Tarea 1 – Bioinfo – 2021B

Pregunta 1

Ir a Blast (http://blast.ncbi.nlm.nih.gov/) en Genbank. Hacer una búsqueda de proteínas ("protein blast"), usando como "proteína" de consulta el nombre del alumno, limitado a los aminoácidos estándar (ACDEFGHIKLMNPQRSTVWY); por ejemplo, yo buscaría ANDRESMREIRA o ANDRESMQREIRA. Si trabajan de a dos, pueden escoger el nombre de uno, o alguna combinación, o básicamente lo que quieran.

Tomen nota de la base de datos que usan y de los parámetros de la búsqueda (abajo en "Algorithm parameters"; ojo que probablemente el servidor ajustará sus parámetros para acomodarse a la longitud -breve- de su query), de modo que yo eventualmente pueda repetir luego sus resultados.

Para el mejor match obtenido por la búsqueda:

- Muestre dos dotplots entre su query (en la forma que se usó en la búsqueda) y un segmento de la proteína encontrada, que incluya el match pero sea más largo que el query (digamos, 3 veces la longitud). El primer dotplot hágalo para las identidades (dotplot simple), y el segundo hágalo indicando dos identidades en ventana de largo 3.
- Muestre el alineamiento que dio BLAST entre su nombre y el segmento de la proteína.
- ¿Qué función cumple la proteína? [Para la función: trate de averiguar; por ejemplo, si es una “pitufasa”, trate de investigar que significa “pitufar”, o qué es la “pitufosis”]. ¿Se sabe con certeza (base experimental), o es una conjetura? ¿Se conoce su estructura 3D? ¿Está etiquetada como miembro de alguna familia de proteínas?
- El organismo en que la encontró: ¿qué es? (clasificación, descripción, imagen si es que existe [Use The Google!]).
- ¿Dónde y cómo está codificada la proteína? (Indique la ubicación de la CDS -"coding sequence"- de DNA que la codifica, y acaso está en la hebra primaria o secundaria, si es continua o tiene interrupciones, si está en cierto cromosoma, si es parte del genoma mitocondrial, etc.; **incluya** el código de acceso o URL de la secuencia de DNA).
- ¿Qué origen tienen los datos? (¿Proyecto de secuenciamiento masivo? ¿Paper específico sobre el gen? ¿Estudio filogenético?)

Considerando ahora además los 5 matches siguientes en su búsqueda inicial (por lo tanto, 6 secuencias encontradas) que sean en especies distintas (de modo que si hay más de uno en la misma especie, se ignoran las "repeticiones"),

- ¿En qué organismos están? ¿Son especies cercanas a las del primer match?
- Dibuje el "árbol familiar" que muestre las relaciones entre los 6 organismos, de acuerdo a la información en Genbank.

Pregunta 2

Programe en su lenguaje favorito. Necesitará (al menos) funciones que hagan lo siguiente:

- Generar una secuencia aleatoria de 200 bases equiprobables e independientes. Nos interesa la evolución de una especie alienígena en que las bases del DNA son 6, no 4: {A,C,G,T,B,D}.
- Una función que aplique una mutación a una secuencia; la mutación se escoge entre inserción, borrado y reemplazo de manera equiprobable, y su lugar de aplicación se elige al azar a lo largo de la secuencia. El borrado borra una letra, la inserción inserta una letra (escogida equiprobable), y el reemplazo reemplaza una letra por cualquiera de las otras 5 (de manera equiprobable).
- Una función que calcule la distancia de Levenshtein entre dos secuencias (implementando Needleman-Wunsch). Puede buscar y reciclar alguna online (las hay, especialmente en Python), pero asegúrese de que está funcionando con parámetros correctos.

Con esas funciones, hará lo siguiente:

a) Generar una secuencia, y aplicar M mutaciones; para M entre 0 y 300, grafique la relación entre M y D, donde D es la distancia de Levenshtein entre la secuencia final y la secuencia inicial.

b) Genere una secuencia, clónela, y a cada copia aplíquele M mutaciones (de modo que tendrá dos secuencias crecientemente distintas). Grafique la relación entre M y D', donde D' es la distancia entre las dos secuencias que están mutando.

c) Genere 10.000 pares de secuencias (largo 200 c/u) y evalúe su distancia de Levenshtein; haga un histograma de la distribución de estos valores, y calcule la media y σ.

d) Considerando (b) y (c), ¿por sobre qué valor de M diría usted que el parentesco entre las secuencias es indetectable (es decir, se entra en la *twilight zone*)? Tome en cuenta que cuando la distancia ya está dentro de la twilight zone, hace rato que la capacidad de juzgar parentesco ya se perdió.

Importante: en su entrega incluya no sólo los resultados sino también el código y posibles archivos intermedios. Si recicló alguna función hallada online (por ejemplo, para Needleman-Wunsch) inclúyala (o link).