

Introducción a la Bioinformática

Trabajo práctico final



Instituto Tecnológico de Buenos Aires

Buenos Aires, Argentina

2015

Integrantes:

● Alejandro Bezdjian	abezdjia@itba.edu.ar	52108
● Matias Fraga	mfraga@itba.edu.ar	53214
● Tomás Mehdi	tmehdi@itba.edu.ar	51014

Ejercicio 5: Trabajo con Bases de Datos Biológicas.

- A. A partir del gen o proteína de interés para ustedes dar su link a NCBI-Gene como una entrada de Entrez. Expliquen brevemente lo que hace la proteína y por qué la eligieron.

Entrada de Entrez

<http://www.ncbi.nlm.nih.gov/gene/3930>.

La proteína codificada por este gen pertenece a la familia ERG4/ERG2 y se encuentra localizada en la envoltura del núcleo de la membrana interna y ancla la lámina y la heterocromatina a dicha membrana. Las mutaciones de este gen se asocian a una enfermedad autosómica recesiva conocida como displasia esquelética de Greenberg (HEM). También han habido casos en los que se la vinculó a la anomalía Pelger-Huet. Por otro lado, pueden producirse solapamientos cuando se identifican dos alternativas de transcripción de la misma proteína codificada.

Inicialmente investigamos sobre la enfermedad de Reynaud (o fenómeno de Reynaud) ya que un conocido posee dicha enfermedad. Pero en la primera de las 2 facetas que posee la enfermedad no había genes identificados por lo que nos vimos obligados a buscar otra alternativa. Sin ir muy lejos, muchas confusiones habían surgido entre este fenómeno y la enfermedad de Reynolds (no son el mismo nombre!!) y varias páginas remarcaban tener cuidado y no confundir ambas enfermedades, debido a que Reynaud es uno de los síntomas de la enfermedad de Reynolds, por lo que decidimos investigar sobre esta última.

- B. ¿Cuántos genes / proteínas homólogas se conocen en otros organismos? Utilicen la información que está en la base de datos de HomoloGene y en la bases de datos Ensembl. Describan los resultados en ambas bases de datos, y en qué se diferencian. Mencionen sobre qué tan común creen son estos genes o proteínas y a qué grupos taxonómicos pertenecen (sólo en las bacterias, en los vertebrados, etc.).

Links a HomoloGene

<http://www.ncbi.nlm.nih.gov/homologene/136116>

<http://www.ncbi.nlm.nih.gov/homologene/2455>

En la base de datos de HomoloGene se encuentran 18 proteínas homólogas conocidas. Dos de ellas son exclusivas de una rana llamada *Xenopus tropicalis*. Las otras 16 involucran chimpancés, mono Rhesus, perros, vacas, ovejas, ratones, pollos entre otros.

Link a Ensembl

http://www.ensembl.org/Homo_sapiens/Gene/Compara_Tree?db=core;g=ENSG0000143815;r=1:225401502-225428925

En la base de datos de **Ensembl** hallamos que el gen posee 186 homólogos como se observa en el árbol de genes que se encuentra en el *link*. Los grupos taxonómicos son vertebrados, reptiles (principalmente *actinopterigios*), cordados y marsupiales. En los vertebrados y los cordados se encuentra la mayor cantidad de genes homólogos.

La principal diferencia entre ambas bases de datos es que **Ensembl** presenta mucha más información y diversas posibilidades de visualizar dicha información (como por ejemplo el árbol de genes homólogos mencionado anteriormente) mientras que **HomoloGene** presenta información mucho más limitada respecto al gen.

- C. ¿Cuántos transcritos y cuántas formas alternativas de splicing son conocidos para este gen /proteína? ¿Cuáles de estos splicing alternativos se expresan? ¿Tienen funciones alternativas? Buscar evidencia de esto en las base de datos de NCBI y en los transcritos de Ensembl ¿Cómo el número de splicings alternativos diferente entre las dos bases de datos y cuál piensan que es más precisa y por qué?

Link a NCBI:

<http://www.ncbi.nlm.nih.gov/gene/3930#genomic-regions-transcripts-products>

Link a Ensembl:

http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000143815;r=1:225401502-225428925

NCBI presenta dos transcritos conocidos para esta proteína, con información muy precisa y detallada de las variantes así como también una descripción de las diferencias. Por otra parte **Ensembl** dispone de una tabla con los transcritos (presenta 8) y muestra información sobre la proteína transcrita pero también sobre otras proteínas. Finalmente consideramos que **NCBI** posee información más precisa ya que presenta la información solicitada y no completa con información de otras proteínas que no se vinculan con lo requerido.

- D. ¿Con cuántas otras proteínas interactúan el producto génico de su gen? ¿Existe un patrón o relación entre las interacciones? Mencione las interacciones interesantes o inusuales. Usted encontrará las interacciones de su gene/proteína tanto en la base de datos NCBI Gene como en la base de datos UniProt . Compare las dos tablas entre sí. ¿Hay proteínas que interactúan únicas para cada tabla?

Link a NCBI:

<http://www.ncbi.nlm.nih.gov/gene/3930#interactions>

Link a UniProt:

<http://www.uniprot.org/uniprot/Q14739#interaction>

NCBI muestra que existen 38 interacciones. especificando con que gen y el tipo de interacción. Cabe destacar que **NCBI** utiliza resultados de bases de datos externas. Por su parte **UniProt** muestra dos interacciones binarias así como también 45 interacciones (también producto de información en bases de datos externas). Llama la atención que **UniProt** posee 3 publicaciones relacionadas en el apartado de interacciones y separa las dos interacciones binarias pero con poca información vinculada a esas interacciones.

- E. Expliquen brevemente de qué componente celular forma parte su proteína (pista: se puede estudiar la información de Gene Ontology - GO), ¿A qué procesos biológicos pertenece (pista idem)? y ¿En qué función molecular trabaja esta proteína? Los términos ontológicos de genes los pueden encontrar tanto en NCBI Gene y en la base de datos UniProt como haciendo una búsqueda en AmiGO.

Forma parte de los siguientes componentes:

- núcleo
- envoltura nuclear
- membrana nuclear interna
- componente integral de la membrana nuclear interna
- membrana
- componente integral de la membrana
- membrana nuclear

Entonces es un componente de la membrana nuclear.

Participa de los siguiente procesos biológicos:

- proceso de biosíntesis del colesterol
- proceso metabólico de pequeñas moléculas
- proceso de oxidación-reducción

En el procesamiento del colesterol, en el proceso metabólico y en las reacciones de oxido-reducción.

Cumple las siguiente funciones:

- unión de ADN
- unión de proteínas
- unión laminar/de láminas
- uniones de poly(A) ARN
- *chromo shadow domain binding*
- *oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor*

- F. Discutan brevemente en qué estructura o vías metabólicas específicas (pathways) estaría participando su gen / proteína? (Reactome, KEGG son algunas bases de datos de pathways).

Link a *Ensembl*

http://www.ensembl.org/Homo_sapiens/Gene/Compare_Tree?db=core;g=ENSG0000143815;r=1:225401502-225428925

- G. Entrar en la base de datos de variantes genéticas dbSNP e intentar interpretarlo o encontrar info sobre alguna variante (reference SNP - rsXXXX) asociada con la patología investigada en su gen de interés. ¿Qué variante es? ¿Hay información sobre la frecuencia que tiene esta variante en la población? ¿Qué grupo étnico parece ser el más afectado?

Reference SNP (refSNP) Cluster Report: rs200180113				** With Pathogenic allele **			
RefSNP		Allele		HGVS Names			
Organism: human (Homo sapiens)		Variation Class: SNV: single nucleotide variation		NC_000001.10:g.225599113G>A			
Molecule Type: Genomic		RefSNP Alleles: A/G (FWD)		NC_000001.11:g.225411411G>A			
Created/Updated in build: 137/146		Allele Origin: A:germline G:germline		NG_008099.1:g.22407C>T			
Map to Genome Build: 107/Weight		Ancestral Allele: G		NM_002296.3:c.1114C>T			
Validation Status:		Variation Viewer:		NM_194442.2:c.1114C>T			
		Clinical Significance: With Pathogenic allele [ClinVar]		NP_002287.2:p.Arg372Cys			
		MAF/MinorAlleleCount: NA		XM_005273125.1:c.1114C>T			
		MAF Source:		XM_011544185.1:c.1114C>T			
				XM_011544186.1:c.1114C>T			

Population Diversity (Alleles in RefSNP orientation)							
Sample Ascertainment				Genotypes		Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	HWP	A	G
ss1686077088	ExAc_Aggregated_Populations		121412	AF		0.000	1.000