penalización de gaps? ¿Se te ocurre alguna otra forma de penalización que no haya sido tenido en cuenta en este ejemplo?

Penalidad  Identidad 0.2

A	N	A	-	-	<input checked="" type="checkbox"/>
A	N	A	N	A	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

☒

Se relacionan directamente.

A mayor penalización impuesta al gap, menor es el puntaje obtenido en la identidad.  
Por ejemplo, el siguiente caso muestra lo que se comenta:


Penalidad  Identidad -0.2

A	-	-	N	A	<input checked="" type="checkbox"/>
A	N	A	N	A	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>


☒

Al aumentar la penalidad, la identidad baja.

Se me ocurre que puede existir diferentes tipos de penalidades al comparar aminoácidos entre sí. No será igual compararlo con uno, que con otro.

 **PARA PENSAR:** Entonces, pensando en un alineamiento de ácidos nucleicos ¿Cuáles te parece que son las implicancias de abrir un gap en el alineamiento? ¿Qué implicaría la inserción o delección de una región de más de un residuo?

Cuanto más gaps se tengan que agregar en el alineamiento, mas nos alejamos de que sean secuencias iguales.

 **RETO IV:** En la siguiente tabla probá distintos alineamientos para las secuencias nucleotídicas. Podrás ver las traducciones para cada secuencia.

Probá varias combinaciones, tomá nota de las observaciones y de las conclusiones que se desprendan de estas.

Consigna: Alineá "TGCGAGG" y "TGCCGAAGG" y mirá las traducciones


Penalidad  Identidad 0.3333333333333333

T	G	C	G	A	G	G	-	-	
C			E			-			
T	G	C	C	G	A	A	G	G	
C			R			R			

Consigna: *Intentá* alinear "AGGGGA" y "TGCAGAGGG" y mirá las traducciones

Penalidad  Identidad 0.2222222222222222

A	G	G	G	G	A	-	-	-	
R			G			-			
T	G	A	G	A	G	G	G	-	
X			E			-			

 **PARA PENSAR:** ¿Dá lo mismo si el gap que introducís cae en la primera, segunda o tercer posición del codón? ¿Cómo ponderarías las observaciones de este ejercicio para evaluar el parecido entre dos secuencias?

Según las pruebas realizadas, da igual en que posición dentro del codón, este colocado el gap. Este ejercicio nos hace dar cuenta de que alinear secuencias no es tan simple como parecía. Se deben tener en cuenta distintos factores de penalización o de acierto y en base a eso calcular la identidad del alineamiento. Creo que resulta muy útil este tipo de herramientas para detectar similitudes entre secuencias.

👉 **RETO V:** Estuvimos viendo que el alineamiento de secuencias no es trivial y requiere contemplar los múltiples caminos posibles, teniendo en cuenta al mismo tiempo la información biológica que restringe ese universo de posibilidades.

¡Es momento de llevar entonces estos conceptos a lo concreto!

Te proponemos pensar los pasos a seguir en un alineamiento de dos secuencias cortas, teniendo en cuenta una matriz genérica de scoring (puntuación) que contemple las complejidades que estuvimos viendo, es decir que penalice de distinto modo una inserción o deleción, una discordancia (mismatch) o una coincidencia (match). Escribilos o esquematizalos en un diagrama de flujo.

👉 **PARA PENSAR:** ¿En qué consiste la programación dinámica? ¿Por qué crees que es útil en este caso?

- Partimos de dos secuencias.
- Las colocamos sobre una matriz, una secuencia sobre la vertical y la otra sobre la horizontal.
- La primer fila y columna contiene los valores acumulados de la secuencia hasta ese momento comparada con los gaps.
- Completo la matriz:
  - Por cada celda calculo el match/mismatch entre los caracteres pertenecientes a la fila y columna correspondiente.
  - Sumo ese valor a los tres obtenidos en las celdas adyacentes.
  - Me quedo con el de mayor valor.
  - Lo guardo en la celda que estoy y paso a la siguiente.
  - El orden para completar la matriz será de izquierda a derecha y una vez completa la fila se pasará a la siguiente.
- Cuando la matriz se encuentra completa, comenzando de la celda inferior derecha, retrocedo por las adyacentes siempre eligiendo la de mayor score.
- De esta forma voy obteniendo el alineamiento

La programación dinámica consiste en buscar la solución a problemas complejos de la forma mas optima posible. Resulta útil en este caso, debido a la complejidad de computo que puede tener tratar de alinear secuencias grandes. De esta forma obtenemos una solución de una manera óptima, que con otro tipo de metodología podríamos tardar muchísimo mas en conseguir.

👉 **RETO VI:** Utilizando la herramienta interactiva [desarrollada por el Grupo de Bioinformática de Freiburg](#) probá distintos *Gap penalties* para el ejemplo propuesto y observá lo que ocurre.

Interpretando la recursión, explicá con tus palabras de dónde salen los valores de la matriz que se construye. ¡Esquematiza tus conclusiones!

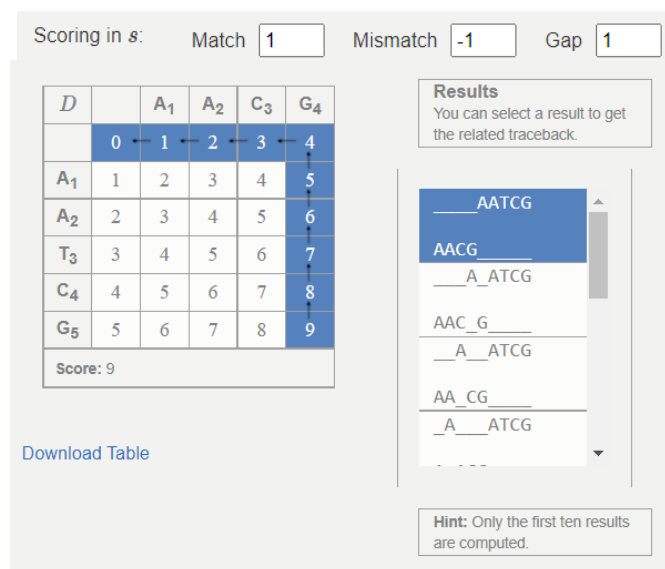
Asimismo existen herramientas que permiten tanto comparaciones de secuencias de a pares y o realizar alineamientos múltiples:

- A pares de secuencias: mide la similitud entre dos secuencias.
- Alineamiento múltiple: compara más de dos secuencias al mismo tiempo.

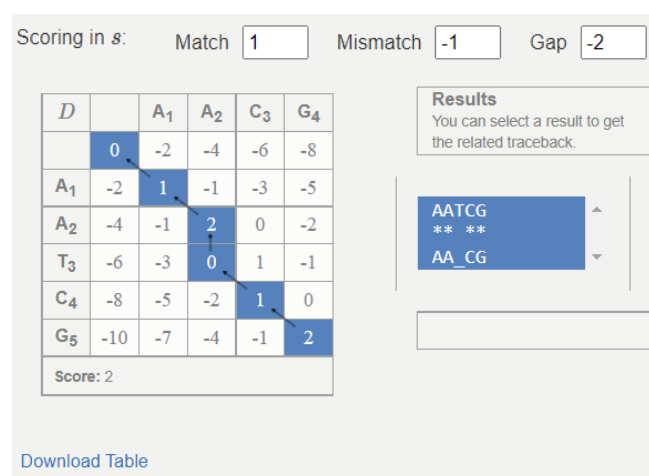
En ambos casos el alineamiento puede ser local o global, lo que supondrá algunas limitaciones de uso para cada caso.

👉 PARA PENSAR: ¿En qué casos serán de utilidad uno u otro tipo de alineamientos? ¿Qué limitaciones tendrá cada uno?

Al ir cambiando el puntaje propuesto para las gap penalties, notamos que al aumentar el puntaje, el alineamiento tiende a preferir los gap, debido a que esto aumenta el score general.



En cambio, al ir disminuyendo las gap penalties, el alineamiento que se obtendrá tendrá mayor porcentaje de match/mismatch, ya que colocar un gap disminuirá mucho el score general.



Cada tipo de alineamiento será mejor o peor opción dependiendo del caso de estudio. Por ejemplo, un alineamiento de pares de secuencias puede servir para hacer una prueba de paternidad. En cambio, si se

está averiguando a que especie pertenece una proteína, entiendo que será mucho mejor hacer un alineamiento de secuencias múltiples, para así tener muchas más posibilidades de comparación.

La principal limitación del alineamiento de a pares, es la cantidad de secuencias que permite alinear. Este alineamiento en muchos estudios no resultaría muy útil.

En cambio, en el alineamiento múltiple, la gran limitante será la complejidad de calculo que tiene el comparar tantas secuencias unas con otras.

👉 PARA PENSAR: Ingresá al servidor del NCBI y mirá los distintos programas derivados del [BLAST](#) que se ofrecen ¿Para qué sirve cada uno? ¿En qué casos usarías cada uno?

- Basic Local Alignment Search Tool (BLAST): Encuentra regiones de similitud local entre secuencias biológicas.
- BLAST Link (BLink): Muestra los resultados de una búsqueda BLAST precalculada de una proteína frente a todas las demás secuencias de proteínas en NCBI.
- BLAST Microbial Genomes: Encuentra regiones de similitud local entre secuencias de consulta y secuencias de genomas microbianos completos.
- BLAST RefSeqGene: Encuentra regiones de similitud local entre secuencias de consulta y secuencias genómicas en el conjunto RefSeqGene/LRG.
- Concise Microbial Protein BLAST: Encuentra regiones de similitud local entre proteínas de consulta y proteínas de genomas microbianos (procariotas) completos.
- Gene Expression Omnibus (GEO) BLAST: Encuentra regiones de similitud local entre secuencias de consulta y secuencias GenBank incluidas en plataformas de microarrays o SAGE en la base de datos GEO.
- Genome BLAST: Encuentra regiones de similitud local entre las secuencias de consulta y las secuencias del genoma.
- Primer-BLAST: Utiliza Primer3 para diseñar cebadores de PCR para una plantilla de secuencia.
- PSSM Viewer: Muestra y manipula matrices PSSM de registros CDD y PSI-BLAST.
- SNP Database Specialized Search Tools: Busca en la base de datos SNP por genotipo, método, población, remitente, marcadores y similitud de secuencia usando BLAST.
- SmartBLAST: Encuentra proteínas similares a la consulta.
- IgBLAST: Busca inmunoglobulinas y secuencias de receptores de células T.
- MOLE-BLAST: Establece taxonomía para secuencias ambientales o no cultivadas.

👉 RETO VII: calculá el E-value y porcentaje de identidad utilizando el programa BLAST de la siguiente secuencia input usando 5000 hits, un e-value de 100 y tomando aquellos hits con un mínimo de 70% cobertura. Observe y discuta el comportamiento de : E-value vs. % id, Score vs % id, Score vs E-value

LLLLKGEIELEIDAKWNEKAAEOWCIIOLAEKZQVAAOBGZWMAGEZBBALGTIGFIEFIAG  
AACGGCGAMGZAMZBYIHFCZDAEDBAAVEIMNHBYIOAAVBYMDEAZIOIIEAHDCAMIIKONLA

El E-value indica si el alineamiento se ha dado por casualidad. A mayor valor, menos significativo es. El %identidad representa si un gran numero de secuencias fueron alineadas de la misma forma. Y el score, es el puntaje que se le otorga al alineamiento. La relación E-value vs %id, en nuestro ejemplo nos muestra que al ir aumentando %id, el E-value disminuye. Por lo que nos da un indicador de que el alineamiento mejora. Respecto la relación score vs e-value, en nuestro ejemplo, al disminuir el e-value, también está disminuyendo el score. Por último, para la relación score vs %id, mantiene una relación directa, conforme uno aumenta el otro también. Valores altos de estos dos parámetros, nos indica que es una muestra mucho mas significativa.

👉 RETO VIII: Realizá nuevas búsquedas usando la mitad de la secuencia problema y para un cuarto de la secuencia original. Compará los gráficos obtenidos. ¿Qué conclusiones puede sacar?

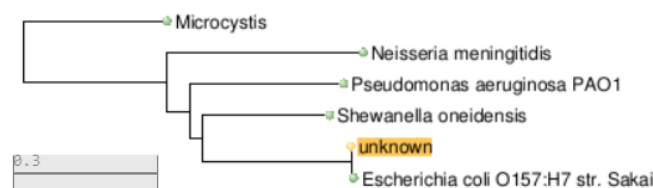
Para este caso, notamos que el E-value es mucho mayor. Esto nos indica que, al ser una secuencia mucho mas corta, la posibilidad de que los match sean por casualidad aumenta notablemente. También podemos comentar que los score disminuyeron mucho.

👉 RETO IX: Utilizando BLAST utilice búsquedas de similitud secuencial para identificar a la siguiente proteína:

MIDKSAFVHPTAIVEEGASIGANAHIGPFCIVGPHVEIGEGTVLKSHVVVNGHTKIGRDNEIYQFASIGEVENQ  
DLKYAGEPTRVEIGDRNRIRESVTIHRGTVQGGGLTKVGSNDLLMINAHIAHDCTVGNRCILANNATLAGHV  
SVDDFAIIGGMTAVHQFCIIGAHVMVGGCSGVAQDVPPYVIAQGNHATPFGVNIEGLKRRGFSREAITAIRN  
AYKLIYRSGKTLDEVKPEIAELAETPEVKAFTDFFARSTRGLIR

👉 PARA PENSAR: ¿Cuál es la función de la proteína? ¿A qué grupo taxonómico pertenece? A un nivel de significancia estadística adecuado ¿cuántas secuencias similares se encuentran?

Al realizar la búsqueda encontramos que la proteína probablemente sea Acetylglucosamine O-acetyltransferase [Escherichia coli].





👉 RETO X: Realizá una nueva corrida del BLASTp, utilizando la misma secuencia , pero ahora contra la base de datos PDB. ¿Se obtienen los mismo resultados? ¿Qué tipo de resultados(hits) se recuperan? ¿Cuándo nos podría ser útil este modo de corrida?

A simple vista notamos que no se obtuvieron los mismos resultados. Solamente se obtuvieron 32 secuencias. Este tipo de corrida podría sernos útil para obtener un primer resultado de un numero acotado de secuencias y en base a eso comenzar a sacar conclusiones, para luego hacer un análisis más exhaustivo.