

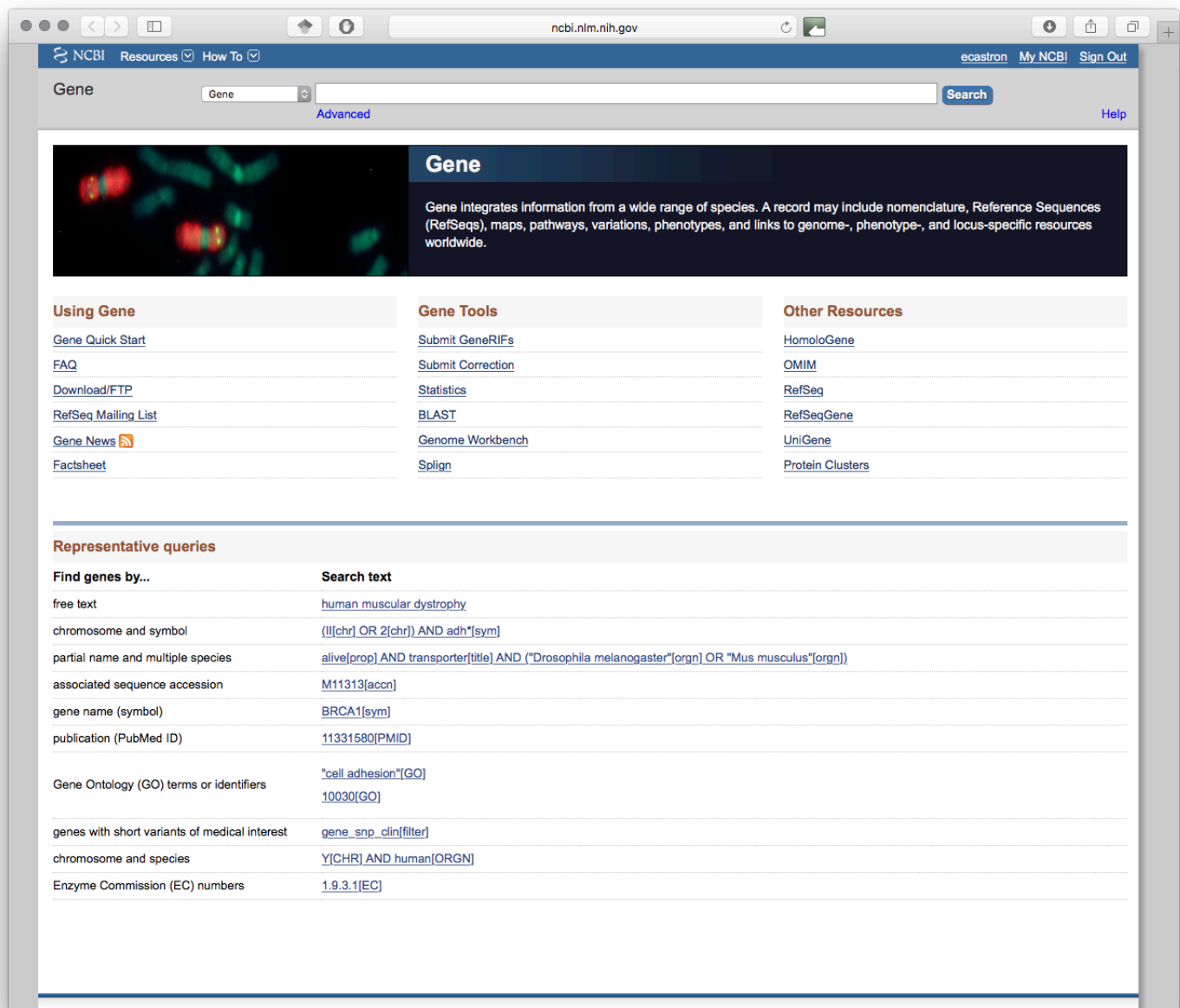
Laboratorio 01 - Bases de datos biológicas

En este laboratorio vamos a explorar bases de datos biológicas a través de un ejemplo guiado. La idea de este práctico es que uses y te interiorices con la mayor parte de las herramientas que las bases de datos te pueden ofrecer. Sigue las instrucciones y responde las preguntas. tus respuestas van a formar parte de la nota de laboratorio 1.

Parte 1: Enfermedades genéticas y genes

- Busca una enfermedad genética humana en alguna base de datos de literatura, i.e., [Pubmed](#), [Google Scholar](#), o en bases de datos de enfermedades genéticas, e.g., [OMIM](#)
- Describe en no más de 4 líneas la causa y lo que provoca la enfermedad que escogiste
- ¿Cuál(es) gene(s) han sido relacionados con esta enfermedad?

Ve a la NCBI Gene database (<http://www.ncbi.nlm.nih.gov/gene>) y busca el gen que está involucrado en la enfermedad seleccionada. Algunos ejemplos pueden ser CFTR, SGCG, IDDM3, HBB. También puedes tratar directamente con el nombre de una enfermedad o condición (e.g., duchenne muscular dystrophy).



La base de datos de genes de NCBI concentra información de distintas fuentes para producir una "ficha" con vínculos con otras bases de datos.

Responde:

¿Tiene nombres alternativos el gen?

¿En qué cromosoma está? ¿Cuántos exones tiene? ¿Cuántas isoformas de transcritos?

¿Qué tipo de proteína es codificada por este gen? ¿Cuál es su número EC?

¿Qué genes están inmediatamente río arriba/abajo?

De tus conocimientos de genética básica, probablemente ya sabes que un gen puede tener múltiples alelos, y

a su vez estos alelos pueden estar asociados a distintos fenotipos, e.g., enfermedades, severidad, etc. Lista cuántas variantes génicas tiene tu gen y a qué tipo de sustituciones corresponden (pista: revisa la sección de Variation en la página de tu gen).


Responde:

¿Cuál es la longitud de tu gen?
¿Cuántas variantes de tu gen hay descritas y en qué posiciones?
¿Las sustituciones corresponden a cambios sinónimos o no sinónimos?
¿Existen ortólogos de tu gen en otras especies? ¿Cuántos?
¿Y paralógos? ¿Hay pseudogenes? ¿Cuántos?

Parte 2: Rutas y procesos metabólicos

Los genes generalmente codifican proteínas, las cuales a su vez están involucradas en rutas y procesos metabólicos. Existen bases de datos especializadas que compilan información referente a metabolismo. Las más populares son [BioCyc](#), [REACTOME](#), y [Kegg](#). Otras bases de datos agrupan proteínas involucradas en procesos biológicos más gruesos como [Clusters of Orthologous Groups \(COGs\)](#) o [Gene Ontology \(GO\)](#). Visita los vínculos a estas páginas y familiarízate con sus funcionalidades (no más de 10 minutos).

Ahora sigamos trabajando con el gen que escogiste en la parte 1. Una vez en la página principal de Kegg, haz clic en el vínculo Kegg gene. Busca el gen que escogiste en Kegg. Tienes que ingresar el código del organismo (*Homo sapiens* es hsa) y el nombre del gen en minúscula (por ejemplo: gapdh) de forma que para buscar la gliceraldehído-3-fosfato deshidrogenasa deberías ingresar hsa:gapdh.



KEGG GENES Database

Molecular building blocks of life in the genomic space

KEGG2 PATHWAY BRITE MODULE KO GENOME GENES SSDB Organisms

Enter org:gene (Example) syn:ssr3451

Gene Catalogs

KEGG GENES is a collection of gene catalogs for all complete genomes (see [release history](#)) generated from publicly available resources, mostly NCBI RefSeq and GenBank. They are subject to SSDB computation and KO assignment (gene annotation) by KOALA tool. KEGG DGENES and MGENES are supplementary gene catalogs for draft genomes and metagenomes, which are given automatic KO assignment by [BlastKOALA](#) and [GhostKOALA](#), respectively, with GENES used as a reference data set. The collections of viral genomes and plasmids in RefSeq are also included in KEGG GENES with the standard annotation procedures.

Furthermore, a KEGG original protein sequence database is being developed as the GENES Addendum category. Protein sequences whose functions are experimentally characterized are collected from PubMed references and used to define new KOs that have not been covered by complete genomes (see [KO](#)).

Category	DBGET	Remark
Complete genomes	GENES	Complete genomes with KOALA and manual annotations
Viruses		Viral genomes with KOALA and manual annotations
Plasmids		Plasmids with KOALA and manual annotations
Addendum <i>New!</i>		PubMed-based collection of functionally characterized proteins
Draft genomes	DGENES	Draft genomes with automatic (BlastKOALA) annotation
Metagenomes	MGENES	Metagenomes with automatic (GhostKOALA) annotation

Search

☐ bfind mode ☐ bget mode

Search

☐ bfind mode ☐ boet mode

Después de presionar *entry* una nueva ventana se va a abrir donde encontrarás un montón de información. Revisa esta información para responder las siguientes preguntas:

Responde:

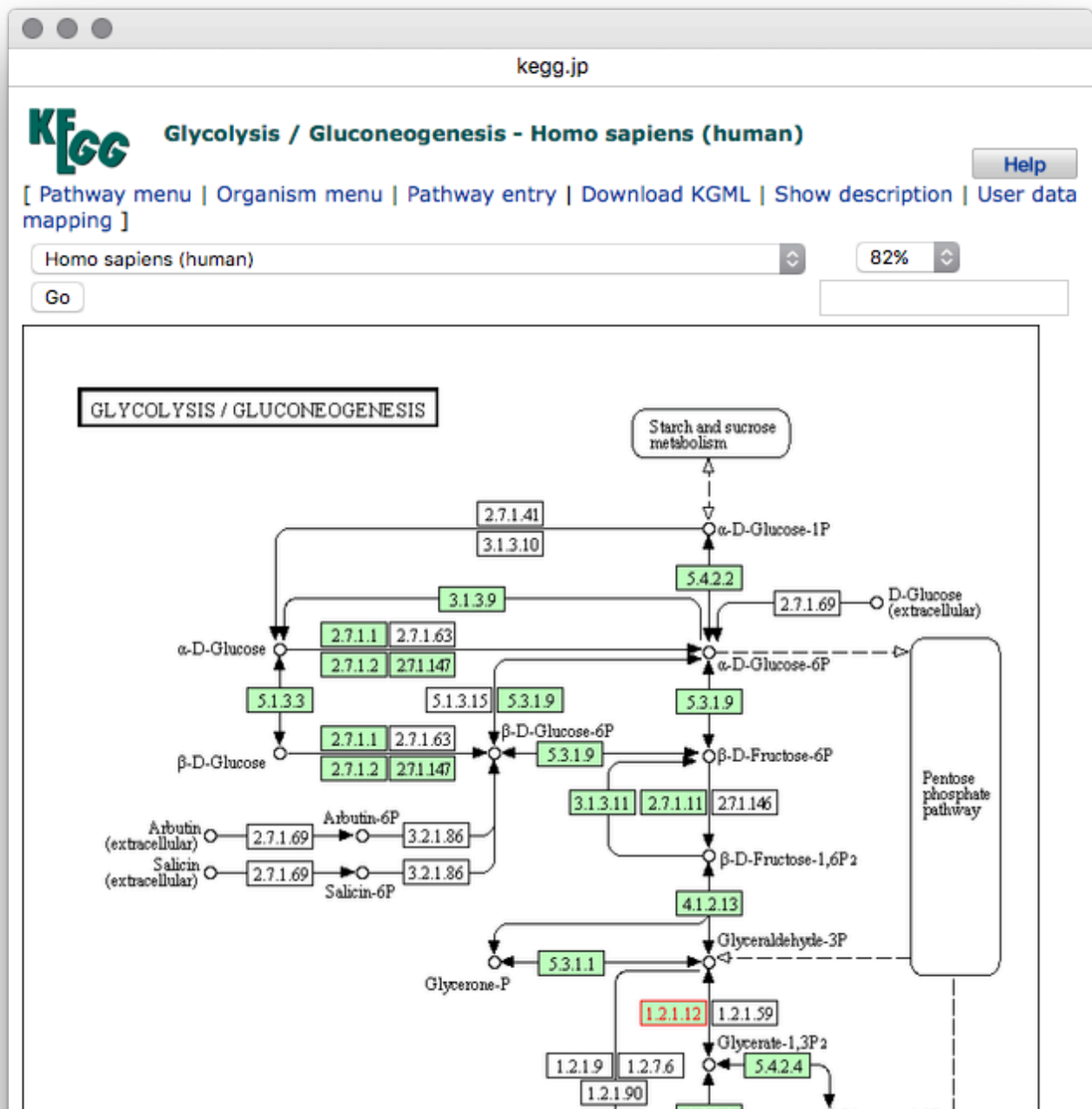
Anteriormente encontraste nombres alternativos de tu gen ¿Existen otros reportados por Kegg? ¿Cuáles?

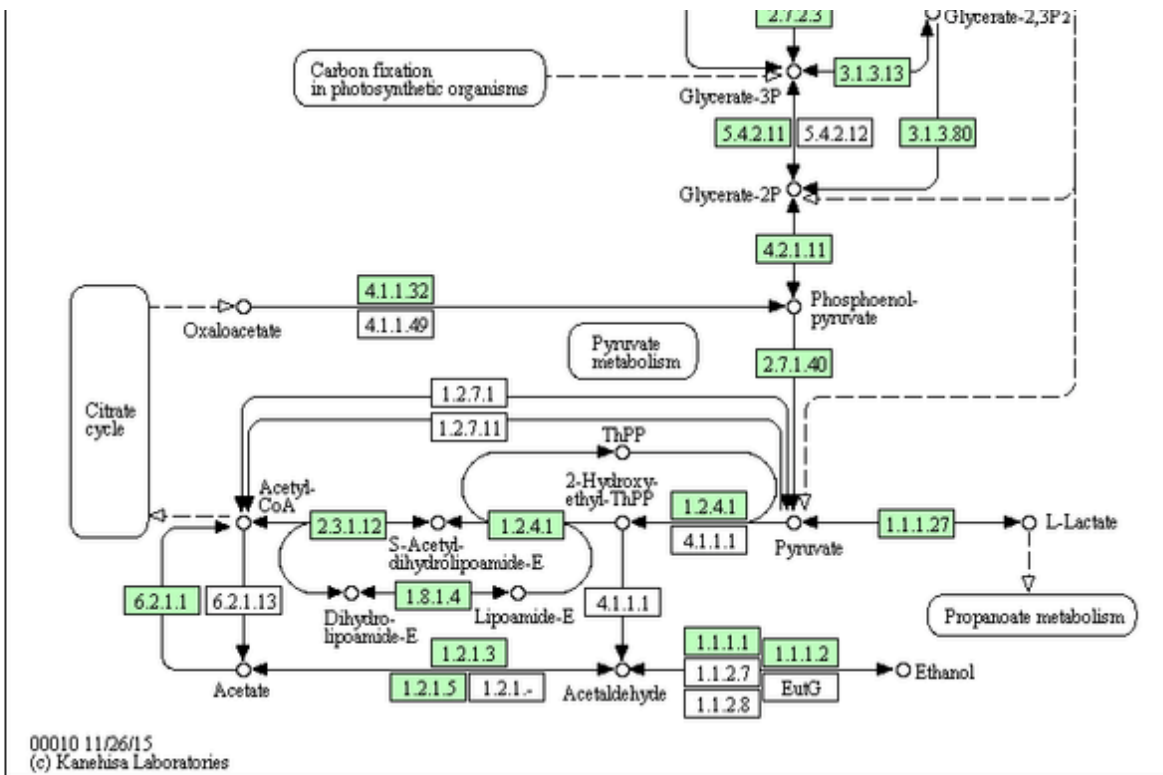
¿En qué rutas metabólicas participa la proteína codificada por tu gen?

¿Cuál es el número de identificación de tu gen (entry number)?

En general, cada unidad dentro de una base de datos tiene un número o código de identificación único. De esta forma, uno puede obtener exactamente lo que quiere dentro de un océano de otras cosas ¿En qué otras bases de datos está tu gen presente y cuáles son sus números de acceso?

La figura más clásica que se puede obtener es el diagrama con una ruta metabólica. Probablemente tu gen está involucrado en más de una. Escoge una y toma un pantallazo para que lo incluyas en tu informe.





Responde:

- ¿En qué otras rutas metabólicas está involucrado tu gen?
- ¿Qué significan los cuadros verdes en el diagrama?
- ¿Con qué rutas se cruza la ruta metabólica?

Ahora vamos a la página de [Gene Ontology \(GO\)](https://www.geneontology.org/). Una ontología es un tipo de anotación genómica que tiene que ver con atributos de regiones genómicas o de genes. En este sentido, el gen del citocromo c, por ejemplo, no es parte de GO pero su actividad biológica, i.e., su atributo, sí está (oxidoreductasa). Cualquier persona que quiera hacer un experimento de transcriptómica o genómica, secuenciar un genoma de un organismo, etc. debería tener un entendimiento profundo sobre GO. Un buen comienzo es [este artículo](#) en PLOS Computational Biology.

Ahora, desde el website de GO, lee la documentación para poder responder las siguientes preguntas:

Responde:

¿Cuántos "dominios" forman la anotación GO?

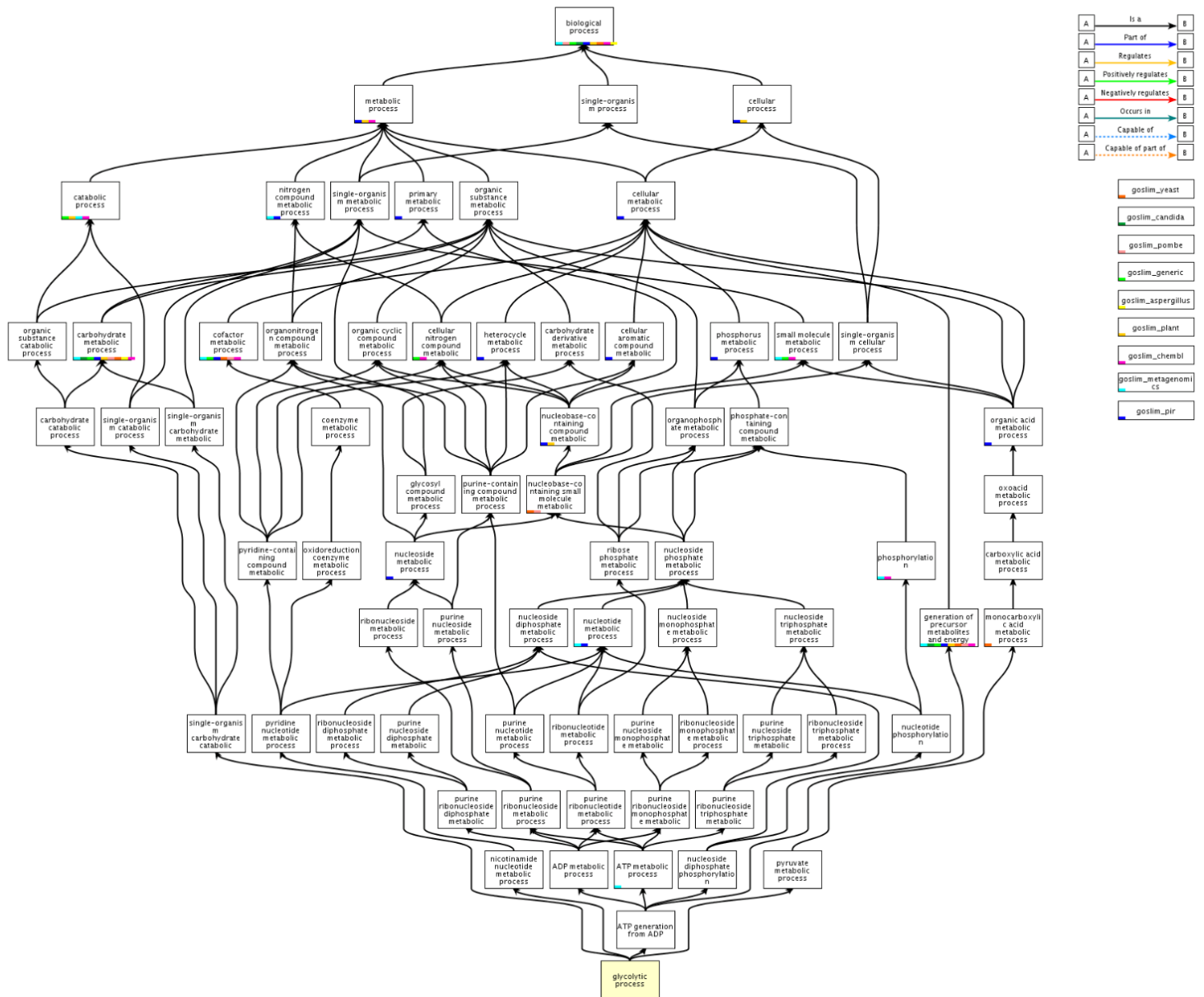
Ve a "Tools" --> "AmiGO 2" y escribe en la casilla de búsqueda GO:0006096

¿A qué corresponde este término y qué información te entrega la página?

Haz clic en "Graph Views" y examina el gráfico. Anota 10 sub-categorías GO a la cual GO:0006096 pertenece

Annotations Graph Views Inferred Tree View Ancestors and Children Mappings

View this term in QuickGO.



Parte 3: Descargando secuencias, convirtiendo formatos

De alguna u otra forma, probablemente ya has tenido algún tipo de interacción con NCBI descargando secuencias nucleotídicas o aminoácidas. En esta sección nos vamos a enfocar en descargar secuencias y

sus números de acceso.

Ve a la página de [NCBI](https://www.ncbi.nlm.nih.gov/) y selecciona la base de datos de nucleotidos. En la casilla de búsqueda escribe GAPDH y presiona Enter.

Responde:

¿Cuántos items fueron encontrados? ¿cuántos en animales?

Probablemente tus resultados fueron una mezcla de fragmentos de genes, regiones codificantes parciales, genes completos, etc. Filtra tus datos por mRNA, animales, Ref Seq.

Haz clic en la entrada para la secuencia de GAPDH de gallina.

¿Cuál es la longitud del gen?

¿Cuál es la referencia bibliográfica más reciente?

¿Cuál es el número de acceso?

Descarga la secuencia en formato fasta y agrégala a tu informe



```
>lcl|NM_204305.1_cds_NP_989636.1_1 [gene=GAPDH] [protein=glyceraldehyde-3-phosphate dehydrogenase] [protein_id=NP_989636.1] [location=58..1059]
ATGGTGAAAGTCGGAGTCAACGGATTTGGCCGTATTGGCCGCCTGGTCACCAGGGCTGCCGTCTCTG
GCAAAGTCCAAGTGGTGGCCATCAATGATCCCTTCATCGATCTGAACTACATGGTTTACATGTTCAAATA
TGATTCTACACACGGACACTTCAAGGGCACTGTCAAGGCTGAGAACGGGAACTTGTGATCAATGGGCAC
GCCATCACTATCTTCCAGGAGCGTGACCCCAGCAACATCAAATGGGCAGATGCAGGTGCTGAGTATGTTG
TGGAGTCCACTGGTGTCTTACCACCATGGAGAAGGCTGGGGCTCATCTGAAGGGTGGTGCTAAGCGTGT
TATCATCTCAGCTCCCTCAGCTGATGCCCCATGTTTGTGATGGGTGTCAACCATGAGAAATATGACAAG
TCCCTGAAAATTGTCAGCAATGCATCGTGACCAACCAACTGCCTGGCACCCCTTGGCCAAGGTCATCCATG
ACAACTTTGGCATTGTGGAGGGTCTTATGACCACTGTCCATGCCATCACAGCCACACAGAAGACGGTGGA
TGGCCCTCTGGGAAGCTGTGGAGAGATGGCAGAGGTGCTGCCAGAACATCATCCAGCGTCCACTGGG
GCTGCTAAGGCTGTGGGGAAAGTCATCCCTGAGCTGAATGGGAAGCTTACTGGAATGGCTTCCGTGTGC
CAACCCCCAATGTCTCTGTTGTTGACCTGACCTGCCGTCTGGAGAAACCAGCCAAGTATGATGATATCAA
GAGGGTAGTGAAGGCTGCTGCTGATGGGCCCCTGAAGGGCATCCTAGGATACACAGAGGACCAGGTTGTC
TCCTGTGACTTCAATGGTGACAGCCATTCTCCACCTTTGATGCGGGTGCTGGCATTGCACTGAATGACC
ATTCGTCAAGCTTGTTTCTGGTATGACAATGAGTTTGGATACAGCAACCGTGTTGTGGACTTGATGGT
CCACATGGCATCCAAGGAGTGA
```


Como nota adicional, la interfaz gráfica de NCBI no es de mucha ayuda cuando tienes que descargar muchas secuencias o genomas completos. En este caso se recomienda usar el portal FTP de NCBI:

<ftp.ncbi.nlm.nih.gov>

Para finalizar esta parte, vamos a convertir la secuencia que bajaste a otro formato. Existen muchos formatos de secuencias porque responden a distintos atributos que sus autores consideraron importantes de registrar y también por razones históricas. Uno de los formatos más usados es el FASTA, sin embargo en filogenética uno súper popular y requerido por muchos programas es el formato NEXUS.

Ve a la página de [Seqret](#) del EBI. Luego copia y pega la secuencia de GAPDH en fasta en la casilla.

Adjunta la secuencia en formato nexus a tu informe.

The screenshot shows a web browser window with the address bar displaying `ebi.ac.uk`. The page is the EMBL-EBI Seqret interface. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below this is a large teal header with the text "EMBOSS Seqret". Under the header, there are tabs for Input form, Web services, and Help & Documentation, along with Share and Feedback links. The main content area shows the results for a job with ID `emboss_seqret-l20160309-222519-0225-88553367-oy`. There are two tabs: "Tool Output" (selected) and "Submission Details". Under the "Tool Output" tab, there is a "Download" button and a text area containing the following NEXUS format sequence:

```
#NEXUS
[TITLE: Written by EMBOSS 09/03/16]

begin data;
dimensions ntax=1 nchar=1002;
format interleave datatype=DNA missing=N gap=-;

matrix
NM_204305.1_cds_NP_9 ATGGTGAAAGTCGGAGTCAACGGATTGGCCGTATTGGCCGCTGGTCAC
NM_204305.1_cds_NP_9 CAGGGCTGCCGTCTCTCTGGCAAAGTCCAAGTGGTGGCCATCAATGATC
NM_204305.1_cds_NP_9 CCTTCATCGATCTGAACATACATGGTTTACATGTTCAAATATGATTCTACA
NM_204305.1_cds_NP_9 CACGGACACTTCAAGGGCACTGTCAAGGCTGAGAACGGGAAACTTGTGAT
NM_204305.1_cds_NP_9 CAATGGGCACGCCATCACTATCTTCCAGGAGCGTGACCCAGCAACATCA
NM_204305.1_cds_NP_9 AATGGGCAGATGCAGGTGCTGAGTATGTTGTGGAGTCCACTGGTGTCTTC
NM_204305.1_cds_NP_9 ACCACCATGGAGAAGGCTGGGGCTCATCTGAAGGTTGGTCTAAGCGTGT
NM_204305.1_cds_NP_9 TATCATCTCAGCTCCCTCAGCTGATGCCCCATGTTTGTGATGGGTGTCA
NM_204305.1_cds_NP_9 ACCATGAGAAATATGACAAGTCCCTGAAAATTGTCAGCAATGCATCGTGC
NM_204305.1_cds_NP_9 ACCACCAACTGCCTGGCACCCTTGGCCAAGTTCATCCATGACAACCTTGG
NM_204305.1_cds_NP_9 CATTGTGGAGGGTCTTATGACCACTGTCCATGCCATCACAGCCACACAGA
NM_204305.1_cds_NP_9 AGACGGTGGATGGCCCTCTGGGAAGCTGTGGAGAGATGGCAGAGGTGCT
NM_204305.1_cds_NP_9 GCCCAGAACATCATCCAGCGTCCACTGGGGCTGCTAAGGCTGTGGGGAA
```

```

NM_204305.1_cds_NP_9 AGTCATCCCTGAGCTGAATGGGAAGCTTACTGGAATGGCTTTCCGTGTGC
NM_204305.1_cds_NP_9 CAACCCCAATGTCTCTGTTGTTGACCTGACCTGCCGTCTGGAGAAACCA
NM_204305.1_cds_NP_9 GCCAAGTATGATGATATCAAGAGGGTAGTGAAGGCTGCTGCTGATGGGCC
NM_204305.1_cds_NP_9 CCTGAAGGGCATCCTAGGATACACAGAGGACCAGGTTGTCTCCTGTGACT
NM_204305.1_cds_NP_9 TCAATGGTGACAGCCATTCTCCACCTTTGATGCGGGTGCTGGCATTGCA
NM_204305.1_cds_NP_9 CTGAATGACCATTTCGTCAAGCTTGTTCCTGGTATGACAATGAGTTTGG
NM_204305.1_cds_NP_9 ATACAGCAACCGTGTGTGGACTTGATGGTCCACATGGCATCCAAGGAGT
NM_204305.1_cds_NP_9 GA
;

end;
begin assumptions;
options deftype=unord;
end;

```

Parte 4: Buscando artículos científicos en línea

Buscar literatura relativa a un tema científico no tiene que ser tedioso ni complicado. Al contrario, conociendo herramientas dedicadas para esto puede hacerte la vida muy simple.

En general el primer impulso es buscar en los clásicos buscadores directamente. Sin embargo el uso de alertas, tablas de contenido electrónicas, y por sobre todo modificadores de búsquedas en línea hacen que estar actualizado en tu tema de investigación sea algo instantáneo.

- Crea una cuenta gratuita en NCBI y Google Scholar
- Escoge un área de investigación, e.g., bacterial genomics, human genetics, etc.
- Ahora crea una alerta de búsqueda en NCBI PubMed

En tu informe de laboratorio incluye un pantallazo de tu alerta. Si es que recibes una alerta en tu correo electrónico, también puedes adjuntarla en tu informe.

Las páginas de búsqueda clásicas son:

- [Pubmed](#), [Highwire](#), [Google Scholar](#), [Scopus](#) para artículos científicos
- [USPTO](#), [ESPACENET](#), [INAPI](#) para patentes. Google Scholar también permite filtrar resultados por patentes

Ahora vamos a la página de la revista [Nature Genetics](#). El objetivo es configurar una tabla de contenidos electrónica, i.e., que cada vez que la revista publique un número nuevo te llegue la tabla de contenidos de ese número a tu correo electrónico. Puedes encontrar el vínculo en la esquina superior derecha, bajo el *current issue*, E-alert



En tu informe agrega un pantallazo del correo electrónico de confirmación a la suscripción.

Para finalizar esta parte vamos a practicar el uso de *operadores de búsqueda en Google Scholar*. Abre una ventana y entra al Google Scholar. Por ejemplo busca Human Microbiome (puedes buscar cualquier otro término).

Ahora usa comillas para realizar tu búsqueda "Human Microbiome"

¿Son los resultados idénticos o no?

El uso de comillas restringe los resultados a resultados donde la frase que buscas aparece de manera exacta.

También puedes usar * para representar una palabra que falta, e.g., "Human Microbiome *"

¿En qué cambiaron los resultados de la búsqueda?

También puedes condicionar tus búsquedas a rangos de números como precios, años, etc. Prueba con 14 inch...17 inch laptops en google.com

¿Qué encuentras en los resultados? Prueba sin el rango también

Para buscar artículos científicos también es útil restringir los resultados de búsqueda a tipos de archivo. Prueba con "human microbiome" filetype:pdf

Describe tus resultados

Otro truco útil es usar signos - y +. Por ejemplo trata buscar "PCR amplification" +temperature, y luego "PCR amplification" -temperature

¿En qué cambian los resultados de la búsqueda?

Finalmente, prueba los operadores Boolean que representan opciones de inclusión. Por ejemplo, trata con Soil OR Human pathogens y luego trata con Soil AND Human pathogens

De nuevo, ¿en qué cambian los resultados de la búsqueda?

Puedes seguir practicando tus habilidades para buscar artículos científicos. Los ejemplos de esta sección fueron tomados desde [acá](#).

No olvides agregar tus respuestas al informe de laboratorio.

Trabajo de laboratorio para la próxima semana

El trabajo de laboratorio para la próxima semana consta de dos partes. La primera parte ya la tienes parcialmente lista. Simplemente tienes que responder las preguntas que aparecen a través de esta guía y enviar un informe a tu ayudante con copia a eduardo.castro@unab.cl. Recuerda que si te la juegas y entregas el informe como documento Markdown tendrás 5 décimas adicionales en tu nota. Información sobre cómo escribir documentos Markdown está en la web por ejemplo [aquí](#) o [aquí en inglés](#).

La segunda parte del trabajo de laboratorio tiene que ver con leer un capítulo de libro y un artículo científico. Estos serán evaluados la próxima clase como parte del primer control de laboratorio.

- El capítulo 2 del libro *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* que se llama *Evolutionary Concept in Genetics and Genomics*. Puedes acceder al capítulo gratis a través de NCBI --> Bookshelf [aquí](#).
- El siguiente artículo científico --> [Gabaldón, T., & Koonin, E. V. \(2013\). *Functional and evolutionary implications of gene orthology*. Nature Reviews Genetics, 14\(5\), 360-366.](#)