Introducción a la Bioinformática

Trabajo Práctico Final

Alumno: Martinez Correa, Facundo Nahuel

Legajo: 49139

Introducción

Para este Trabajo Práctico Final se decidió analizar la enfermedad de Charcot - Marie - Tooth. Esta tiene la cualidad de ser ser un síndrome de características totalmente genéticas, resultado de la mutación aberrante de ciertos genes.

Existen nueve clasificaciones distintas de esta enfermedad que difieren en el gen o la proteína afectada. En particular en este trabajo se tratará la variedad CMTX1, producto de una mutación en el gen GJB1 que sintetiza la proteína *Unión gap beta-1* (GJB1 por sus siglas en inglés). La misma pertenece a una familia de proteínas que se encargan de regular y controlar la transferencia de señales de comunicación entre las membranas de las células, principalmente en el hígado y en el sistema nervioso periférico.
Mutaciones en este gen en particular pueden resultar en neuropatías periféricas, desmielinización de los oligodendrocitos y las células de Schwann, causando transmisión diferida de la comunicación nerviosa del sistema nervioso periférico, lo que acaba con síntomas tales como atrofia muscular y problemas de sensibilidad en las extremidades de los miembros.

Ejercicios

Los scripts con los que fueron realizados los ejercicios se encuentran en el siguiente repositorio de Github:

https://github.com/fnmartinez/bionifo

El mismo es de público acceso.

Todos los scripts de PERL contienen documentación interna y explicación de cómo ejecutarlos.

Ejercicio 1

Para este ejercicio se descargó el archivo en formato GenBank con el reporte completo del mRNA del gen GJB1 de la siguiente dirección: https://www.ncbi.nlm.nih.gov/nuccore/NC_000023.11?report=genbank&from=71215212&to=71225215

Y fue guardado en el archivo gjb1.gb, dentro del directorio input/ex1. Luego se ejecutó la siguiente línea de comando:

```
$> ./src/ex1.pm -input-file input/ex1/gjb1.gb
-output-directory output/ex1/
```

Con lo que se obtuvieron los seis ORF del gen en formato FASTA. Además se generó un séptimo archivo con el ORF más largo, siendo éste el de mayor probabilidad de contener la cadena de aminoácidos de la proteína que sintetiza.

```
$> ls output/ex1/

10004_gjb1-fwd-1.fas  gjb1-fwd-1.fas  gjb1-fwd-2.fas
gjb1-fwd-3.fas  gjb1-rev-1.fas  gjb1-rev-2.fas
gjb1-rev-3.fas
```

Ejercicio 2

Parte A

Usando el resultado del ejercicio uno, se procedió a usar el script ex2a.sh para usar el programa blastp de forma local, con el algoritmo BLAST+ y la

base de datos de swissprot. Si se tienen correctamente instaladas las dependencias de BioPerl para correr de forma local, se puede usar el el script ex2a.sh. Si no, siempre se puede usar este mismo último script con la opción --remote para que lo ejecute de forma remota.

El script mencionado se ejecutó de la siguiente manera:

```
$> ./src/ex2a.sh output/ex1/10004_gjb1-fwd-1.fas >
output/ex2a/blast.out
```

Esto brindó como resultado una lista de las proteínas que más probabilidades tenían de ser sintetizadas por el ORF suministrado.

El resultado de dicho blast puede ser encontrado en el repostiorio.

Parte B

Tomando los 10 mejores matches de proteínas, listados a continuación:

Sequence	Score (Bits)	EValue
P08034.1	586	0.0
Q60HF7.1	584	0.0
Q6WGK6.1	579	0.0
P08033.1	579	0.0
O18968.1	549	3e-176
P08983.2	400	1e-124
A2VE67.1	344	2e-105
P46691.1	341	2e-104
O93533.1	342	5e-104
Q8MIT9.1	340	5e-104

Se procedió a realizar un alineamiento multisecuencia con la finalidad de entender la evolución de la proteína. Para ello se utilizó el sitio de Clustal Omega, del EMBL-EBI, que está hosteado en la siguiente dirección:

http://www.ebi.ac.uk/Tools/msa/clustalo/

El resultado de dicho alineamiento puede verse en el anexo de este informe

como en el repositorio.

Ejercicio 3

Se creó un script para analizar la salida del BLAST, mediante la búsqueda de patrones en las secuencias encontradas. El mismo puede ser utilizado desde la consola de la siguiente manera:

```
$> ./src/ex3.pm --input-file blast.out --pattern arabidopsis
```

Ejercicio 4

Se creó un script para verificar con otro set de herramientas los análisis posteriores. Para esto se instaló localmente EMBOSS y con el uso de este script se pretendío llegar al mismo resultado que antes. Es decir, con un archivo FASTA con la secuencia del mRNA, encontrar el ORF que sintetiza la proteína y luego encontrarla en la base de datos de PROSITE. Para esto se usaron los programas getorf y patmatmotif.
El script puede ser corrido de la siguiente manera:

```
$> ./src/ex4.sh input/ex4/gjb1.fasta output/ex4/report
```

Ejercicio 5

A. El gen de interés, como fue mencionado en la introducción, es el GJB1:
   https://www.ncbi.nlm.nih.gov/gene/2705
   Este, sintetiza la proteína GJB1 o Unión Gap beta-1 proteína:
   https://www.ncbi.nlm.nih.gov/protein/AAH39198.1
   Fue elegida por ser la causante un tipo de Síndrome de Charcot-Marie-Tooth.
B. Se conocen genes y proteínas homólogas en todos los vertebrados ya que es común entre ellos.
   En HomoloGene se nos hace referencia a otros genes homólogos de otras especies, proteínas que estos generan, etc. Pero todas son referencias a sitios externos. Mientras que Ensembl tiene datos propios de estos genes y algunos links externos.
C. Existen cuatro formas alternativas de splicing, de los cuales todos se expresan. Las tres primeras GJB1-001, GJB1-002 y GJB1-003, todas expresan la misma proteína (P08034). Mientras que el splice GJB1-004 expresa la proteína (C9JWU8) que sirve en la comunicación celular.
   NCBI no muestra alternativas de splicing, en tanto que Ensembl sí.
D. Según NCBI hay catorce interacciones conocidas, en tanto que en en

UniProt se indican ocho, de las cuales siete ya estaban informadas en NCBI. No parece haber ningún patrón.

E. La proteína forma parte del sistema de comunicación de la célula, ya que interviene en procesos tales como el transporte entre células, la señalización entre célula y célula. Y forma parte del retículo endoplasmático.
F. La proteína se ve involucrada en:
    - La oligomerización de los conexines a conexones
    - El transporte de conexines a través del pasaje excretor
    - Ensamblaje de las juntas Gap
G. De acuerdo a ClinVar, todas las variaciones que no involucran a otros genes están directamente relacionadas con el Síndrome de Charcot Marie Tooth relacionado al cromosoma X.

Anexo

Resultado del Alineamiento Multisecuencia

Formato FASTA

```
>O93533.1 RecName: Full=Gap junction beta-6 protein; AltName:
Full=Connexin-31; Short=Cx31
MDWGALQTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAERVWGDEQDDFICNTLQPGC
KNVCYDHFFPISHIRLWALQLIFVSTPALLVAMHVAYRRHEKKRQFRKGDQKCEYKDIEE
IRTQRFRIEGTLWWTYTCSIFFRLVFEAVFMYAFYFMYDGFRMPRLMKCSAWPCPNTVDC
FVSRPTEKTVFTIFMIAVSSICILLNVAELCYLLTKFFLRRSRKAGNQKHHP-----NHE
NKEETKQNEMNELISDSCQNTVIGFTSS-----------------
>A2VE67.1 RecName: Full=Gap junction beta-2 protein; AltName:
Full=Connexin-26; Short=Cx26
MDWGGLHTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAKEVWGDEQADFVCNTLQPGC
KNVCYDHYFPISHIRLWALQLIFVSTPALLVAMHVAYYRHEKKRKFIRGEIKTEFKDIEE
IKKQKVRIEGSLWWTYTGSIFFRVIFEAAFMYVFYVMYDGFAMQRLVKCNAWPCPNTVDC
FVSRPTEKTVFTVFMIAVSGICILLNVTELCYLLIRFCSGKSKKPV--------------
---------------------------------------------
>Q8MIT9.1 RecName: Full=Gap junction beta-2 protein; AltName:
Full=Connexin-26; Short=Cx26
MDWGALQTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAKEVWGDEQADFVCNTLQPGC
KNVCYDHYFPISHIRLWALQLIFVSTPALLVAMHVAYRRHEKKRKFIKGEIKSEFKDIEE
IKTQKVRIEGSLWWTYTSSIFFRVIFEAAFMYVFYVMYDGFSMQRLVKCNAWPCPNTVDC
FVSRPTEKTVFTVFMIAVSGICILLNVTELCYLLIRYCSGRSKKPV--------------
---------------------------------------------
>P08983.2 RecName: Full=Gap junction beta-1 protein; AltName:
Full=Connexin-30; Short=Cx30
MNWAGLYAILSGVNRHSTSIGRIWLSVVFIFRIMVLVAAAESVWGDEKSAFTCNTQQPGC
NSVCYDHFFPISHIRLWALQLIIVSTPALLVAMHVAHLQHQEKKELRLS-RHVKDQELAE
VKKHKVKISGTLWWTYISSVFFRIIFEAAFMYIFYLIYPGYSMIRLLKCDAYPCPNTVDC
FVSRPTEKTIFTVFMLVASGVCIVLNVAEVFFLIAQACTRRARRHRDSGS----------
ISKEHQQNEMNLLITGG-----SIIKRSAGQ-----EKGDHCSTS
>O18968.1 RecName: Full=Gap junction beta-1 protein; AltName:
Full=Connexin-32; Short=Cx32
MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC
NSVCYDHFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE
VKRHKVHISGTLWWTYVISVVFRLLFEAAFMYVFYLLYPGYAMVRLVKCDAYPCPNTVDC
FVSRPTEKTIFTVFMLAASGICIILNVAEVVYLIFRACARRAQRRSNPPSRKGSGGFGHR
LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC
>Q6WGK6.1 RecName: Full=Gap junction beta-1 protein; AltName:
Full=Connexin-32; Short=Cx32
MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC
NSVCYDHFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPIHLEE
VKRHKVHISGTLWWTYVISVVFRLLFEAAFMYVFYLLYPGYAMVRLVKCDAYPCPNTVDC
FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIVRACARRAQRRSNPPSRKGS-GFGHR
LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC
>P08033.1 RecName: Full=Gap junction beta-1 protein; AltName:
Full=Connexin-32; Short=Cx32; AltName: Full=GAP junction 28 kDa liver prote
MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC
NSVCYDHFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE
VKRHKVHISGTLWWTYVISVVFRLLFEAVFMYVFYLLYPGYAMVRLVKCEAFPCPNTVDC
FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIIRACARRAQRRSNPPSRKGS-GFGHR
```

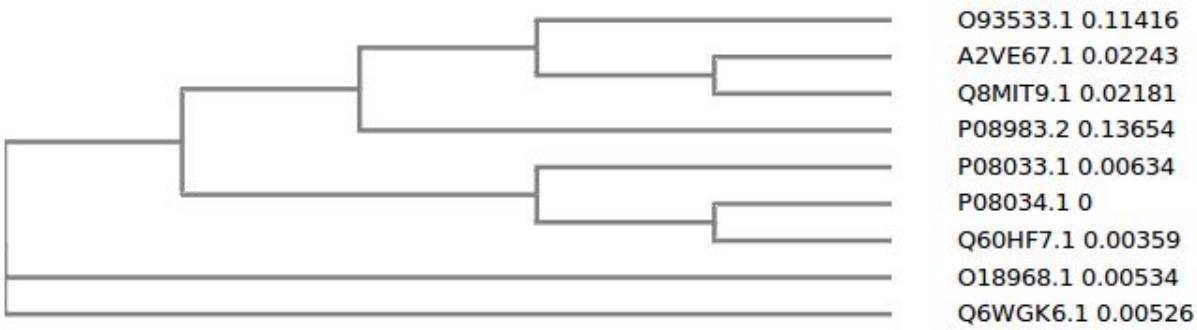
LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC
>P08034.1 RecName: Full=Gap junction beta-1 protein; AltName:
Full=Connexin-32; Short=Cx32; AltName: Full=GAP junction 28 kDa liver prote
MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC
NSVCYDQFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE
VKRHKVHISGTLWWTYVISVVFRLLFEAVFMYVFYLLYPGYAMVRLVKCDVYPCPNTVDC
FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIIRACARRAQRRSNPPSRKGS-GFGHR
LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC
>Q60HF7.1 RecName: Full=Gap junction beta-1 protein; AltName:
Full=Connexin-32; Short=Cx32
MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC
NSVCYDQFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE
VKRHKVHISGTLWWAYVISVVFRLLFEAVFMYVFYLLYPGYAMVRLVKCDVYPCPNTVDC
FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIIRACARRAQRRSNPPSRKGS-GFGHR
LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC

Formato CLUSTALO numerado

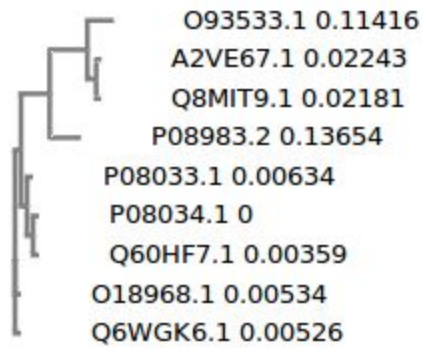
```
O93533.1      MDWGALQTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAERVWGDEQDDFICNTLQPGC    60
A2VE67.1      MDWGGLHTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAKEVWGDEQADFVCNTLQPGC    60
Q8MIT9.1      MDWGALQTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAKEVWGDEQADFVCNTLQPGC    60
P08983.2      MNWAGLYAILSGVNRHSTSIGRIWLSVVFIFRIMVLVAAAESVWGDEKSAFTCNTQQPGC    60
O18968.1      MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC    60
Q6WGK6.1      MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC    60
P08033.1      MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC    60
P08034.1      MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC    60
Q60HF7.1      MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAAESVWGDEKSSFICNTLQPGC    60
              *:* .* ::*.***:***:**::**:*:******:**.**: *****:  * *** ****

O93533.1      KNVCYDHFFPISHIRLWALQLIFVSTPALLVAMHVAYRRHEKKRQFRKGDQKCEYKDIEE    120
A2VE67.1      KNVCYDHYFPISHIRLWALQLIFVSTPALLVAMHVAYYRHEKKRKFIRGEIKTEFKDIEE    120
Q8MIT9.1      KNVCYDHYFPISHIRLWALQLIFVSTPALLVAMHVAYRRHEKKRKFIKGEIKSEFKDIEE    120
P08983.2      NSVCYDHFFPISHIRLWALQLIIVSTPALLVAMHVAHLQHQEKKELRLS-RHVKDQELAE    119
O18968.1      NSVCYDHFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE    119
Q6WGK6.1      NSVCYDHFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPIHLEE    119
P08033.1      NSVCYDHFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE    119
P08034.1      NSVCYDQFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE    119
Q60HF7.1      NSVCYDQFFPISHVRLWSLQLILVSTPALLVAMHVAHQQHIEKKMLRLE-GHGDPLHLEE    119
              :.****::*****:***:****:*************: :* :*: :       : .  .: *

O93533.1      IRTQRFRIEGTLWWTYTCSIFFRLVFEAVFMYAFYFMYDGFRMPRLMKCSAWPCPNTVDC    180
A2VE67.1      IKKQKVRIEGSLWWTYTGSIFFRVIFEAAFMYVFYVMYDGFAMQRLVKCNAWPCPNTVDC    180
Q8MIT9.1      IKTQKVRIEGSLWWTYTSSIFFRVIFEAAFMYVFYVMYDGFSMQRLVKCNAWPCPNTVDC    180
P08983.2      VKKHKVKISGTLWWTYISSVFFRIIFEAAFMYIFYLIYPGYSMIRLLKCDAYPCPNTVDC    179
O18968.1      VKRHKVHISGTLWWTYVISVVFRLLFEAAFMYVFYLLYPGYAMVRLVKCDAYPCPNTVDC    179
Q6WGK6.1      VKRHKVHISGTLWWTYVISVVFRLLFEAAFMYVFYLLYPGYAMVRLVKCDAYPCPNTVDC    179
P08033.1      VKRHKVHISGTLWWTYVISVVFRLLFEAVFMYVFYLLYPGYAMVRLVKCEAFPCPNTVDC    179
P08034.1      VKRHKVHISGTLWWTYVISVVFRLLFEAVFMYVFYLLYPGYAMVRLVKCDVYPCPNTVDC    179
Q60HF7.1      VKRHKVHISGTLWWAYVISVVFRLLFEAVFMYVFYLLYPGYAMVRLVKCDVYPCPNTVDC    179
              :: ::.:*.*:***:*  *:.**::***.*** **.:* *: * **:**..:********

O93533.1      FVSRPTEKTVFTIFMIAVSSICILLNVAELCYLLTKFFLRRSRKAGNQKHHP-----NHE    235
A2VE67.1      FVSRPTEKTVFTVFMIAVSGICILLNVTELCYLLIRFCSGKSKKPV--------------    226
Q8MIT9.1      FVSRPTEKTVFTVFMIAVSGICILLNVTELCYLLIRYCSGRSKKPV--------------    226
P08983.2      FVSRPTEKTIFTVFMLVASGVCIVLNVAEVFFLIAQACTRRARRHRDSGS-----------    229
O18968.1      FVSRPTEKTIFTVFMLAASGICIILNVAEVVYLIFRACARRAQRRSNPPSRKGSGGFGHR    239
Q6WGK6.1      FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIVRACARRAQRRSNPPSRKGS-GFGHR    238
P08033.1      FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIIRACARRAQRRSNPPSRKGS-GFGHR    238
P08034.1      FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIIRACARRAQRRSNPPSRKGS-GFGHR    238
Q60HF7.1      FVSRPTEKTVFTVFMLAASGICIILNVAEVVYLIIRACARRAQRRSNPPSRKGS-GFGHR    238
              *********:**:**:..*.:**:***:*: :*: :    ::::

O93533.1      NKEETKQNEMNELISDSCQNTVIGFTSS-----------------    263
A2VE67.1      ---------------------------------------------    226
Q8MIT9.1      ---------------------------------------------    226
P08983.2      ISKEHQQNEMNLLITGG-----SIIKRSAGQ-----EKGDHCSTS    264
O18968.1      LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC    284
Q6WGK6.1      LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC    283
P08033.1      LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC    283
P08034.1      LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC    283
Q60HF7.1      LSPEYKQNEINKLLSEQDGSLKDILRRSPGTGAGLAEKSDRCSAC    283
```

Árbol Filogenético

Cladograma



Real



Data

```
(
(
(
(
O93533.1:0.11416,
(
A2VE67.1:0.02243,
Q8MIT9.1:0.02181)
:0.03849)
:0.16377,
P08983.2:0.13654)
:0.12477,
(
P08033.1:0.00634,
(
P08034.1:0.00000,
Q60HF7.1:0.00359)
:0.00779)
:0.00517)
:0.00013,
O18968.1:0.00534,
Q6WGK6.1:0.00526);
```