Introducción a la Bioinformática

TP Alineamientos Secuenciales

Enrique Alonso

DESAFIO I: Intentemos, entonces alinear estas dos palabras, para comprender mejor el problema. Alineá en la tabla interactiva las palabras "BANANA" y "MANZANA".

Notamos que podemos alinear, usando criterios diferentes de muchas maneras diferentes las dos palabras. Sin dejar espacios, es decir, alineando las palabras como un conjunto in separable y relacionándolas por la coincidencia de algunas letras en particular. También dejando espacios, por ejemplo separando BAN- ANA para que coincida el inicio con el final logrando más coincidencias.

Se hace más visible la necesidad de tener un criterio para tomar decisiones.

DESAFIO II: En la siguiente tabla interactiva 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.

Se observa que adicional a la penalidad de no coincidir letras entre palabras también existe una penalidad sobre la identidad (que alcanza el valor 1 en caso que las palabras sean idénticas y estén alineadas). Esto es que aunque las letras A-N-A coincidan con las letras de ANANA, existe una penalidad adicional por los gaps, espacios no comparables, porque no hay una de las partes a comparar.

DESAFIO III: Probá en tabla interactiva 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.

Aparentemente la penalidad actualiza como multiplicador de los casos que no son coincidencias, lo cual termina reduciendo la identidad notablemente inversamente al crecimiento de la penalidad.

Por ejemplo, para un alineamiento de las 3 letras de ANA con ANANA tenemos:

PENALIDAD	IDENTIDAD
0	0.6
1	0.4
2	0.2
3	0

DESAFIO IV: Probá en la tabla interactiva 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.

Se hace visible la necesidad de contar con uno o más criterios para alinear codones , o por ejemplo alineando las aminoácidos que representan o parte de ellos.

DESAFIO 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.

Teniendo en cuento lo visto en clase realizamos una matriz genérica con un par de secuencias cortas y obtenemos el mejor caso posible. Replico un caso general. La tabla es un Excel que calcula las celdas según el match y las celdas aledañas.

		A	H	C	N	I	R	V	S
	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	1	0	-1	-2	-3	-4	-5	-6
I	-2	0	0	-1	-2	-1	-2	-3	-4
C	-3	-1	-1	1	0	-1	-2	-3	-4
I	-4	-2	-2	0	0	1	0	-1	-2
N	-5	-3	-3	-1	1	0	0	-1	-2
R	-6	-4	-4	-2	0	0	1	0	-1
C	-7	-5	-5	-1	-1	-1	0	0	-1
K	-8	-6	-6	-2	-2	-2	-1	-1	-1

-

A	H	C	-	N	I	R	V	S
A	I	C	I	N	-	R	C	K

MATCH
MO MATCH
GAP

DESAFIO VI: Utilizando la herramienta interactiva desarrolladas por el Grupo de Bioinformática de Freiburg probá distintos Gap penalties para el ejemplo propuesto y observá lo que ocurre.

Para el mismo par de secuencias utilizadas en el paso anterior

Con mismatch -1

Output:

D		A_1	I_2	C_3	I_4	N_5	R_6	C_7	K_8
	0	-1	-2	-3	-4	-5	-6	-7	-8
A_1	-1	1	0	-1	-2	-3	-4	-5	-6
H_2	-2	0	0	-1	-2	-3	-4	-5	-6
C_3	-3	-1	-1	1	0	-1	-2	-3	-4
N_4	-4	-2	-2	0	0	1	0	-1	-2
I_5	-5	-3	-1	-1	1	0	0	-1	-2
R_6	-6	-4	-2	-2	0	0	1	0	-1
V_7	-7	-5	-3	-3	-1	-1	0	0	-1
S_8	-8	-6	-4	-4	-2	-2	-1	-1	-1

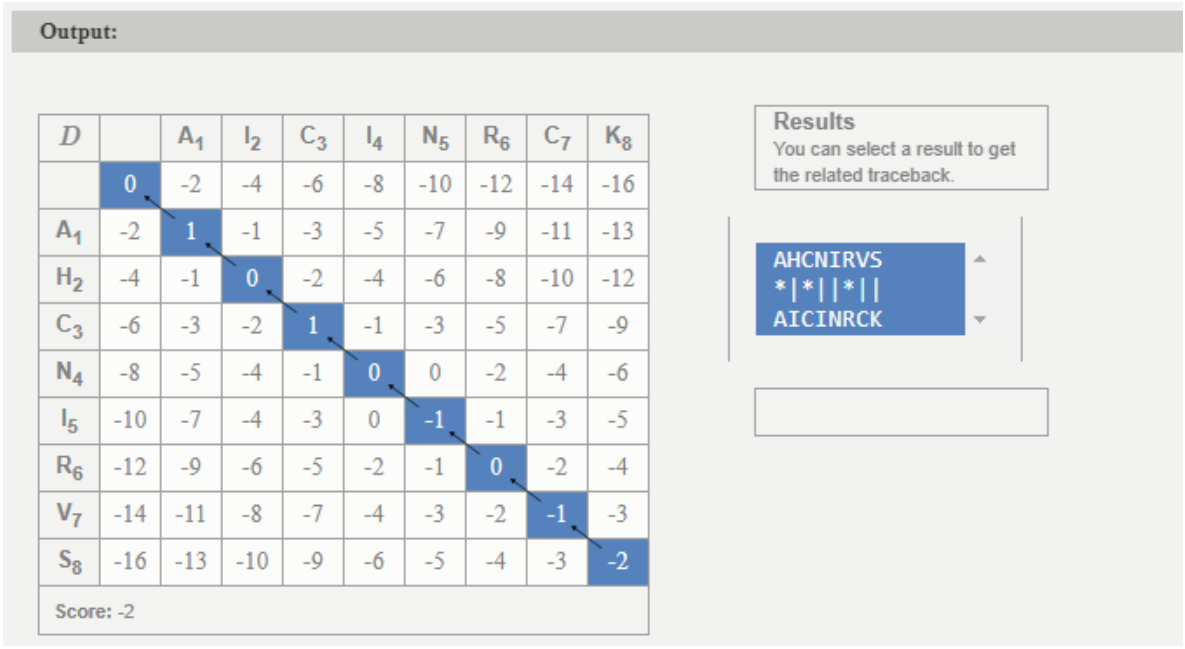
Score: -1

Results
You can select a result to get the related traceback.

```
AHC_NIRVS
*|* * *||
AICIN_RCK
```

```
AHCNI_RVS
*|* * *||
AIC_INRCK
```

Con mismatch -2



D		A_1	I_2	C_3	I_4	N_5	R_6	C_7	K_8
	0	-2	-4	-6	-8	-10	-12	-14	-16
A_1	-2	1	-1	-3	-5	-7	-9	-11	-13
H_2	-4	-1	0	-2	-4	-6	-8	-10	-12
C_3	-6	-3	-2	1	-1	-3	-5	-7	-9
N_4	-8	-5	-4	-1	0	0	-2	-4	-6
I_5	-10	-7	-4	-3	0	-1	-1	-3	-5
R_6	-12	-9	-6	-5	-2	-1	0	-2	-4
V_7	-14	-11	-8	-7	-4	-3	-2	-1	-3
S_8	-16	-13	-10	-9	-6	-5	-4	-3	-2

Se observa que se alteran las primeras filas y columnas numéricas con la modificación de la penalidad por gap. Es claro que toda la matriz se verá alterada ya que por recursión tomará los valores de las filas iniciales por la naturaleza del cálculo empleado. En este caso es Needleman-Wunsch.

DESAFIO VII: calculá el E-value y % identidad utilizando el programa Blast de la siguiente secuencia input usando 20000 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

VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTT
TKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILLISFLIFLIVG

Toda la secuencia:

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	major prion protein preproprotein Prp precursor [Homo sapiens]	282	282	100%	5e-97	100.00%	NP_000302.1
☐	major prion protein precursor [Mus musculus]	242	242	94%	3e-81	89.06%	NP_001265185.1
☐	prion-like protein doppel precursor [Mus musculus]	45.1	45.1	84%	3e-05	21.74%	NP_001119810.1
☐	prion protein gene complex [Mus musculus]	45.1	45.1	84%	3e-05	21.74%	NP_001265187.1
☐	prion-like protein doppel preproprotein [Homo sapiens]	37.7	37.7	76%	0.012	25.23%	NP_036541.2

Solo un fragmento de la secuencia anterior:

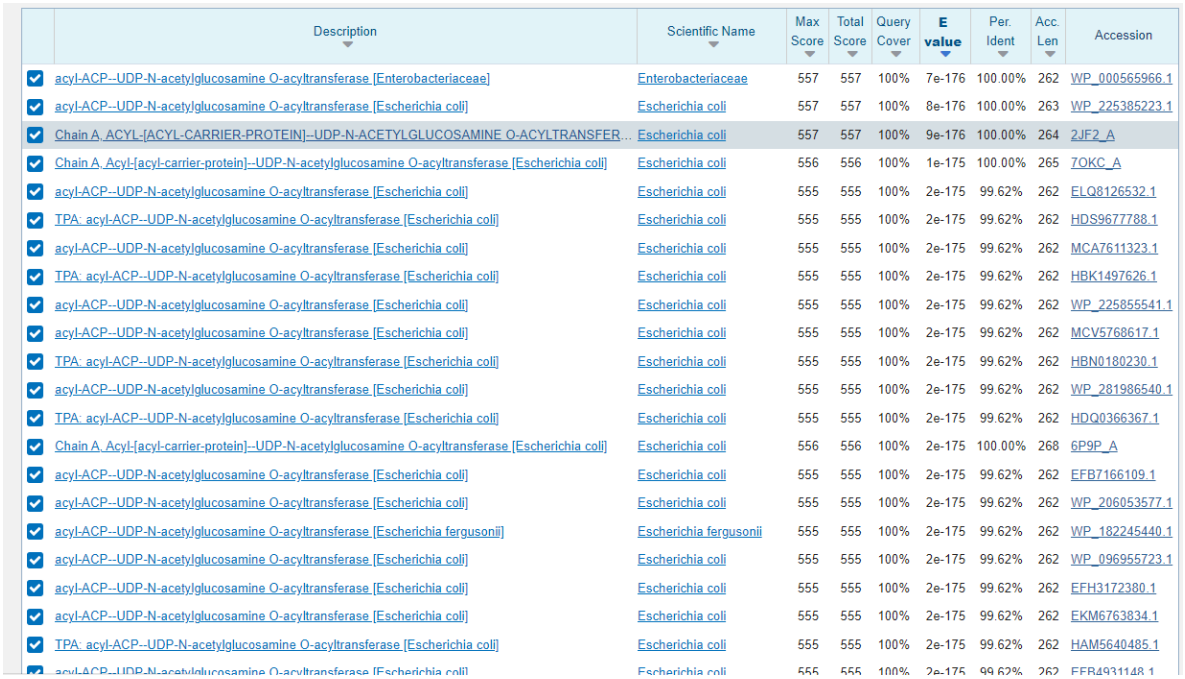
	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	major prion protein preproprotein Prp precursor [Homo sapiens]	106	106	100%	2e-29	100.00%	NP_000302.1
☐	major prion protein precursor [Mus musculus]	96.7	96.7	100%	1e-25	95.65%	NP_001265185.1

DESAFIO 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 sacas?

Se observa que para el fragmento utilizado el porcentaje de identidad se mantiene para los primeros casos, aunque el valor aumenta para [Mus Musculus], calculo que porque cambio la proporción entre el fragmento usado y la muestra de la base de datos. Lo mismo sucede para el Expected Value que en el primer caso es mucho mas alto.

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

MIDKSAFVHPTAIVEEGASIGANAHIGPFCIVGPHVEIGEGTVLKSHVVVNGHTKIGRDNEIYQFASIGEVNQDLK
YAGEPTRVEIGDRNRIRESVTIHRGTVQGGGLTKVGSDNLLMINAHIAHDCTVGNRCILANNATLAGHVSVDDF
AIIGGMTAVHQFCIIGAHVMVGGCSGVAQDVPPYVIAQGNHATPFGVNIEGLKRRGFSREAITAIRNAYKLIYRS
GKTLDEVKPEIAELAETYPEVKAFTDFFARSTRGLIR



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Enterobacteriaceae]	Enterobacteriaceae	557	557	100%	7e-176	100.00%	262	WP_000565966.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	557	557	100%	8e-176	100.00%	263	WP_225385223.1
Chain A, ACYL-[ACYL-CARRIER-PROTEIN]--UDP-N-ACETYLGLUCOSAMINE O-ACYLTRANSFER...	Escherichia coli	557	557	100%	9e-176	100.00%	264	2JF2_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	556	556	100%	1e-175	100.00%	265	7OKC_A
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	ELQ8126532.1
TPA: acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	HDS9677788.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	MCA7611323.1
TPA: acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	HBK1497626.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	WP_225855541.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	MCV5768617.1
TPA: acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	HBN0180230.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	WP_281986540.1
TPA: acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	HDQ0366367.1
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	556	556	100%	2e-175	100.00%	268	6P9P_A
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	EFB7166109.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	WP_206053577.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia fergusonii]	Escherichia fergusonii	555	555	100%	2e-175	99.62%	262	WP_182245440.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	WP_096955723.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	EFH3172380.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	EKM6763834.1
TPA: acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	HAM5640485.1
acyl-ACP--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	555	555	100%	2e-175	99.62%	262	EFB4931148.1

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

No se obtienen los mismos resultados. En la búsqueda de PDB hay muchos hits “Chain”

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Chain A, UDP N-ACETYLGLUCOSAMINE O-ACYLTRANSFERASE [Escherichia coli K-12]	Escherichia coli…	557	557	100%	1e-179	100.00%	262	1LXA_A
Chain A, ACYL-[ACYL-CARRIER-PROTEIN]--UDP-N-ACETYLGLUCOSAMINE O-ACYLTRANSFERASE [Escheric…	Escherichia coli	557	557	100%	1e-179	100.00%	264	2JF2_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	556	556	100%	2e-179	100.00%	265	7OKC_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Escherichia coli]	Escherichia coli	556	556	100%	3e-179	100.00%	268	6P9P_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Proteus mirabilis HI4320]	Proteus mirabilis…	422	422	100%	1e-128	72.66%	270	6OSS_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Acinetobacter baumannii]	Acinetobacter ba…	292	292	97%	2e-82	52.69%	265	4E6U_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Acinetobacter baumannii]	Acinetobacter ba…	292	292	97%	5e-82	52.69%	294	4E6T_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Moraxella catarrhalis BBH18]	Moraxella catarrh…	289	289	97%	2e-81	53.33%	257	5JXX_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Burkholderia thailandensis E264]	Burkholderia thail…	283	283	97%	5e-79	52.12%	283	4EQY_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Pseudomonas aeruginosa PA7]	Pseudomonas ae…	276	276	98%	4e-77	54.47%	258	5DEM_A
Chain A, Acyl-[acyl-carrier-protein]-UDP-N-acetylglucosamine O-acyltransferase [Pseudomonas aeruginosa PA7]	Pseudomonas ae…	276	276	98%	4e-77	54.47%	261	7OJ6_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Francisella tularensis subsp. nov…	Francisella tulare…	218	218	97%	5e-58	44.02%	280	5F42_A
Chain A, Putative acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Bacteroides fragilis NCT…	Bacteroides fragil…	201	201	100%	2e-52	42.97%	275	4R36_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Leptospira interrogans]	Leptospira interro…	197	197	96%	2e-51	43.02%	259	3HSQ_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Helicobacter pylori]	Helicobacter pylori	180	180	95%	3e-46	41.60%	270	1J2Z_A
Chain A, UDP-N-acetylglucosamine O-acyltransferase domain-containing protein [Arabidopsis thaliana]	Arabidopsis thali…	174	174	99%	6e-44	38.23%	305	3T57_A
Chain A, Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase [Campylobacter jejuni subsp. jeju…	Campylobacter je…	169	169	96%	1e-42	36.61%	266	3R0S_A
Chain A, UDP-N-acetylglucosamine acyltransferase [Psychrobacter cryohalolentis K5]	Psychrobacter cr…	69.4	69.4	83%	7e-12	27.15%	189	8E62_A
Chain A, Udp-3-o-[3-hydroxymyristoyl] Glucosamine N-acyltransferase [Chlamydia trachomatis]	Chlamydia tracho…	60.4	60.4	81%	6e-09	27.82%	374	2IU8_A
Chain A, UDP-3-O-acylglucosamine N-acyltransferase [Pseudomonas aeruginosa PAO1]	Pseudomonas ae…	59.6	80.2	69%	9e-09	32.80%	368	6UEC_A
Chain A, UDP-3-O-acylglucosamine N-acyltransferase [Pseudomonas aeruginosa PAO1]	Pseudomonas ae…	59.2	79.7	69%	1e-08	32.80%	369	6UED_A
Chain A, UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase [Pseudomonas aeruginosa]	Pseudomonas ae…	59.2	79.3	69%	1e-08	32.80%	372	3PMO_A