Introducción a la Bioinformática

Trabajo Práctico - Enfermedad de Huntington

Primer Cuatrimestre de 2022

Integrantes

- Dallas, Tomás 56436
- La Mattina, Luca 57093
- Reyes, Santiago 58148
- Rolandelli, Alejandro 56644

Índice

Introducción

Objetivo del trabajo

El objetivo del presente trabajo práctico es poder adquirir las primeras habilidades en el campo de la Bioinformática. Para ello se eligió una enfermedad asociada a un gen en particular y se lo exploró siguiendo las direcciones de los ejercicios provistos por la cátedra.

Herramientas utilizadas

Se utilizó Python como lenguaje de programación para el desarrollo del trabajo, utilizando principalmente la librería Biopython. Además se utilizaron los siguientes programas:

- EMBOSS
- Docker

Requerimientos

Para ejecutar los scripts correspondientes a cada ejercicio se necesita disponer de ciertas dependencias y configuraciones iniciales. A fines prácticos se utilizó pipenv para el manejo de dependencias, las instrucciones para establecer el ambiente inicial se encuentran en el README del proyecto.

Enfermedad de Huntington

Descripción de la enfermedad

La enfermedad de Huntington es una enfermedad neurológica, hereditaria y degenerativa. La enfermedad ocasiona trastornos motrices, cognitivos y psiquiátricos con una amplia gama de signos y síntomas. Tiene una progresión muy lenta, normalmente de 15 a 20 años, pero existen variantes con progresión más acelerada. La expresión de los síntomas varía de paciente a paciente, especialmente en las etapas iniciales de la enfermedad.

Síntomas

Trastornos motrices

Presenta trastornos del movimiento, puede incluir tanto movimientos involuntarios como deterioro de los voluntarios. Por ejemplo:

- Movimientos espasmódicos o de contorsión involuntarios.
- Problemas musculares como rigidez o contracturas.
- Movimientos oculares lentos o anormales.
- Dificultad para hablar o tragar.

Trastornos cognitivos

Algunos de los signos de deterioro cognitivo relacionados con la enfermedad son:

- Dificultades para organizarse o enfocarse en tareas.
- Tendencia a quedarse sumido en pensamiento, conducta o acción.
- Falta de control de impulsos.
- Falta de conciencia sobre las conductas y habilidades propias.
- Lentitud para procesar pensamientos.
- Dificultad para aprender información nueva.

Trastornos psiquiátricos

El trastorno psiquiátrico más frecuente asociado con la enfermedad es la depresión, parece ocurrir debido a lesiones en el cerebro. Otros trastornos frecuentes asociados son:

- Trastorno obsesivo compulsivo
- Manía
- Trastorno bipolar

Etiología

La enfermedad se produce mediante un único factor hereditario, un defecto genético que se encuentra a nivel del cromosoma 4. El gen Huntingtin (HTT) es el que está relacionado a la enfermedad, y este produce una proteína llamada Huntingtina. Este se expresa en numerosos tejidos y si bien la función exacta de la proteína es desconocida, se cree que tiene un rol importante en las neuronas y es esencial para el correcto desarrollo de las mismas previo al nacimiento. Dentro de las células, la proteína se encuentra tanto en el núcleo como en el citoplasma y está asociada con funciones de señalización celular, transporte vesicular, endocitosis y de prevenir la apoptosis. Algunos estudios sugieren que tiene un rol en la reparación del ADN dañado.

Se desconocen exactamente las bases fisiopatológicas de la enfermedad, pero se cree que este exceso de tripletes CAG hace que las proteínas interaccionen entre sí de manera hidrofóbica, facilitando la formación de precipitados y acúmulos proteicos, especialmente en el cerebro. La gravedad de los síntomas es directamente proporcional a la cantidad de repeticiones del triplete, y esta es inversamente proporcional a la edad de presentación de la enfermedad. Como es el caso en este tipo de enfermedades de expansión de tripletes, cambios ligeros al número de repeticiones no produce la enfermedad, pero ese aumento se transmite a generaciones futuras, hasta finalmente inducir la enfermedad.

Efecto de la cantidad de repeticiones del triplete:

- Normal: 26 o menos repeticiones CAG.
- Intermedio: 27-35 repeticiones CAG. No tiene riesgo de desarrollar síntomas de EH, pero debido a la inestabilidad en la repetición de este triplete es posible que tenga hijos con una mayor cantidad de tripletes CAG que los ponga en el rango de la EH.
- Alto riesgo: 36-39 repeticiones CAG. La gente en este rango de repeticiones tiene un alto riesgo de presentar síntomas de EH, y al igual que en el caso anterior, aumenta el riesgo que sus futuras generaciones presenten la enfermedad.
- Afectado: Si un individuo presenta 40 o más repeticiones del triplete en el gen HTT este presentará síntomas de EH en algún momento de su vida. 60 o más repeticiones caen en la categoría de enfermedad de Huntington juvenil.

Una vez que la enfermedad está establecida, tiene una herencia autosómica dominante (cada descendiente tiene un 50% de heredar la enfermedad). La enfermedad además presenta una impronta genética de tipo paterno. La cantidad de repeticiones del triplete CAG se determina en la concepción y no suele cambiar, pero si el gen es heredado del padre, es probable que la cantidad de tripletes CAG aumente (respecto al del padre), mientras que si es heredado de la madre la cantidad de tripletes tiende a mantenerse estable.

Variantes

La enfermedad de Huntington juvenil es una forma de la enfermedad de Huntington caracterizada por la aparición de signos y síntomas antes de los 20 años y ocurre debido a un gran número de repeticiones del triplete CAG en el gen.

Procesamiento de Secuencias

Se procesó un archivo de secuencias Genbank de un mRNA del gen HTT. Para obtener el archivo de secuencias se utilizó la base de datos de genes del NCBI. El programa toma el archivo en formato GeneBank y analiza las secuencias de nucleótidos en sus registros. Luego, para cada una de ellas escribe un archivo en formato FASTA como <id>.fasta.

Para correr el programa se debe ejecutar el siguiente comando:
python gb_to_fasta.py [-h] -i I [-o O]
Argumentos:

-i I, --in-file I: Archivo de entrada GeneBank con extensión .gb
-o O, --out-dir O: Directorio de salida para las secuencias FASTA
el directorio de salida por defecto es out/fasta/

Para este ejercicio se encuentra disponible el archivo GeneBank correspondiente al gen HTT en data/genebank/HTT_NM_001388492_iso1.gb.

También se cuenta con un archivo data/genebank/multi_record.gb con 2 registros para probar que el programa funciona con varios registros.

BLAST

Modo de ejecución

Para este ejercicio se busca hacer un script que realice un BLAST de una o varias secuencias, escribiendo su salida en un archivo. El programa se conecta con el API del NCBI y realiza el BLAST para cada secuencia, escribiendo el resultado del informe en un archivo en formato texto.

Para correr el programa se debe ejecutar el siguiente comando
python blast.py [-h] -i I [-o O] -m {nucleotide,protein}
Argumentos:

-i I, --in-dir I: Directorio de entrada con secuencias FASTA
-o O, --out-dir O: Directorio de salida para los reportes BLAST de cada secuencia
-m {nucleotide, protein}: Modo de ejecución, nucleótido o proteína.
el directorio de salida por defecto es out/blast/

Se puede correr el programa utilizando como entrada el FASTA del ejercicio anterior, este se encuentra dentro del directorio results/fasta/NM_001388492.1.fasta

Análisis de resultados

El programa produce un archivo coincidente a los primeros 50 hits (secuencias de nucleótidos de mayor coincidencia) de correr un BLAST de un mRNA del gen HTT. Dentro de los primeros tres resultados se encuentra la secuencia del gen NM_001388492.1 como la más parecida, con un match del 100%. Esto es de esperarse, ya que esta es la variante del mRNA HTT que levantó del archivo GenBank. Como segunda opción encuentra NM_002111.8 que es otra variante, en este caso con 6 gaps y 0 mutaciones. Como tercera opción se encuentra la secuencia del gen de referencia que también es muy parecida, con 10 gaps nada más.

Varios de los resultados obtenidos corresponden a primates lo cual es razonable teniendo en cuenta que el genoma humano y el de los primates son muy parecidos. Los siguientes resultados son de chimpancés y bonobos que no solo entre sí como especie son muy similares, sino que son las especies vivas más relacionadas a los humanos, lo cual explica también el score tan alto para estos genes.

Limitaciones de implementación

Originalmente hubo un problema en la implementación del ejercicio ya que se estaba utilizando la librería Biopython para realizar el BLAST de manera remota, pero por un detalle de la implementación de la librería se estaba introduciendo un timeout que hacía que los resultados tarden varios minutos en aparecer, a diferencia de utilizar directamente la interfaz web. Por lo tanto para este ejercicio se utilizó el código de la librería pero eliminando esos timeouts para bajar el tiempo de ejecución a uno más razonable. El problema con esto es que solo se puede usar a esta velocidad durante fines de semana, sino el API devuelve un error así que al fin se usa la librería como viene con sus tiempos.

Multiple Sequence Alignment

Utilizando la búsqueda de blast realizada en el ejercicio anterior, se analizaron los FASTA de 3 especies distintas más la del Homo Sapiens y se realizaron sus posteriores alineamientos múltiples para cada secuencia de cada especie contra la del Homo Sapiens. Las 3 especies elegidas fueron:

- Gorila
- Chimpancé
- Zorro ártico

El alineamiento de secuencias múltiples es un procedimiento en el cual se puede representar y comparar dos o más secuencias de ARN para encontrar zonas de similitud. Tales zonas podrían indicar relaciones funcionales entre los genes o proteínas, como así también mutaciones. Del MSA obtenido se pueden inferir patrones genéticos que puedan determinar o no el mismo origen evolutivo para los genes, e incluso realizar un árbol filogenético para ver si en algún momento se compartió un antepasado común.

Se utilizó la herramienta online para realizar MSA hosteada en ebi.ac.uk, el algoritmo elegido fue el CLUSTAL debido a que corre de forma más rápida aunque con menor precisión que otros algoritmos para secuencias que no tengan mucha relación entre sí, pero dado que en su mayoría (Gorila y Chimpancé) las especies elegidas poseen relaciones muy fuertes con la del Homo Sapiens, se decidió por este algoritmo dada su velocidad.
Se desarrolló, además, un script para poder analizar la salida de cada análisis y poder ver con mayor profundidad las similitudes y/o diferencias entre las secuencias para cada par de especies. Estas salidas se encuentran en la carpeta ej3 del repositorio de Github, como también el script y los archivos de input.
La salida obtenida del script realizado para los 3 pares de MSA que se realizaron con el algoritmo de CLUSTAL se puede observar a continuación:

Gorilla
Secuencias con similitudes: 214
Matches totales: 3116
Matches con score >= 0.5 (match mayor al 50%) : 9
Matches con score < 0.5 (match menor al 50%) : 111

Pan homo (Chimpancé)
Secuencias con similitudes: 214
Matches totales: 3114
Matches con score >= 0.5 (match mayor al 50%) : 7
Matches con score < 0.5 (match menor al 50%) : 113

Vulpes (Zorro ártico)
Secuencias con similitudes: 214
Matches totales: 2872
Matches con score >= 0.5 (match mayor al 50%) : 125
Matches con score < 0.5 (match menor al 50%) : 162

Analizando la salida de cada MSA pudimos observar que entre el Gorila/Homo Sapiens y Chimpancé/Homo Sapiens hay muchísima similitud entre las secuencias, existen alrededor de 3100 matches totales (100% de score), y en menor medida matches con un score menor al 50%. Ahora bien, analizando el MSA para el zorro ártico se puede observar que los matches totales son 2800, lo que parece ser lógico dada la poca relación existente entre el Zorro y Homo Sapiens comparando con el Gorila y Chimpancé.

BLAST Output

El script utiliza la librería Biopython para parsear el archivo de entrada BLAST y, por cada reporte, si el pattern está dentro del título del alignment copia los datos del reporte en el archivo .txt de salida. Luego con el id del reporte, el tipo de retorno FASTA y si es proteína o nucleido realiza una búsqueda utilizando efetch de Entrez y lo escribe en el archivo de salida.

Para correr el programa se debe ejecutar el siguiente comando
python ej4/ej4.py [-h] -i I [-o O] -m {nucleotide,protein} -p pattern
Argumentos:

-i I, --in-file I: Archivo de entrada con secuencias FASTA
-o O, --out-file O: Archivo de salida para los resultados
-m {nucleotide, protein}: Modo de ejecución, nucleótido o proteína.
-p, --p: Patrón de búsqueda
el directorio de salida por defecto es out/blast-output/ej4.txt

Este requiere que el resultado de salida del blast sea XML, por lo que se generó una versión del mismo en XML. Se encuentra en results/blast-xml/NM_001388492.1.out

EMBOSS

Para esta aplicación se utilizó una solución dockerizada, ya que hubo problemas para correr EMBOSS en algunos de nuestros sistemas operativos. Para poder correr este script correctamente se necesita tener docker instalado y corriendo en la computadora.

El programa levanta el archivo de entrada de tipo fasta y utiliza la aplicación *transeq* de EMBOSS para generar una secuencia de proteínas posible. Esta se guarda en un archivo temporal y después se la utiliza como entrada de la aplicación *patmatmotifs* de EMBOSS para generar un análisis de las secuencias de proteínas obtenidas anteriormente.

Para correr el programa se debe ejecutar el siguiente comando
cd ej5 # Para entrar en la carpeta del ejercicio
python ej5.py [-h] -i I [-o O]
Argumentos:

-i I, --in-file I: Archivo de entrada FASTA
-o O, --out-file O: Archivo de salida patmatmotifs
el directorio de salida por defecto es ej5.patmatmotifs

Para probar este script conviene usar el archivo FASTA que se obtuvo previamente. Se debe ejecutar desde la carpeta ej5 dentro del directorio del proyecto ya que se utiliza el directorio data que se encuentra ahí adentro para hacer el traspaso de archivos entre el contenedor de docker y el host. Es posible que la primera ejecución del script parezca que se colgó, pero eso es porque está descargando las imágenes necesarias de docker y eso puede tomar un tiempo.

Trabajo con bases de datos biológicas

Gen y proteína de interés

Los siguientes son enlaces al gen y mRNA humano HTT que utilizamos para el trabajo:

- https://www.ncbi.nlm.nih.gov/gene/3064
- https://www.ncbi.nlm.nih.gov/nuccore/NM_001388492.1

Se decidió investigar esta enfermedad ya que era una enfermedad genética que conocíamos de nombre por la serie "HOUSE". Uno de los personajes "Remy Hadley" padecía de la enfermedad y fue un factor importante en la serie. La explicación de la proteína se encuentra al principio del informe.

Genes homólogos

Se encontraron los siguientes genes homólogos en HomoloGene

HTT, *H.sapiens*
huntingtin
HTT, *P.troglodytes*
huntingtin
HTT, *C.lupus*
huntingtin
HTT, *B.taurus*
huntingtin
Htt, *M.musculus*
huntingtin
Htt, *R.norvegicus*
huntingtin
HTT, *G.gallus*
huntingtin
htt, *X.tropicalis*
huntingtin
htt, *D.rerio*
huntingtin

Utilizando Ensembl encontramos la siguientes lista que contiene mucha más información:

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
Primates (*23 species*) Humans and other primates	☐	23	0	0	0
Rodents and related species (*24 species*) Rodents, lagomorphs and tree shrews	☐	23	0	0	1
Laurasiatheria (*38 species*) Carnivores, ungulates and insectivores	☐	35	1	0	2
Placental Mammals (*90 species*) All placental mammals	☐	86	1	0	3
Sauropsida (*27 species*) Birds and Reptiles	☐	26	1	0	0
Fish (*65 species*) Ray-finned fishes	☐	54	7	0	4
All (*200 species*) All species, including invertebrates	☑	180	11	0	9

notamos que el gen se encuentra en una variedad muy amplia de especies, aunque está mayormente concentrado en mamíferos.

Splicing

Utilizando la base de datos de Ensembl encontramos **23 variantes de splicing.** Los resultados son visibles en el siguiente link y a continuación:r

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000355072.11	HTT-201	13472	3142aa	Protein coding	CCDS43206	P42858	NM_001388492.1	MANE Select v0.95 Ensembl Canonical GENCODE basic APPRIS P2 TSL:1
ENST00000681528.1	HTT-223	14033	3086aa	Protein coding		A0A7P0TAC5	-	GENCODE basic APPRIS ALT2
ENST00000680956.1	HTT-222	13943	3056aa	Protein coding		A0A7P0TA78	-	GENCODE basic APPRIS ALT2
ENST00000509618.1	HTT-205	429	112aa	Protein coding		H0YA07	-	TSL:3 CDS 5' incomplete
ENST00000649131.1	HTT-215	289	97aa	Protein coding		A0A3B3ISR3	-	CDS 5' and 3' incomplete
ENST00000680360.1	HTT-221	13984	2191aa	Nonsense mediated decay		A0A7P0Z417	-	-
ENST00000680239.1	HTT-219	13834	2880aa	Nonsense mediated decay		A0A7P0TAN5	-	-
ENST00000650588.1	HTT-217	503	75aa	Nonsense mediated decay		A0A3B3IU25	-	CDS 5' incomplete
ENST00000650595.1	HTT-218	390	75aa	Nonsense mediated decay		A0A3B3IU25	-	CDS 5' incomplete
ENST00000647962.1	HTT-213	1324	No protein	Processed transcript		-	-	-
ENST00000649900.1	HTT-216	587	No protein	Processed transcript		-	-	-
ENST00000513806.1	HTT-212	432	No protein	Processed transcript		-	-	TSL:5
ENST00000648150.1	HTT-214	331	No protein	Processed transcript		-	-	-
ENST00000510626.5	HTT-207	14438	No protein	Retained intron		-	-	TSL:1
ENST00000680291.1	HTT-220	7184	No protein	Retained intron		-	-	-
ENST00000506137.1	HTT-202	781	No protein	Retained intron		-	-	TSL:3
ENST00000509751.1	HTT-206	723	No protein	Retained intron		-	-	TSL:3
ENST00000512909.1	HTT-209	611	No protein	Retained intron		-	-	TSL:3
ENST00000512068.1	HTT-208	401	No protein	Retained intron		-	-	TSL:3
ENST00000508321.1	HTT-203	386	No protein	Retained intron		-	-	TSL:3
ENST00000509043.1	HTT-204	380	No protein	Retained intron		-	-	TSL:2
ENST00000513326.5	HTT-210	373	No protein	Retained intron		-	-	TSL:2
ENST00000513639.5	HTT-211	259	No protein	Retained intron		-	-	TSL:2

Analizando la tabla encontramos que más de la mitad de las entradas no poseen ORFs o no son proteínas; las restantes, 5 de ellas poseen funcionalidades de transporte intracelular (Ejemplo) mientras que para las restantes no se encontró ninguna funcionalidad.

Interacciones

Se encontraron 539 interacciones, aunque de esas 539, sólo alrededor de 60 tienen una descripción apropiada de cada interacción.

Se destaca la interacción entre CBP (CREB binding protein) y la proteína Huntinting. Esta interacción reprime la transcripción de los promotores p53, p21 y MDR-1, lo que da lugar a una disfunción neuronal y muerte de las células que contienen la enfermedad.