



BIOINFORMÁTICA - 16.21
INSTITUTO TECNOLÓGICO DE BUENOS AIRES

TP CUATRIMESTRAL

Profesores

Mariano Nicolás BREGLIA
Adrián Federico PEREZ

GRUPO 3

Alumnos

Agustín LUNA SIMONDI	62053
Sofia BOUZO	61522
Josue Francisco LASZESKI	62502
Sebastian BERG WÖHLK	67173

Fecha de entrega: 18 de Noviembre de 2024

Índice

1. Objetivo	3
2. Herramientas, Tecnologías y Metodologías	3
2.1. Metodologías ágiles	3
2.2. Herramientas Informáticas	4
3. Introducción	4
3.1. Síndrome de Marfan	5
4. Ejercicios	5
4.1. Ejercicio 1: Procesamiento de secuencia	5
4.2. Ejercicio 2: BLAST	6
4.2.1. Interpretación de los resultados de BLAST	6
4.3. Ejercicio 3: Alineamiento Múltiple de Secuencias (MSA)	7
4.3.1. Resultados del MSA	7
4.4. Ejercicio 4: EMBOSS	7
4.5. Ejercicio 5: Primers	7
4.6. Ejercicio 6: Bases de Datos biológicas	8
4.6.1. Genes homólogos	9
4.6.2. Splicing alternativo	9
4.6.3. Interacción con otras proteínas	10
4.6.4. Ontología del gen	11
4.6.5. Vías metabólicas	12
4.6.6. Variantes genéticas	12
5. Conclusiones	12
6. Referencias	14
A. Anexo	15

1. Objetivo

Para este trabajo cuatrimestral se busca no solamente incorporar los conocimientos impartidos en la asignatura, sino además utilizar herramientas y tecnologías útiles tanto en el ámbito académico como laboral. De esta manera, se simula una situación real, en donde la consigna a resolver toma el papel de un trabajo propiamente dicho. Asimismo, se asume el rol de un equipo de empleados dispuestos a llevar a resolver los requisitos de nuestros **Stakeholders**. En este caso, la cátedra.

Por otro lado, se propone también utilizar metodologías que se emplean en el mundo laboral, lo cual permitirá tanto la división de tareas como la organización mismas. A su vez, en este trabajo tiene su participación **Mila Langone**, quien toma el rol de la **Product Owner (PO)**. Asume el rol de intermediaria entre el equipo y los Stakeholders, a la vez que ayuda a guiar en el uso de las metodologías ágiles.

Por último, al resolverse la mayoría de los ejercicios mediante programación, se busca que las soluciones sean **reproducibles y parametrizables**, de modo que terceros puedan usar el código sin inconvenientes (realizando las instalaciones y configuraciones pertinentes). Por esta razón, no solamente se hace un enfoque en efectivamente abordar cada ejercicio, sino que también que estos puedan ser reutilizados sin ningún problema, arribando a una solución fiable con un input variable.

De esta manera, como producción final de este trabajo se busca desarrollar un repositorio en **GitHub**, accesible para terceros y de uso totalmente libre. Adicionalmente, el repositorio se presenta debidamente ordenado, separando las carpetas de las configuraciones (*config*), la entrada inicial (*input*) y los códigos en si (*source*). A su vez cada código se acompaña de un **archivo Markdown** que explica tanto el propósito del ejercicio, así como las diversas funcionalidades implementadas, los *inputs* esperados, y sus respectivos *outputs*. Referencias para la presentación final y el repositorio pueden encontrarse en el Anexo del presente informe.

2. Herramientas, Tecnologías y Metodologías

2.1. Metodologías ágiles

Como se explica previamente, una de las metas de este trabajo cuatrimestral es incorporar habilidades pertenecientes al ámbito laboral, como lo son las metodologías ágiles. Se trata de un conjunto de enfoques y prácticas de gestión de proyectos que buscan optimizar la colaboración, adaptabilidad, eficiencia y flexibilidad [1]. Se emplea para establecer una organización para equipos en trabajos de varias etapas, estableciendo un entorno dinámico, donde las tareas pueden modificarse sobre la marcha.

En este caso, se lleva una metodología de **Scrum**, en donde el PO guía en cómo llevar a cabo dicha propuesta. Principalmente se realizaron reuniones cortas pre-estipuladas denominadas "daily", en donde el PO orienta en cara a los objetivos y tareas a cumplir. A su vez, uno de los integrantes del equipo asume el rol del **Scrum Master**, cuya función es delegar y dividir las diferentes tareas.

Para la organización de tareas y tiempos se construye un **Backlog**, enmarcado por **Sprints**. El Backlog permite tener en consideración el listado completo de tareas pendientes. Las entradas se añaden al Backlog de acuerdo a lo discutido en las daily, tras identificar la estructura de cada actividad. Por otro lado, el Sprint es la realización de cada una de estas tareas en un ciclo, generalmente corto, que ronda entre las 2 y 4 semanas.

Como herramienta para implementar Scrum se decide emplear **Notion**, tanto para el armado del Backlog como para la ejecución de los Sprints. Notion es una plataforma práctica y sencilla de utilizar, que permite que los miembros del equipo consulten el progreso del proyecto y añadir trazabilidad a la realización del trabajo. Algo importante a mencionar es que Notion también

otorga la posibilidad de incorporar herramientas adicionales, tales como **Kanban**, que divide las tareas en un tablero para gestionar el flujo de trabajo. En Kanban, se subdividen las tareas en las categorías de *Pendiente*, *En Progreso* y *Completado*.

Una vez completados los Sprints, se realiza con el PO el **Review** y **Retrospective**. En el Review, como equipo se realiza una revisión de lo que se ha logrado, evidenciando los resultados obtenidos. En la Retrospective, por otro lado, se toma un enfoque más orientado hacia el proceso, evaluando el trabajo del equipo más allá del valor agregado por el trabajo realizado.

2.2. Herramientas Informáticas

El desarrollo y ejecución de los ejercicios planteados se realiza dentro de un entorno de programación Linux, utilizando Bash como intérprete de consola y Python como lenguaje de programación para cada script. Es necesario contar con los siguientes comandos, o similares que suplan igual función:

- `wget`, para descargar archivos pertinentes
- `gunzip`, para descomprimir archivos obtenidos por `wget`

A su vez, es requisito instalar los siguientes paquetes, que deben de poder ser llamados desde archivos `.sh` interpretados por Bash:

- Muscle [2]
- Blast [3]
- Emboss Suite [4]

Por último, en parte de los scripts implementados por Python se utilizan las funciones de la librería **BioPython**, que debe ser instalada para poder utilizarse.

Para una descripción más detallada de las dependencias, se recomienda consultar el repositorio del proyecto. Finalmente, es importante mencionar que el código podría ser ejecutado desde otros sistemas operativos, siempre y cuando se encuentre con un terminal bash y los comandos utilizados correctamente configurados. No obstante, se recomienda utilizar máquinas virtuales, tales como VMBox o Windows Subsystem for Linux (WSL).

3. Introducción

Para poder comenzar con el trabajo el primer paso fue seleccionar una enfermedad asociada a trastornos genéticos en humanos. Para esto mismo, usamos la base de datos **Online Mendelian Inheritance in Man (OMIM)** [5], que recopila información detallada sobre enfermedades genéticas humanas y los genes asociados a ellas.

En un principio, se seleccionaron diversas enfermedades como el Alzheimer, Síndrome de Down, Diabetes, Albinismo, entre otras. Luego de una discusión se decidió quedarse con el **Síndrome de Marfan**, [6] asociándola con Abraham Lincoln, quien padecía de esta patología. La decisión tomada permitió al equipo trabajar sobre una enfermedad poco vista por parte de los integrantes, a diferencia de las anteriormente mencionadas.

Asimismo, una vez decidida la patología, es necesario seleccionar uno de los genes asociados. En este caso se eligió trabajar sobre la fibrilina 1 (FBN1) [7]. Por último se buscó el gen en NCBI para finalmente obtener el transcritpo del mismo en formato GenBank [8]. Este constituye el *input* para todo el flujo de trabajo.

3.1. Síndrome de Marfan

El síndrome de Marfan es un trastorno genético que padece aproximadamente a 1 de cada 5000 personas. En esta enfermedad se ve afectado el tejido conectivo, incluyendo a la piel, los huesos, los vasos sanguíneos y otros órganos. El aspecto del cuerpo de un paciente que padece esta patología se puede visualizar en la Figura 1. Este síndrome es causado por mutaciones en el gen *FBN1*, que es esencial para la formación de fibras elásticas en el tejido conectivo. Las personas con este síndrome suelen presentar características distintivas como extremidades largas y delgadas, escoliosis, problemas cardiovasculares (como dilatación de la aorta) y anomalías oculares. La gravedad y los síntomas pueden variar ampliamente dependiendo del caso en particular.

El diagnóstico es clínico y genético, y las recomendaciones incluye la vigilancia cardiovascular y ortopédica, así como el tratamiento de las complicaciones para mejorar la calidad y esperanza de vida.

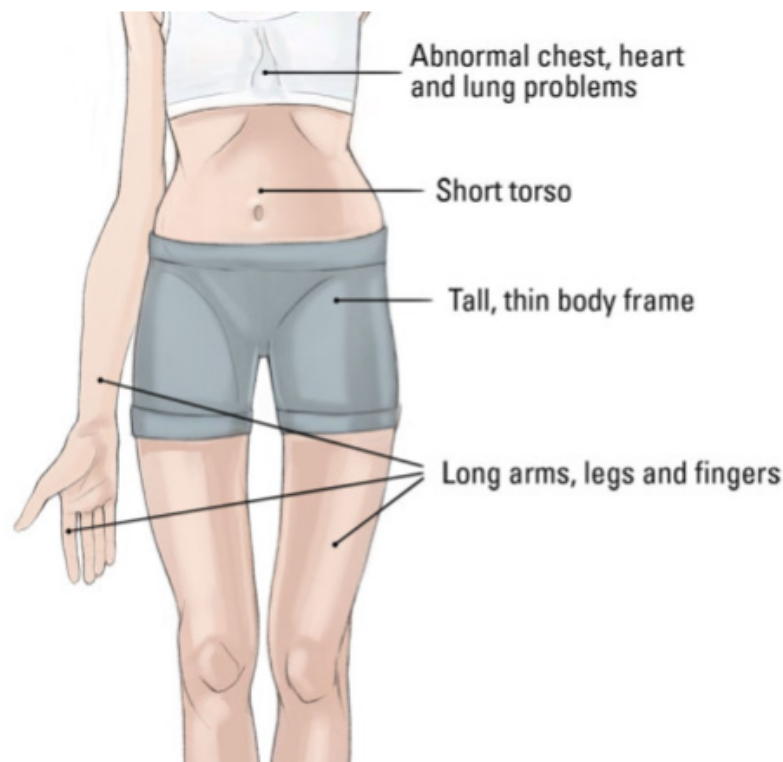


Figura 1: Contexturas y afecciones que padece una persona con síndrome de Marfan [9]

4. Ejercicios

4.1. Ejercicio 1: Procesamiento de secuencia

En este primer ejercicio se plantea hacer una traducción a partir de una secuencia nucleótido del mRNA de nuestro gen en formato GenBank. Para esto, es requisito tener en consideración los diferentes marcos de lecturas posibles (ORF, de *Open Reading Frame*). En un principio, se obtuvo la traducción de los 6 marcos principales.

Si bien una de las heurísticas desarrolladas en la cursada es que aquel ORF que produzca una cadena mas larga de aminoácidos es probablemente la que se exprese en la practica, se decidió tomar una aproximación distinta. Se aprovechó que el archivo GenBank del transcripto contiene tanto la secuencia del mRNA como la secuencia peptídica traducida. Por lo tanto se seleccionó, como el ORF correcto aquel que al traducirlo sea idéntica a la secuencia reportada dentro del mismo archivo GenBank. Entre las ventajas de utilizar esta aproximación, se incorporan datos experimentales que han dilucidado cuál ORF es el correcto.

Una vez hecho esto, se guarda este ORF en un archivo en formato FASTA con el ID de acceso en el encabezado para poder usarlo en futuras tareas.

4.2. Ejercicio 2: BLAST

Para el siguiente ejercicio es necesario confeccionar un script que lleve a cabo un BLAST, tanto de manera local como de forma remota. La entrada está constituida por el archivo .fasta producido en el primer ejercicio.

Tanto para el caso remoto como local se emplea la base de datos de **SwissProt** [10]. En el caso del BLAST local, se hace previamente una descarga de la base de datos para poder trabajar dentro del mismo ordenador y se ejecuta *makeblast* (dicho paso se encuentra implementado dentro del archivo **makeconfig.sh** del repositorio). En el caso remoto se emplean los servidores de NCBI para acceder al SwissProt y realizar un BLAST sin la necesidad de descargar la base de datos.

Mas allá de la modalidad, para ambas consultas se obtienen los resultados en formato XML y se extraen las 10 mejores secuencias encontradas (basadas en puntaje). Finalmente, estas secuencias se guardan en formato FASTA para su uso posterior.

Se incorporó la funcionalidad de realizar ambos tipos de consultas para todas las secuencias presentes en el archivo .fasta que se envíe como entrada, escribiendo informes diferentes para cada secuencia. No obstante, dado el flujo de trabajo planteado en la consigna y a que se ha analizado un único ORF de solo un mRNA, el resultado final son dos archivos .fasta correspondientes a las consultas local y remota para el ORF seleccionado.

4.2.1. Interpretación de los resultados de BLAST

En líneas generales, los valores estadísticos que arroja el BLAST son:

1. **Bit score:** puntaje, normalizado por el tamaño de la base de datos, que indica la calidad del alineamiento. Cuanto más alto es, se espera que haya significado biológico con más probabilidad.
2. **E-value:** arroja el valor esperado para el número de alineamientos con un cierto puntaje que se espera que ocurran por azar, considerando el tamaño de la base de datos. Cuanto más bajo, el alineamiento se considera más significativo.

El BLAST realizado arroja como resultado secuencias catalogadas como fibrillin-1. Por supuesto, la primera de todas corresponde con la proteína humana, que muestra un grado de similitud del 100 % (es la misma secuencia). A continuación, siguen fibrillin-1 correspondientes a diversos mamíferos, destacándose bovinos (*Bos taurus*) y ratones comunes (*Mus musculus*). Luego siguen proteínas similares (fibrillin-2 y fibrillin-3) y finalmente secuencias de otras proteínas, con una cobertura, puntaje y similitud menores.

Notar que muchas de las proteínas halladas con un puntaje alto son homólogas. Las primeras son ortólogas (gen de la fibrilina-1 en distintos mamíferos) y luego encontramos parálogas (gen de otras fibrilinas en el ser humano).

4.3. Ejercicio 3: Alineamiento Múltiple de Secuencias (MSA)

Con uno de los archivo FASTA de las 10 mejores secuencias obtenidas del ejercicio anterior (es indistinto cuál), mas la secuencia de consulta (aquella que contiene el ORF elegido traducido) se prosiguió a realizar un alineamiento múltiple.

Para esto mismo se utilizó el programa de **MUSCLE**, el cual se invoca dentro del código desarrollado en Python. Los resultados se escribieron tanto en un archivo FASTA como en otro CLUSTAL.

4.3.1. Resultados del MSA

Pueden visualizarse los resultados del MSA en la Figura 2. Dado que los hits de la búsqueda en BLAST se guardan indexados por puntaje, es de esperar que se muestre una mayor similitud en aquellas secuencias con un índice menor.

```

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

hit_8      -----
hit_9      -----
hit_10     -----
hit_7      LARGPLARLLLAWSALLCMAGGQGRWD---GALEAAGPGRVRR--GSPGILQGPNVCG
hit_5      -VRPAVAGSEGGF--MGPEYRDEGAVA-----ASRVRR--GQQEILRGPNVCG
hit_6      -VSATAGSEGGF--LAPEYREEGAIV-----ASRVRR--GQQDVLRGPNVCG
hit_4      MRRGGLEVALAFALLLESYTSHGADANLEAGSLKETRANRAKRRGGGGHDALKGPNVCG
hit_3      MRRGRLLLEVALGFTVLLASYTSHRAEANLEAGNGKETRASRAKRRGGGGHDALKGPNVCG
hit_1      MRRGRLLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRRGGGGHDALKGPNVCG
NM_000138.5 MRRGRLLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRRGGGGHDALKGPNVCG
hit_2      MRRGGLEVALGFTVLLASYTSHGADTNLEAGNVKETRANRAKRRGGGGHDALKGPNVCG
  
```

Figura 2: Primer Segmento del MSA obtenido en el Ejercicio 3

4.4. Ejercicio 4: EMBOSS

El uso del programa *patmatmotifs* permite la búsqueda de motivos en una secuencia query que se indica como entrada. La salida es, por defecto, un archivo de texto que detalla la ubicación de los distintos motivos predefinidos por una base de datos, junto con una descripción para cada uno de ellos en una segunda parte del reporte. Esta descripción consiste en un análisis literario resumido, pero completo, que detalla funciones asociadas y proteínas donde se suele hallar, entre otros aspectos. Para poder ejecutar *patmatmotifs*, es necesario descargar un archivo .dat que contiene los motivos a procesar (generalmente de Prosite [11]) y ejecutar *prosextract*.

Al realizar este ejercicio, se utilizó la base de datos de Prosite. Realizando los pasos descritos up supra para el archivo .fasta con el ORF elegido en el Ejercicio 1, se obtuvieron los motivos principales para la proteína de interés. Un resumen de la función prevista para los motivos más relevantes se describe en la Tabla 1. Si se desea una descripción completa, se debe ejecutar el código contenido en el repositorio.

4.5. Ejercicio 5: Primers

A partir del transcrito utilizado en el Ejercicio 1, se creó un script que permitiese diseñar 5 primers para detectar a la variante del mensajero dentro de muestras clínicas.

Para esto se tomaron en consideración los siguientes aspectos:

Tabla 1: Funciones asociadas a diferentes motivos para FBN1

Motivo	Función
Sitios de Amidación e Hidroxilación	Marcación de la proteína
RGD	Adhesión Celular
Dominio EGF	Estimulación de la división celular
Dominio de Unión a Calcio similar al Dominio EGF	Interacciones Proteína-Proteína, plegamiento tridimensional correcto
Dedo de Zinc	Interacción con nucleótidos
Sitio de Fosforilación con Tirosina	Marcación de la proteína

- **Cálculo del contenido de GC:** El script obtiene el porcentaje de guanina (G) y citosina (C) en una secuencia de ADN.
- **Cálculo de la temperatura de melting (Tm):** Se calcula la Tm de la secuencia a partir de las bases presentes.
- **Validación de extremos:** La función verifica que el primer no comience ni termine con las bases G o C.

Además de esto, mediante un archivo JSON se pueden modificar las siguientes características del primer:

- Longitud mínima
- Longitud máxima
- Porcentaje mínimo de GC
- Porcentaje máximo de GC
- Temperatura máxima de Tm

De esta manera, el diseño del primer puede ser sencillamente modificado por el usuario. Los parámetros descritos se toman en consideración para el cálculo de la Tm y del contenido de GC.

Una vez corrido el script se obtiene como salida un archivo de texto en donde se encuentra la secuencia de los diferentes primers, el porcentaje de GC y su Tm.

4.6. Ejercicio 6: Bases de Datos biológicas

Para esta última parte nos avocamos a dar información mas detallada del gen y su asociación con la enfermedad al indagar en bases de datos como NCBI. En este caso, como fue mencionado previamente, elegimos trabajar con el síndrome de Marfan, una enfermedad que está asociado a mutaciones en el gen FBN-1 (<https://www.ncbi.nlm.nih.gov/gene/2200>), ubicado en el cromosoma 15, que en los seres humanos codifica la fibrilina-1. Esta es una glicoproteína de matriz extracelular de gran tamaño que actúa como componente estructural de microfibrillas de unión al calcio.

4.6.1. Genes homólogos

Se buscaron genes homólogos a la proteína FBN-1 tanto en NCBI-Gene como en Ensembl.

0 items

604 genes for: jawed vertebrates (Gnathostomata)

0 selected

SEARCH THE TAXONOMY TREE

Enter taxonomic name

- ▶ jaws vertebrates
 - ▶ birds
 - ▶ turtles
 - ▶ alligators and others
 - ▶ lizards & snakes
 - ▶ mammals
 - ▶ amphibians
 - ▶ coelacanth
 - ▶ lungfishes
 - ▶ bony fishes
 - ▶ cartilaginous fishes

Species	Gene	Architecture	aa
<input type="checkbox"/> <i>Homo sapiens</i> human	FBN1 fibrillin 1		2,871
<input type="checkbox"/> <i>Mus musculus</i> house mouse	Fbn1 fibrillin 1		2,873
<input type="checkbox"/> <i>Rattus norvegicus</i> Norway rat	Fbn1 fibrillin 1		2,872

Figura 3: Resultados de la búsqueda de ortólogos en NCBI-Gene

En NCBI-Gene (previamente conocido como HoloGene) se hallaron 604 resultados dentro de un conjunto específico de especies (*Gnathostomata*) (Figura 3).

Por otro lado, en Ensembl se hallaron en total 245 ortólogos, 176 especies que tienen un ortólogo 1:1, 17 que tienen un ortólogo 1:muchos y 52 especies que no tienen ninguno (Figura 4). En cuanto a parálogos, se hallaron 2, ambos provenientes del ancestro común de los cordados.

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
Primates (26 species)	<input type="checkbox"/>	22	1	0	3
Humans and other primates					
Rodents and related species (30 species)	<input type="checkbox"/>	24	0	0	6
Rodents, lagomorphs and tree shrews					
Laurealtheria (42 species)	<input type="checkbox"/>	38	0	0	4
Carnivores, ungulates and insectivores					
Placental Mammals (103 species)	<input type="checkbox"/>	89	1	0	13
All placental mammals					
Sauropsida (59 species)	<input type="checkbox"/>	27	0	0	32
Birds and Reptiles					
Fish (66 species)	<input type="checkbox"/>	50	12	0	4
Ray-finned fishes					
All (245 species)	<input checked="" type="checkbox"/>	176	17	0	52
All species, including invertebrates					

Figura 4: Resultados de la búsqueda de ortólogos en Ensembl

NCBI y Ensembl tienen enfoques diferentes en la identificación de genes homólogos, lo que explica las variaciones en los resultados. Por ejemplo, emplean metodologías y definiciones de ortólogos diferentes. Ensembl utiliza árboles filogenéticos y una metodología más conservadora, mientras que NCBI combina similitud de secuencias y microsintenia, permitiendo relaciones transitivas y un conjunto más amplio de genes homologados. Esto explica por qué NCBI reporta más ortólogos, aunque Ensembl ofrece datos más curados y específicos para relaciones evolutivas directas [12].

4.6.2. Splicing alternativo

En NCBI se hallaron 4 resultados para splicing alternativo (ver Figura 5). La secuencia 1 es considerada como referencia (canónica). Las otras 3 son variantes de splicing alternativas - isoformas a, b y c.

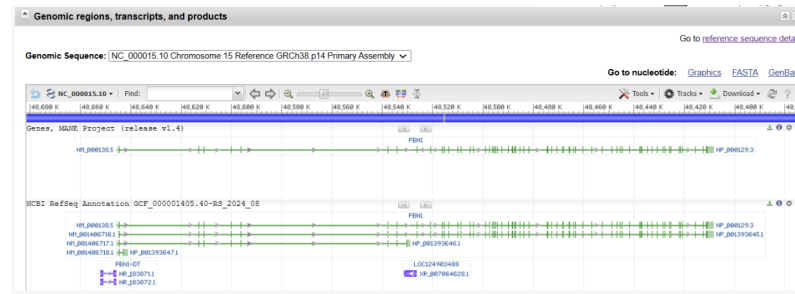


Figura 5: Resultados de la búsqueda de transcritos en Ensembl

La Figura 6 muestra los resultados de la búsqueda en Ensembl, donde encontramos que hay 14 transcritos, de los cuales 5 tienen una proteína relacionada y el resto no.

- *Protein coding*: Estos producen un producto proteico que puede ser accedido siguiendo los enlaces de UniProt.
- *ENST00000316623.10*: Es el transcrito canónico de fibrilina, mientras que el otro es una proteína mucho más pequeña, con solo 55 aminoácidos.
- *Nonsense mediated decay*: Estos tres transcritos probablemente contienen codones de parada prematuros, lo que lleva a la degradación de la proteína dentro de las células.

Gene: **FBN1** ENSG00000166147

Description: fibrillin 1 [Source:HGNC Symbol;Acc:HGNC:3603]

Gene Synonyms: FBN, MASS, MFS1, OCTD, SUG, WMS

Location: Chromosome 15: 48,400,313-48,845,721 reverse strand

GRCh38: CM000677.2

About this gene: This gene has 14 transcripts (splice variants), 234 orthologues, 2 paralogs and is associated with 26 phenotypes

Transcripts: [Hide transcript table](#)

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000166223.10	FBN1-201	11609	2871aa	Protein coding	CCDS32232.0	CS5555.0	NC_001335.5.0	MANE Select Ensembl Canonical GENCODE Primary GENCODE Basic APPRIS P1 TSL:1
ENST00000090055.1	FBN1-205	4109	55aa	Protein coding		HD7N80.0		GENCODE Primary GENCODE Basic TSL:1
ENST00000090133.8	FBN1-204	9910	2412aa	Nonsense mediated decay		HD7N80.0		TSL:1
ENST00000064302.2	FBN1-208	8995	2272aa	Nonsense mediated decay		AG088P1.22.0		TSL:5
ENST00000037463.6	FBN1-202	3765	302aa	Nonsense mediated decay		FBI405.0		TSL:5
ENST00000174128.1	FBN1-214	1219	No protein	Protein coding CDS not defined				TSL:5
ENST000000581426.1	FBN1-207	660	No protein	Protein coding CDS not defined				TSL:5
ENST00000058293.1	FBN1-203	482	No protein	Protein coding CDS not defined				TSL:3
ENST00000060445.1	FBN1-210	5268	No protein	Retained intron				
ENST00000060446.1	FBN1-213	5189	No protein	Retained intron				
ENST00000060276.1	FBN1-211	4389	No protein	Retained intron				
ENST00000060158.1	FBN1-209	2885	No protein	Retained intron				
ENST00000060358.1	FBN1-212	967	No protein	Retained intron				
ENST00000058020.1	FBN1-206	579	No protein	Retained intron				TSL:3

Figura 6: Resultados de la búsqueda de transcritos en Ensembl

La diferencia entre los dos en cuanto a la cantidad radica en que Ensembl típicamente incluye tanto transcritos curados manualmente como predicciones automáticas, mientras que NCBI RefSeq proporciona datos de transcritos curados manualmente y bien validados. Esto lleva a que haya más transcritos en el conjunto de datos de Ensembl. NCBI proporciona un conjunto confiable y conciso de variantes de empalme con relevancia biológica probada, mientras que Ensembl tiene una base de datos más amplia que puede ser útil si el enfoque está en explorar isoformas nuevas o menos caracterizadas.

4.6.3. Interacción con otras proteínas

Para el análisis de las interacciones con otras proteínas, se realizó la búsqueda en UniProt (ver Figura 7), donde se halló que la FBN1 interactúa con al menos 15 proteínas diferentes. Se puede

observar que la mayoría de los resultados están relacionados con proteínas que también participan de la estructura y función de la matriz extracelular, brindando elasticidad y organización al tejido conectivo, como es en el caso de colágeno, la elastina y las fibulinas (FBLN2 y FBLN5). Además, se observó una relación con proteínas como la LTBP2, LTBP1 y TNFSF11, que dependen de la presencia de iones de calcio que estabilizan la estructura proteica, regulando su capacidad de unirse a otras moléculas. Estas proteínas están involucradas en la regulación de TGF- β y la señalización celular.

Interactionⁱ

Subunitⁱ

Fibrillin-1



Interacts with COL16A1 (PubMed:15165854).
 Interacts with integrin alpha-V/beta-3 (PubMed:15062093).
 Interacts with ADAMTS10; this interaction promotes microfibril assembly (PubMed:21402694).
 Interacts with THSD4; this interaction promotes fibril formation (By similarity).
 Interacts (via N-terminal domain) with FBLN2 and FBLN5 (PubMed:15790312, PubMed:17255108).
 Interacts with ELN (PubMed:15790312).
 Forms a ternary complex with ELN and FBLN2 or FBLN5 and a significant interaction with ELN seen only in the presence of FBLN2 or FBLN5 (PubMed:17255108).
 Interacts (via N-terminal domain) with LTBP2 (via C-terminal domain) in a Ca(+2)-dependent manner (PubMed:17293099).
 Interacts (via N-terminal domain) with LTBP1 (via C-terminal domain) (PubMed:17293099).
 Interacts with integrins ITGA5:ITGB1, ITGAV:ITGB3 and ITGAV:ITGB6 (PubMed:12807887, PubMed:17158881).
 Interacts (via N-terminal domain) with BMP2, BMP4, BMP7, BMP10 and GDF5 (PubMed:18339631).
 Interacts (via N-terminal domain) with MFAP2 and MFAP5 (PubMed:15131124).
 Interacts with ADAMTSL5 (PubMed:23010571).
 Interacts with MFAP4 (PubMed:26601954).
 Interacts (via N-terminal domain) with TNFSF11 in a Ca(+2)-dependent manner (PubMed:24039232).
 Interacts (via N-terminal domain) with EFEMP2; this interaction inhibits EFEMP2 binding to LOX and ELN (PubMed:17255108, PubMed:19349279, PubMed:19570982).  

Figura 7: Resultados de la búsqueda de interacciones de la FBN1 en UniProt

4.6.4. Ontología del gen

La FBN1 (fibrilina-1) es una glicoproteína que desempeña un papel crucial en la formación y el mantenimiento de la matriz extracelular, especialmente en el ensamblaje de fibras elásticas, que proporcionan elasticidad a los tejidos conectivos. Esta se encuentra principalmente en la matriz extracelular, particularmente en las microfibrillas, y está asociada con estructuras como el tejido conectivo, la piel y los vasos sanguíneos [13].

Se realizó una búsqueda en AmiGO y en NCBI-Gene para identificar los procesos biológicos en los que interviene y cuál es su función molecular para explicar su influencia en el desencadenamiento del síndrome de Marfan. Los más relevantes fueron:

- Participa en la morfogénesis de estructuras anatómicas, lo cual está relacionado con las malformaciones óseas (escoliosis, deformaciones torácicas) y cardiovasculares (dilatación aórtica) características de pacientes que padecen del síndrome de Marfan.
- Interviene en la adhesión celular mediada por integrinas, afectando la integridad de los tejidos conectivos, ya que se relaciona con la interacción entre la matriz extracelular y las células.
- Está involucrada en la respuesta celular al estímulo del factor de crecimiento transformante beta (TGF- β). El desbalance de este factor es un mecanismo importante en la patogénesis

del síndrome de Marfan, contribuyendo a alteraciones estructurales y funcionales. [13] [14]

4.6.5. Vías metabólicas

Se realizó una búsqueda en KEGG para hallar las vías metabólicas en las que participa esta proteína. Se halló que la FBN1 participa en la vía de señalización del TGF- β , involucrada en la regulación del crecimiento y la diferenciación celular. Esta vía es fundamental para el desarrollo y mantenimiento de tejidos, y su desregulación está asociada a enfermedades como el síndrome de Marfan. Además, FBN1 juega un papel clave en la formación de microfibrillas dentro de la matriz extracelular, las cuales son cruciales para la integridad estructural de tejidos conectivos y musculares. Estas funciones muestran su importancia tanto en el mantenimiento de la matriz extracelular como en la regulación de procesos celulares [15].

4.6.6. Variantes genéticas

La mayoría de las mutaciones que causan el síndrome de Marfan cambian un solo aminoácido en la proteína fibrilina-1, y el resto codifican una proteína que tampoco puede funcionar correctamente. Las mutaciones que dan lugar a este síndrome reducen la cantidad de fibrilina-1 producida por la célula, alteran su estructura o estabilidad, o bien afectan el transporte de fibrilina-1 fuera de la célula. Esto reduce significativamente la cantidad de fibrilina-1 disponible para formar microfibrillas. Sin suficientes microfibrillas, se activan factores de crecimiento TGF- β en exceso, disminuyendo la elasticidad de muchos tejidos, y fomentando el sobrecrecimiento e inestabilidad de los mismos, un síntoma muy característico del síndrome de Marfan. [16]

rs140603		Current Build 156 Released September 21, 2022	
Organism	Homo sapiens	Clinical Significance	Reported in ClinVar
Position	chr15:48503845 (GRCh38.p14)	Gene : Consequence	FBN1 : Stop Gained
Alleles	G>A / G>C / G>T	Publications	6 citations LitVar²
Variation Type	SNV Single Nucleotide Variation	Genomic View	See rs on genome
Frequency	A=0.003238 (857/264690, TOPMED) A=0.000756 (190/251440, GnomAD_exome) A=0.003045 (427/140252, GnomAD) (+ 13 more)		

Figura 8: Resultados de la búsqueda de variaciones de la FBN1 en dbSNP

Se realizó la búsqueda de las variaciones genéticas en dbSNP, donde se hallaron 1480 resultados de variaciones patógenas. Una de ellas es la **rs140603** (ver fig. 8), la cual está asociada a una variación de un único nucleótido (single nucleotide variation, SNV), en donde se halla en vez de una G, una A una T o una C ($G > A / G > T / G > C$). Esta mutación genética introduce un codón de parada prematuro en la secuencia codificante del gen, lo cual da como resultado una proteína truncada que suele ser no funcional, interferir con procesos biológicos normales o incluso el ARNm resultante puede ser degradado. Su frecuencia en la población global es del 0,19 %, mientras que la población más afectada son los Africanos y los Afro-americanos, con un 1,04 % y un 1,06 % respectivamente.[17]

5. Conclusiones

Se llevo a cabo un trabajo completo respecto a una enfermedad de interés en donde se indagaron en el uso de varias herramientas a la vez que se llevó un rol realista de equipo trabajando en un

entorno laboral. Mas allá de haber cumplido con las tareas, también se logró adquirir nuevos conocimientos y habilidades, como lo son las metodologías ágiles, las cuales son de gran valor para la inserción en el ámbito profesional.

Con respecto al desarrollo conceptual, se ha logrado analizar de forma exhaustiva el gen elegido para completar los distintos incisos de la actividad. Se logró de forma exitosa orquestar un flujo de trabajo que incluye traducción de una secuencia nucleotídica en una secuencia peptídica, búsquedas en bases de datos, alineamiento múltiple de secuencias, análisis de motivos, y diseño de Primers. A su vez, se consultaron bases de datos biológicas que permitieron caracterizar la función de la proteína encontrada, sus interacciones, variantes e implicancia clínica de cambios con respecto a su expresión fisiológica.

Por último, se plantea como continuación la automatización del análisis para un conjunto más grande de secuencias, que apalanque LLMs actuales para el procesamiento de información conceptual. De esta forma, se permite explorar diversos caminos de forma eficiente y masiva, detectando aquellas regiones en donde la literatura actual sugiere el desarrollo de nuevas investigaciones.

6. Referencias

- [1] Inesdi Digital Business School. (2023) ¿Qué son las metodologías ágiles? Tipos y ejemplos. [Online]. Available: <https://www.inesdi.com/blog/que-son-las-metodologias-agiles-tipos-y-ejemplos/>
- [2] Robert Edgar, “MUSCLE.” [Online]. Available: <https://www.drive5.com/muscle/>
- [3] NCBI, “BLAST: Basic Local Alignment Search Tool.” [Online]. Available: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
- [4] EBI, “EMBOSS homepage.” [Online]. Available: <http://emboss.open-bio.org/>
- [5] O. M. I. in Man, “Omim - online mendelian inheritance in man,” 2024. [Online]. Available: <https://omim.org/>
- [6] Online Mendelian Inheritance in Man, “OMIM Entry Marfan Syndrome,” <https://www.omim.org/entry/154700?search=%22marfan%20syndrome%22&highlight=%22marfan%20%28syndromic%7Csyndrome%29%22>.
- [7] —, “OMIM Entry - FBN1,” <https://www.omim.org/entry/134797>.
- [8] National Center for Biotechnology Information, “Ncbi gene - gene id: 2200,” <https://www.ncbi.nlm.nih.gov/gene/2200>.
- [9] ZinkMD, “Amet sollicitudin - portfolio,” 2024. [Online]. Available: <https://zinkmd.com/portfolio/amet-sollicitudin/>
- [10] SIB - Swiss Institute of Bioinformatics, “UNIPrOTKB/Swiss-Prot - SIB Swiss Institute of Bioinformatics — Expasy.” [Online]. Available: <https://www.expasy.org/resources/uniprotkb-swiss-prot>
- [11] SIB - Swiss Institute of Bioinformatics, “PROSITE user manual.” [Online]. Available: <https://prosite.expasy.org/prosuser.html>
- [12] NCBI, “How are orthologs calculated?” 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/gene/2200/ortholog/?scope=7776>
- [13] NCBI, “Gene Ontology.” [Online]. Available: <https://www.ncbi.nlm.nih.gov/gene/2200#gene-ontology>
- [14] AmiGO, “Gene Ontology.” [Online]. Available: <https://amigo.geneontology.org/amigo/gene-product/UniProtKB:P35555>
- [15] KEGG, “Orthology FBN-1.” [Online]. Available: <https://www.genome.jp/entry/K06825>
- [16] M. . NIH, “Fbn1 gene,” 2024. [Online]. Available: <https://medlineplus.gov/genetics/gene/fbn1/#conditions>
- [17] dbSNP, “rs140603,” 2022. [Online]. Available: <https://www.ncbi.nlm.nih.gov/snp/rs140603>

A. Anexo

A continuación se incluyen links al repositorio donde está alojado el código desarrollado para el trabajo y la presentación desarrollada para presentar al curso los resultados obtenidos.

Repositorio en GitHub: <https://github.com/josulas/Bioinformatics-Practice.git>

Presentación en Canva: https://www.canva.com/design/DAGT9eSi6pU/jqyOTxY_9rzVRoV-poPojQ/edit