Colegio de
San Francisco de Paula

Biotecnología y bioinformática

Opción B

B5: Bioinformática

International Baccalaureate

Tema 9 de Biología NS

Diploma BI

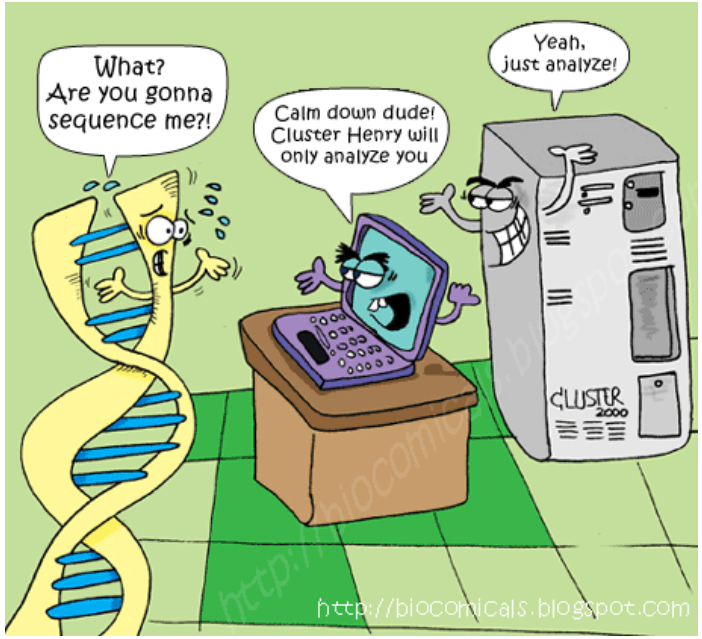


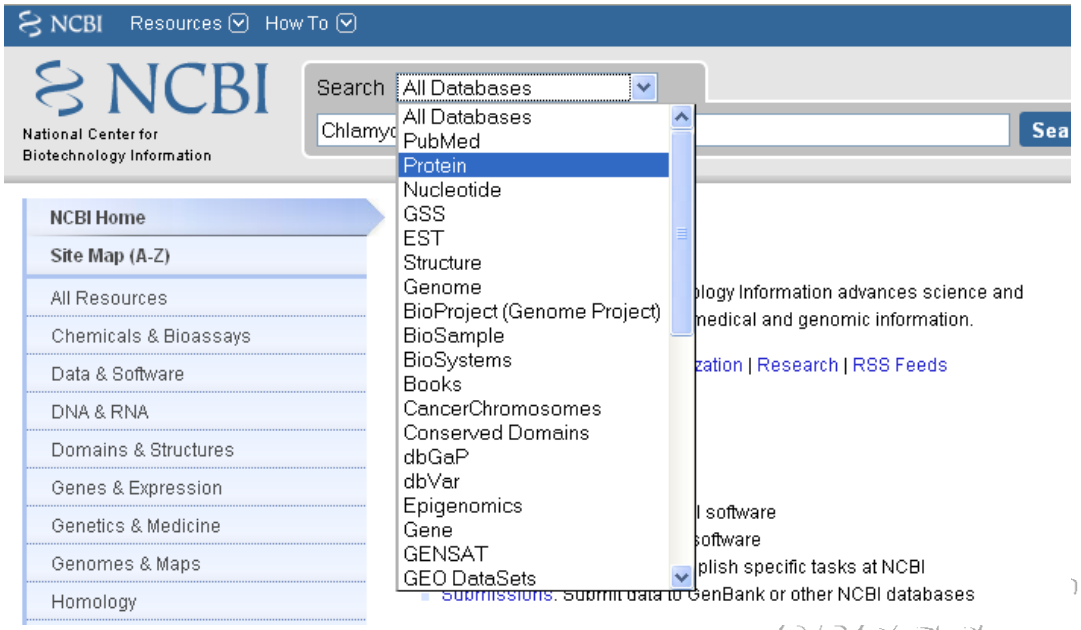
IMAGEN: http://4.bp.blogspot.com

Idea Fundamental: La bioinformática consiste en el uso de computadores para analizar secuencias de datos en investigaciones biológicas.

Bases de datos

- Una base de datos es una colección estructurada de información almacenada en un ordenador, que puede incluir datos en un rango de formatos como imágenes, artículos, información cualitativa o cuantitativa.
- **Las bases de datos facilitan a los científicos un sencillo acceso a la información.**
- Alguna de las bases de datos usadas en bioinformática incluyen:

  - Secuencias de nucleótidos (EMBL).
  - Secuencias de proteínas (SwissProt).
  - Estructura 3D de proteínas (PDB).
  - Expresión génica de chips de ADN.
  - Rutas metabólicas (KEGG).



Bases de datos

- En la actualidad es posible la comprobación de una hipótesis a partir de las bases de datos, y no solo a partir de datos obtenidos directamente por uno mismo.
- Un investigador puede emplear una base de datos para:

  - Añadir los propios datos que ha obtenido para que tengan acceso a ellos otros investigadores.
  - Extraer un subconjunto de datos.
  - Hacer una búsqueda de una secuencia en concreto.

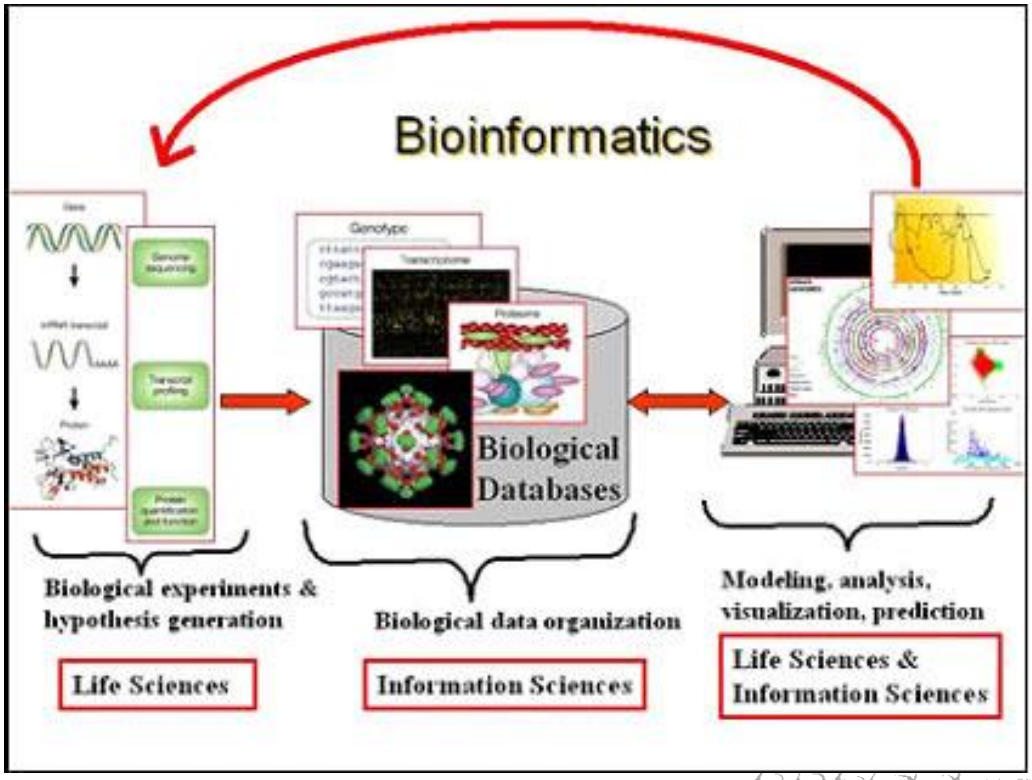


IMAGEN: stellamariscollege.org

Crecimiento de las bases de datos

- Los avances tecnológicos han provocado que el ritmo de creación y publicación de datos está incrementando, **de hecho, el cuerpo de los datos almacenados en las bases de datos aumenta exponencialmente.**
- Los avances en la tecnología de secuenciación genómica, chips de ADN, programas de modelización 3D y en la capacidad de computación, han resultado en un número de proyectos colaborativos de investigación a gran escala que han culminado en este crecimiento de los datos almacenados en las bases de datos.

IMAGEN: ulacir.ac.cr

NATURALEZA CIENCIAS: Cooperación y colaboración entre grupos científicos

- La mayoría de las bases de datos bioinformáticas son públicas y de libre acceso a todos los investigadores. Frecuentemente, cuando se añaden datos a una base de datos, se sincronizan inmediatamente con otras bases de datos.
- **Las bases de datos en internet facilitan a los científicos el acceso libre a la información.**
- El acceso libre y la sincronización facilitan la colaboración y el espíritu de cooperación.
- Sin embargo, hay quien piensa que la comercialización de las bases de datos bioinformáticas es una amenaza para este espíritu.

IMAGEN: ngeeks.com

NATURALEZA CIENCIAS: Cooperación y colaboración entre grupos científicos

- Algunos trabajadores de compañías privadas no facilitan el acceso libre a sus datos, debido a la necesidad de obtener beneficios.
- Así, algunas bases de datos que en el pasado eran de acceso libre, han sido adquiridas por compañías que han comenzado a cambiar el acceso a la información.
- Un ejemplo de ello son las bases de datos de la levadura *Saccharomyce cereviciae* y del gusano *Caenorhabditis elegans*, dos de los organismos modelo eucariotas más usados. Esto generó gran controversia, dado que contenían datos de estudios publicados y comunicaciones personales.

IMAGEN: senescense.info

NATURALEZA CIENCIAS: Cooperación y colaboración entre grupos científicos

- Otro ejemplo de esto puede encontrase en la revista científica Science, que aún habiendo publicado la secuenciación del genoma humano por la empresa Celera en 2001, y del arroz por Syngenta en 2002, permitió que ambas almacenaran dichas secuencias en sus servidores privados, en lugar de hacerlos públicos en GenBank, como se hacía hasta el momento.
- Esto además, está en contra de otro de los principios de la investigación científica, que es que los datos publicados deben ser de libre acceso para que la comunidad científica pueda verificarlo.

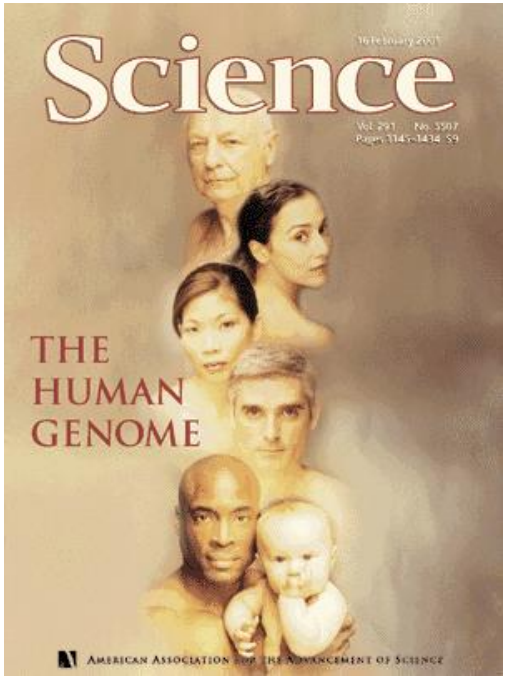
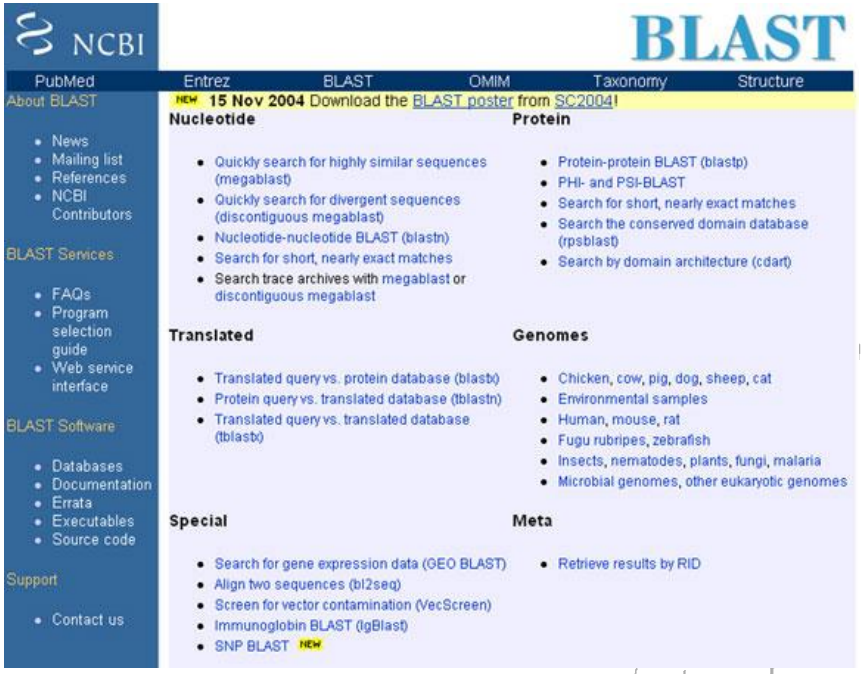


IMAGEN: sciencemag.com

Búsqueda de secuencias similares: BLAST

- Una vez que un investigador identifica por primera vez una secuencia de interés, ya sea tras secuenciar un gen, identificar un marco abierto de lectura o encontrar unos altos niveles de ARNm, el siguiente paso es realizar una **búsqueda con software BLAST para identificar secuencias similares en organismos diferentes.**

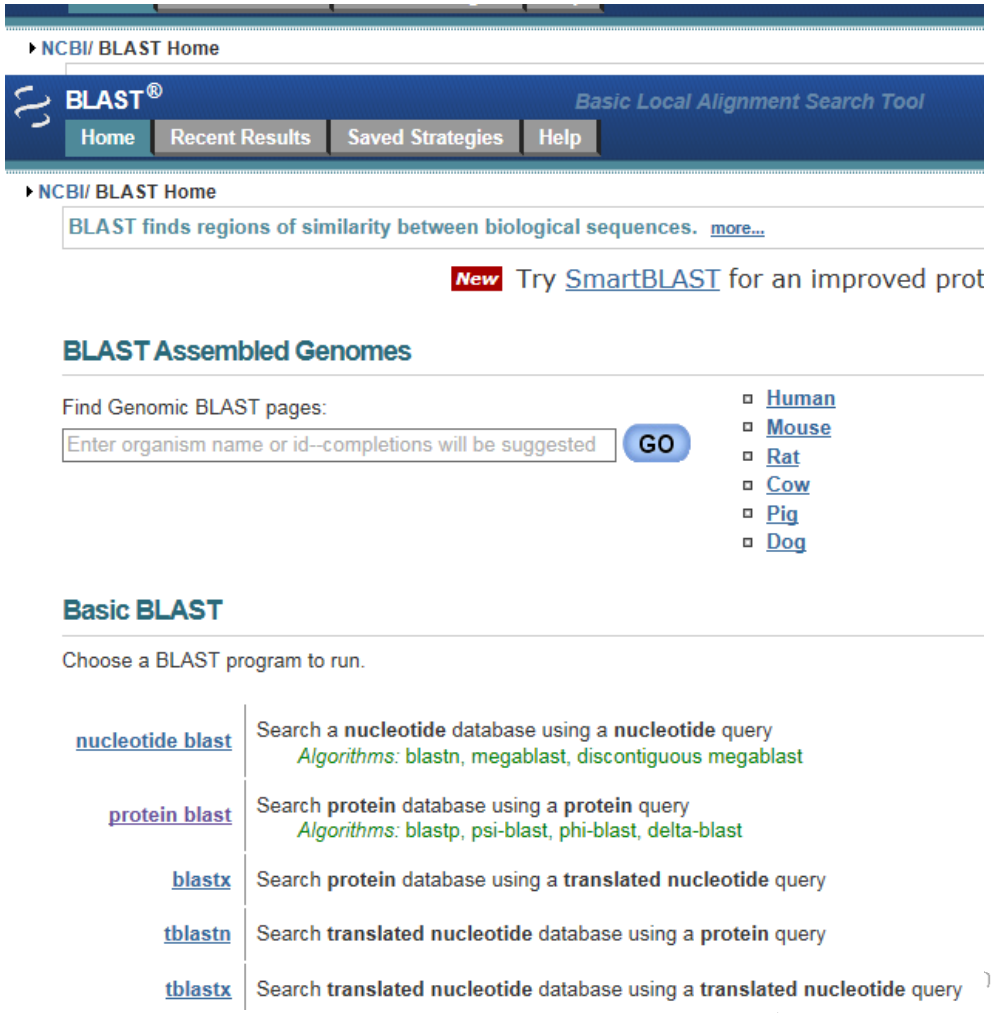
- El acrónimo BLAST viene del inglés "Basic Local Alignment Search Tool". Este programa informático lleva a cabo un algoritmo matemático que permite encontrar regiones de similitud entre secuencias (nucleótidos o aminoácidos) alojadas en las bases de datos.



Búsqueda de secuencias similares: BLAST

- Una vez que un investigador ha identificado un marco abierto de lectura en una secuencia de nucleótidos, puede realizar una **búsqueda con el software BLASTn, que permite una alineación de secuencias de nucleótidos** en las bases de datos con objeto de determinar si existe dicho ORF en otras especies.
- Por otro lado, **el software BLASTp permite una alineación de proteínas.**
- El software tBLASTn permite localizar en el genoma el gen de una determinada proteína.

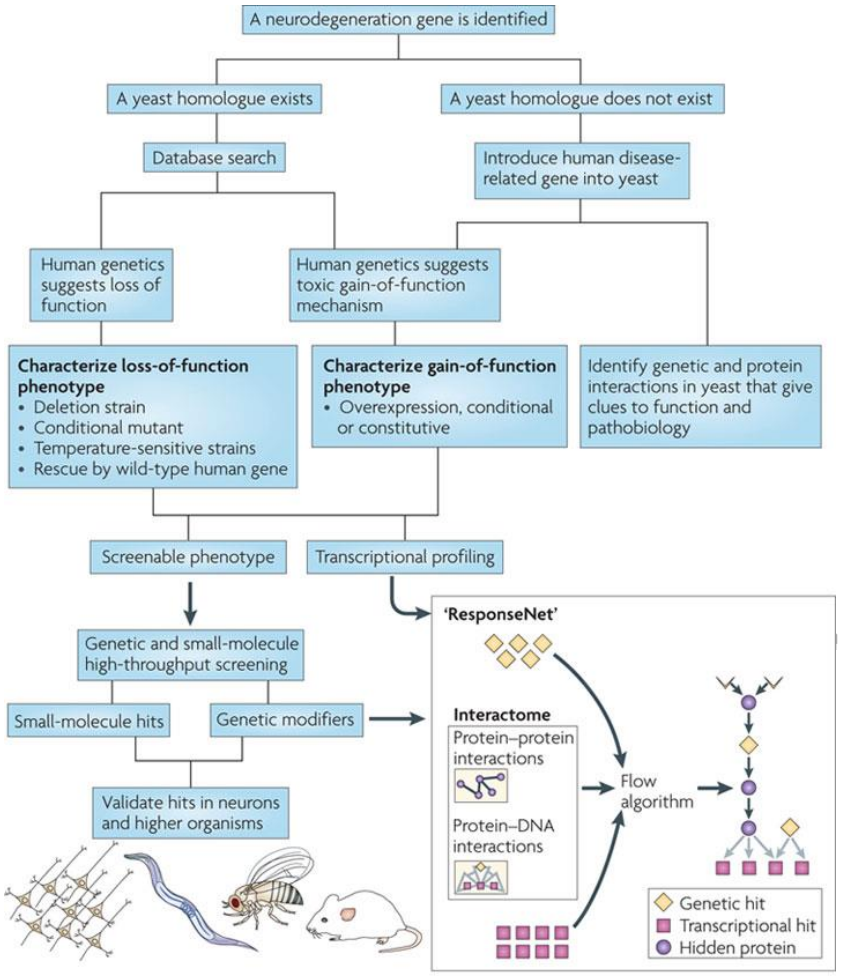
Video1/2



Búsqueda de secuencias similares en la base de datos

- Los investigadores **pueden realizar búsquedas en las bases de datos para comparar secuencias recién identificadas con secuencias que tienen funciones conocidas en otros organismos.**
- Así, si se tiene una secuencia de nucleótidos, puede realizarse una búsqueda BLASTn con objeto de determinar su función a partir de su comparación con **organismos modelo** con secuencias similares y función conocida.

Web1



Nature Reviews | Neuroscience

Organismos modelo

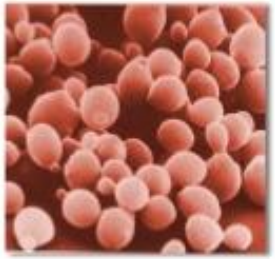
- Un **organismo modelo** es una especie particular que ha sido extensamente estudiada, al haberse asumido que los descubrimientos realizados en este organismo modelo tendrán relevancia en otros organismos.

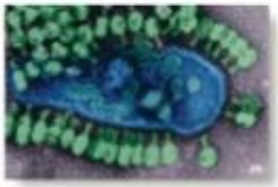
Fenotipo evidente

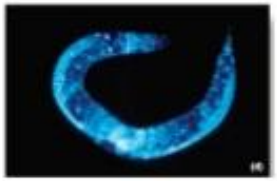
Disponibilidad de mutantes

Ciclo de vida corto

Genoma secuenciado







- Algunos de estos organismos modelos más estudiados *Mus musculus* (ratón común), *Drosophila melanogaster* (mosca de la fruta), *Caenorhabditis elegans* (gusano), *Arabidopsis thaliana* (jaramago), *Escherichia coli* (bacteria) y *Saccharomyces cerevisiae* (levadura).

Organismos modelo

- El genoma de estos organismos ha sido secuenciado, y dado que a lo largo de la diversidad de la vida existen rutas metabólicas y secuencias génicas conservadas, estos organismos se usan como modelos *in vivo* de enfermedades relacionadas con rutas conservadas o con mutaciones en estas secuencias conservadas.
- **La función de los genes puede estudiarse usando organismos modelo con secuencias similares.**
- Se utilizan, por ejemplo, diferentes modelos para investigar los mecanismos moleculares del envejecimiento.

IMAGEN:aging-academic.blogspot.com.es

APLICACIÓN: Uso de la tecnología de bloqueos de genes

- Un método para determinar la función de un gen, es mediante el uso de la **tecnología de bloqueo de genes** (*gene knockout*) en ratones, uno de los organismos modelo.

- Este método implica:

  Paso 1. Reemplazar en células madre embrionarias de un blastocisto de ratón la secuencia funcional de un determinado gen, por una secuencia no funcional del mismo.

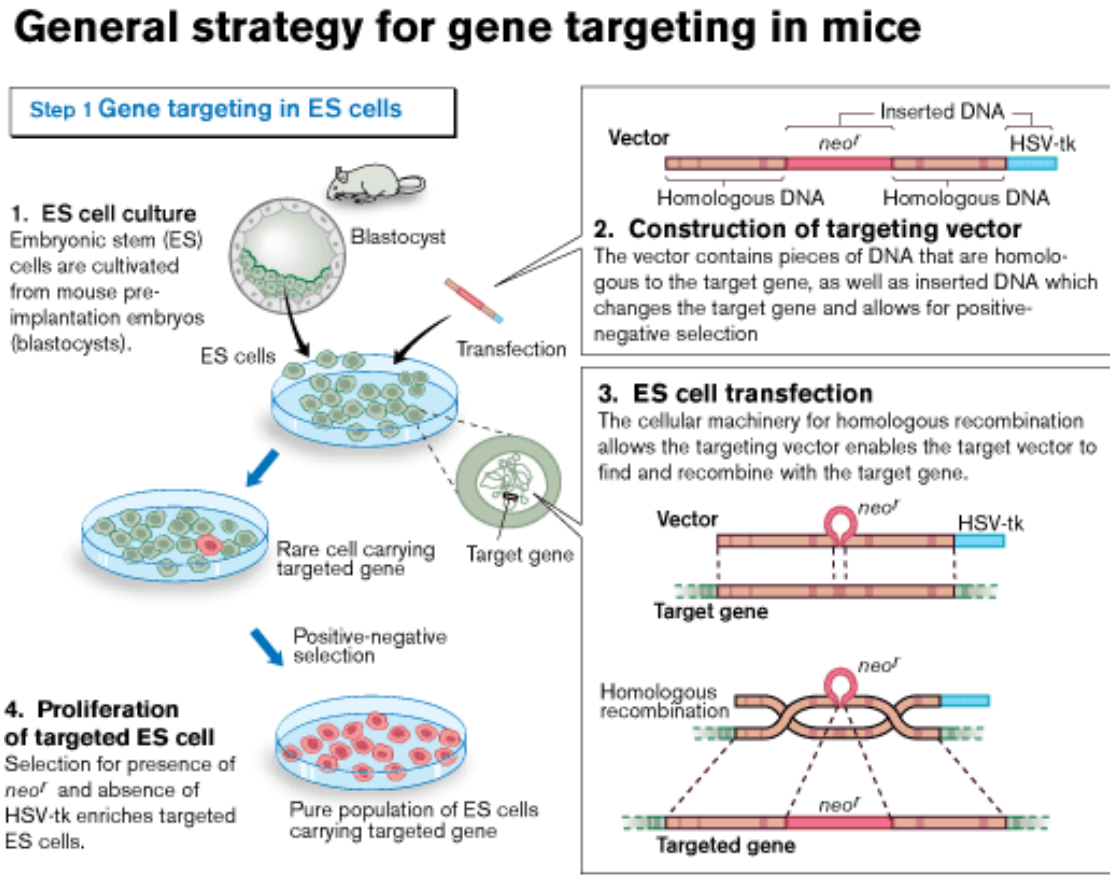


IMAGEN: http://www.nobelprize.org

APLICACIÓN: Uso de la tecnología de bloqueos de genes

Paso 2. Inyectar las células embrionarias con la copia no funcional del gen en otro blastocisto, el cual contendrá una mezcla de células embrionarias con el gen funcional y con el gen no funcional. El blastocisto es implantado es un ratón, cuya progenie se denomina ratón **quimera.**

El ratón quimera es cruzado con un ratón normal. Los heterocigotos obtenidos son cruzados hasta que se obtienen un ratón homocigoto para el gen bloqueado.

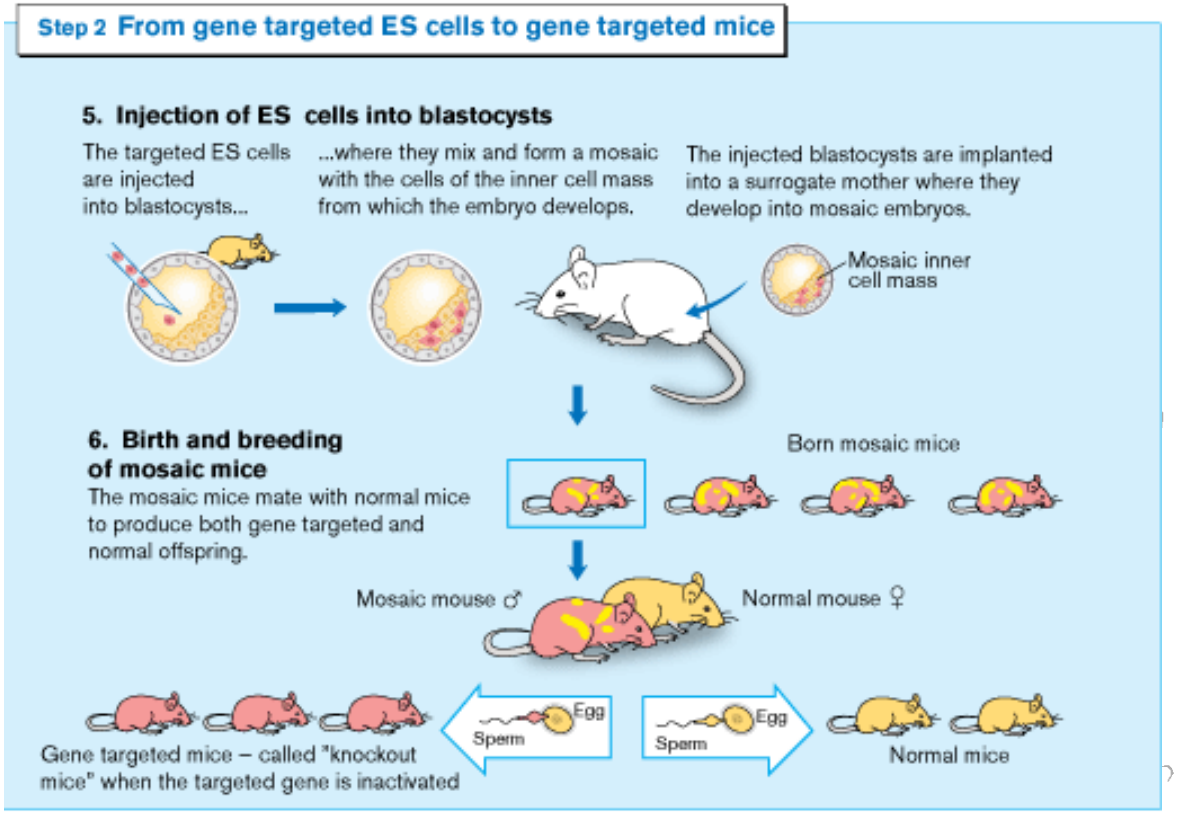


IMAGEN: leptinoblachno.blogspot.com.es/

APLICACIÓN: Uso de la tecnología de bloqueos de genes

- La pérdida de actividad del gen permitirá obtener un fenotipo observable en el ratón, lo que permitirá a los investigadores determinar la probable función del gen.
- Mediante esta técnica se evidenció que la hormona leptina juega un papel en la regulación de la deposición de grasa y el metabolismo energético.

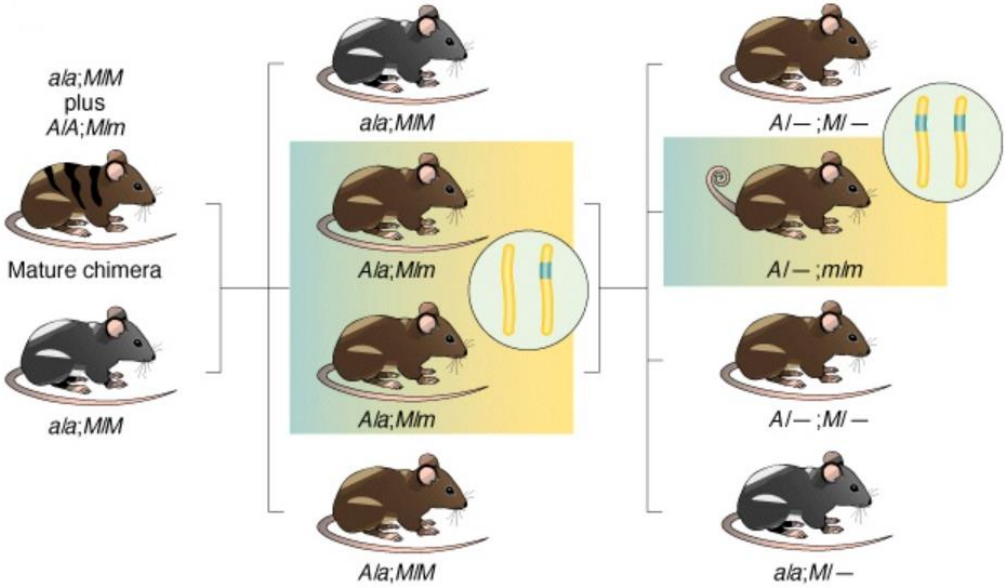


IMAGEN: www.bio.miami.edu/dana/dox/knockout.html

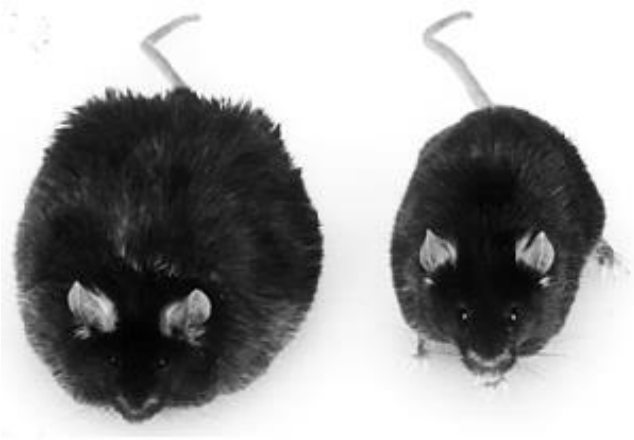


IMAGEN: leptinoblachno.blogspot.com.es/

Web2

Alineación de secuencias

- Las secuencias que son similares entre diferentes organismos, sugieren una relación de tipo evolutivo. Cuanto mayor es esta similitud, más estrecha es la relación, es decir, menos tiempo hace que comparte un mismo ancestro común.
- **Existen softwares de alineación de secuencias que permiten comparar secuencias de distintos organismos.**
- Estos programas, como Clustal Omega o MUSCLE, se basan en algoritmos e indican el grado de significación estadístico del mismo.

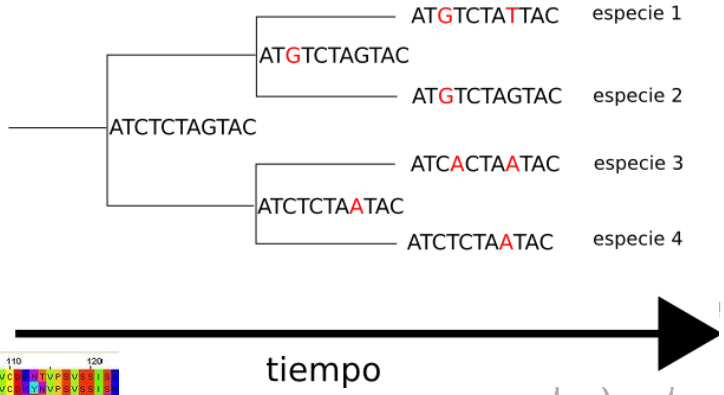


IMAGEN: personales.upv.es

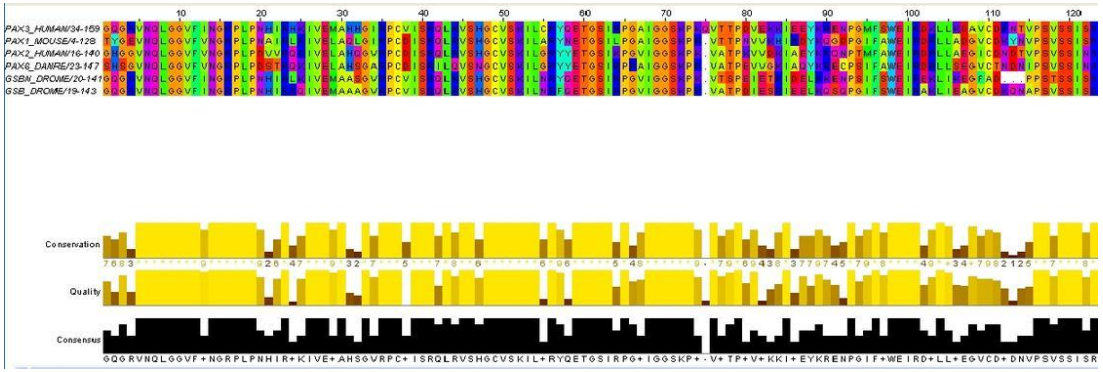
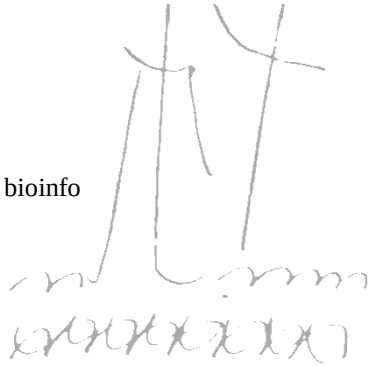
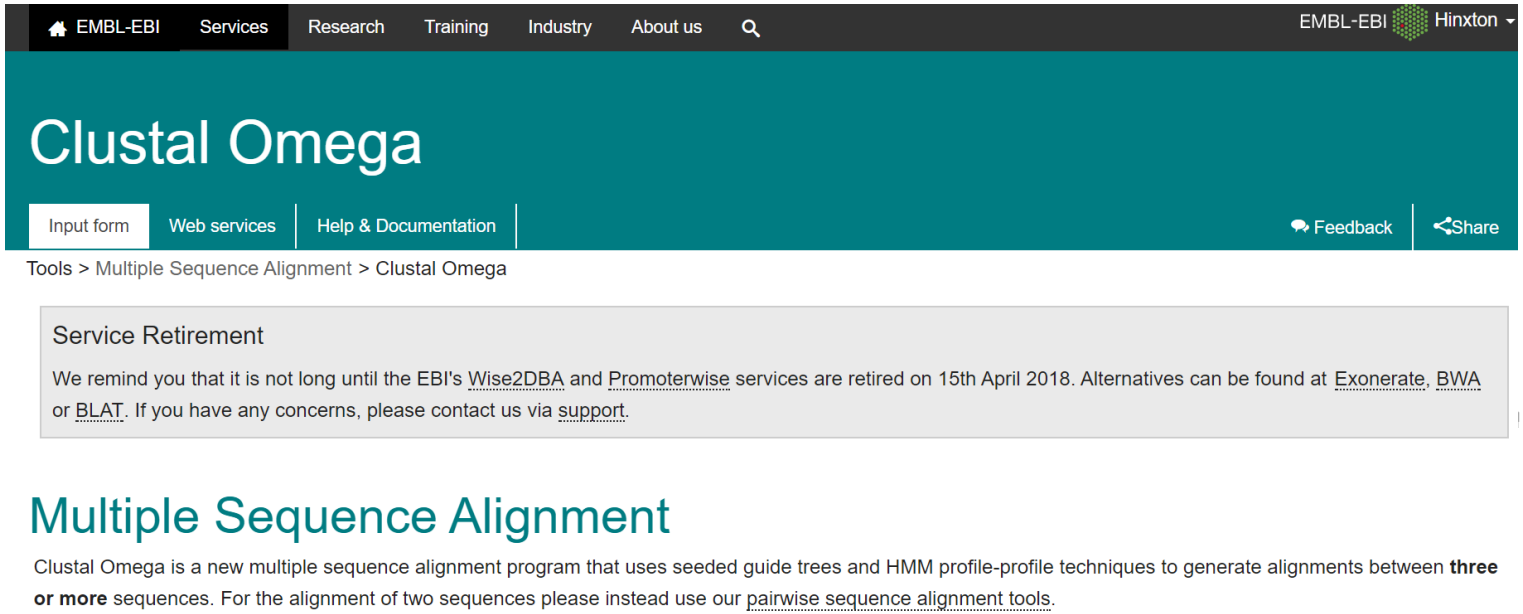


IMAGEN: bioinfo



HABILIDAD: Uso de software para alinear 2 proteínas

- El EMBL (European Molecular Biology Laboratory) posee muchas bases de datos, así como herramientas de bioinformática online. Una de estas herramientas es **Clustal Omega**, un software informático que permite realizar múltiples alineamientos de secuencias.



- Buscar la secuencia de la proteína catalasa en el humano (*homo sapiens*) y la patata (*Solanum tuberosum*) y alinearlas usando Clustal Omega.

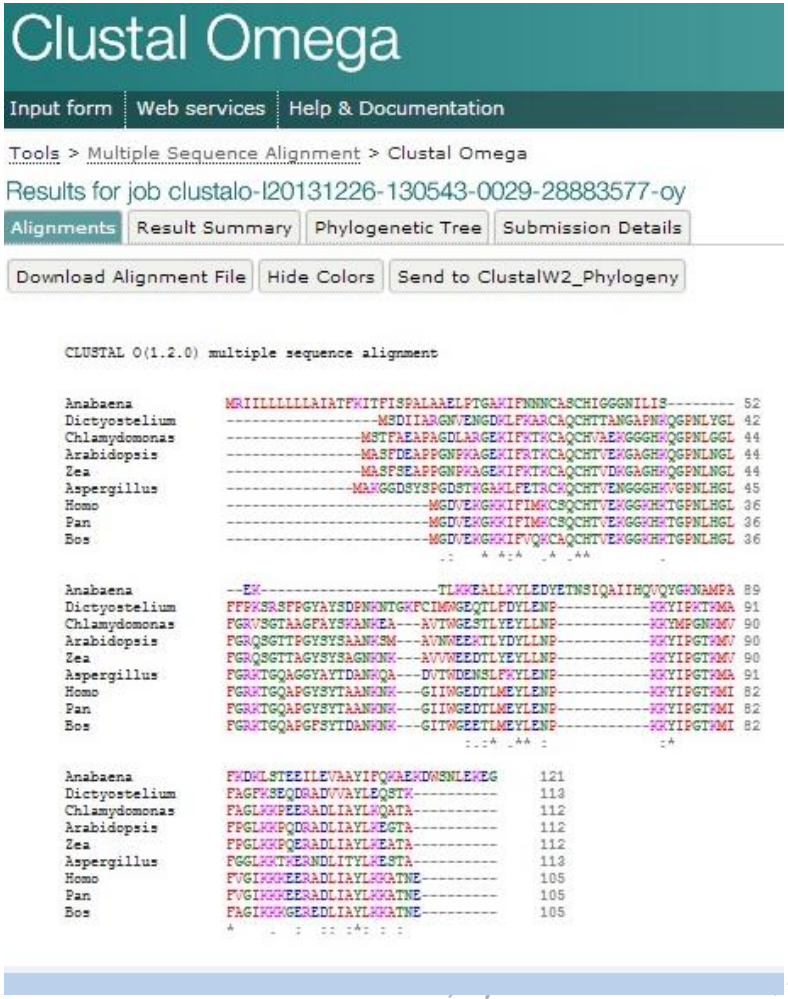
HABILIDAD: Uso de software para alinear 2 proteínas

```
CLUSTAL O(1.2.4) multiple sequence alignment


homo        madsrdpasdqmqhwkeqraaqkadvlttgagnpvgdklnvitvgprgpllvqdvvftde
potato      ----mdps-----ky-rpssaydtpflttnaggpvynnvssltvgprgpvlledyyliek
                 **:     :: .  :* .: .***.**.** :::. :*******:*::*  : ::

homo        mahfdreripervvhakgagafgyfevthditkyskakvfehigkktpiavrfstvages
potato      latfdrekipervvhargasakgffevthdishltcadflrapgaqtpvicrfstvvher
            :* ****:********:**.* *:*******:: : *..:.  * :**:  *****. *

homo        gsadtvrdprgfavkfytedgnwdlvgnntpiffirdpilfpsfihsqkrnpqthlkdpd
potato      gspesirdirgfgvkfynrggnfdlvgnnvpvffnrdaksfpdtiralkpnpkshiqedw
            ** :::** ***.****...**:******.*:** **   **. *:: * **::*:::

homo        mvwdfwslrpeslhqvsflfsdrgipdghrhmngygshtfklvnangeavyckfhyktdq
potato      rtldffsflpeslhtfaffyddvclptdyrhmegfgvhayqlinkegkahyvkfhwkptc
             . **:*: ***** .:*::.*  :* .:***:*:* *:::*:* :*:* * ***:*

homo        giknlsvedaarlsqedpdygirdlfnaiatgkypswtfyiqvmtfnqaetfpfnpfdlt
potato      gvkcmseeeairvggtnhshatkdlydsiaagnypewklfiqtmdpedvdkfdfdpldvt
            *:* :* *:* *:.  : .:. :**:::**:*:**.*.::**.*  ::.:.* *:*:*:*

homo        kvwphkdyplipvgklvlnrnpvnyfaeveqiafdpsnmppgieaspdkmlqgrlfaypd
potato      ktwpedllplipvgrlvlnrnidnffaeneqlafnpghivpgiyysedkllqtrifayad
            *.**..  ******:******  *:*** **:**:*.:: ***  * **:** *:*** *

homo        thrhrlgpnylhipvncpyrarvanyqrdgpmcmqdnqggapnyypnsfgapeqqpsale
potato      tqrhrigpnymqlpvnapkcghh-nnhrdgamnmthrd-eevdyfpsrfdpcrpaeqypi
            *:***:****:::***.*  .:  * :*** * * ..:    :*:*. *.  .   .

homo        hsiqysgevrrfntanddnvtqvrafyvnvlneeqrkrlceniaghlkdaqifiqkka--
potato      pacvlngrrtncvipkennskqageryrswesd-rqdryinkwveslsdprvtheirsiw
             :   .*.  .    :::* .*.   * .  .: ::.*  :: .  *.* ::  : ::

homo        vknftevhpdygshiqalldkynaekpknaihtfvqsgshlaarekanl
potato      isylsqadkscgqkvasrlt----vkptm--------------------
            :. :::.. . *.:: : *      **.
```

- Cuando la proteína es más grande en una especie que en otra, en aquellos huecos donde no hay aminoácidos en esa posición aparece un guión (-).
- Los asteriscos indican aquellos aminoácidos conservados, es decir, aquellos sitios donde ambas proteínas muestran el mismo aminoácido.
- Cuando los aminoácidos son diferentes pero parecidos en forma o en propiedades aparecen (·) o (:), respectivamente.
- Cuando los aminoácidos son diferentes, aparece un hueco.

Alineación de secuencias múltiples

- La filogenia hace referencia a la historia evolutiva de una especie o grupo de especies.
- Un árbol filogenético es un diagrama que describe la filogenia existente entre un grupo de especies.
- **La alineación de secuencias múltiples se usa en estudios de filogenética**.
- Cuando se comparan múltiples secuencias de aminoácidos o nucleótidos, frecuentemente se identifican **secuencias consenso**, es decir, secuencias comunes en todos los organismos.
- Aquellas secuencias relacionadas evolutivamente, y que por tanto presenta una alta similitud, se denominan **secuencias homólogas**.

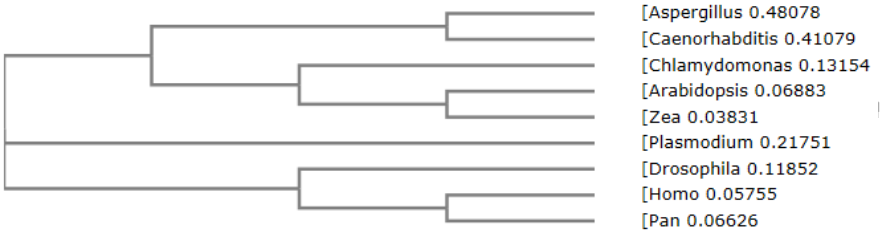


Video3

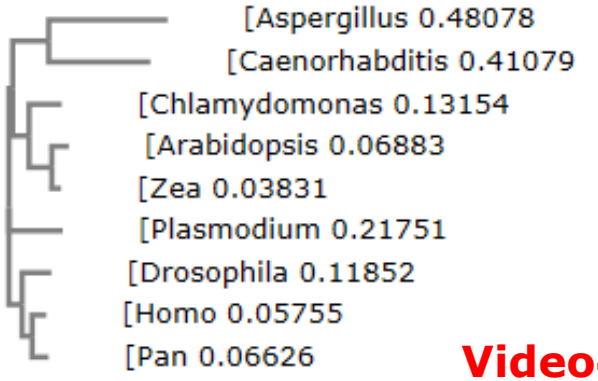
HABILIDAD: Uso de software para elaborar cladogramas y filogramas

- Un **cladograma** es un diagrama usado en la cladística que esquematiza la filogenia o historia evolutiva más probable de un grupo de organismos.
- Un cladograma muestra un patrón de ramificación donde la longitud de sus ramas no representa tiempo o la cantidas relativa de cambios que han ocurrido a lo largo de la rama.

Cladograma



Filograma



Video4

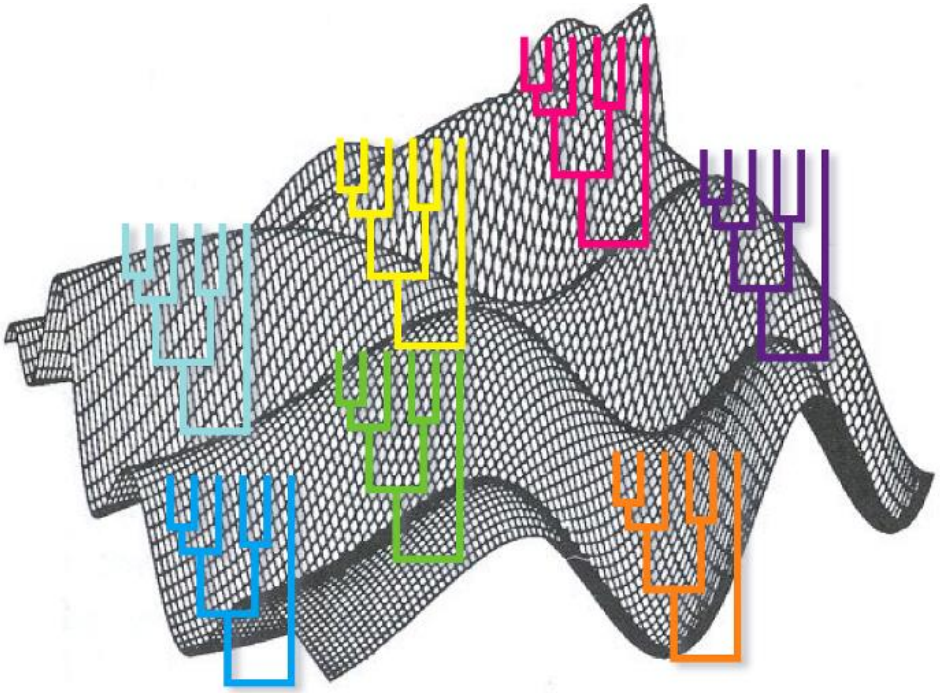
- Un **filograma** es un árbol filogenético donde la longitud de cada rama es proporcional a la cantidad de cambios producidos.

HABILIDAD: Uso de software para elaborar cladogramas y filogramas

- Existen softwares gratuitos online, como Clustal Omega, que permiten la elaboración de **cladogramas** y **filogramas**.
- Para la elaboración de un cladograma, se llevan a caso los siguientes pasos:

1. Se alinean las secuencias de una misma proteína en diferentes especies con objeto de cuantificar las diferencias y semejanzas.

2. Se utilizan algoritmos matemáticos, como el de mínimos cuadrados, máxima parsimonia o máxima verosimilitud, para generar el cladograma, seleccionando el mejor modelo.



Video4

Marcadores de secuencia expresada (EST)

- Cuando un gen se expresa, el ARNm transcrito se encontrará en la célula. Este ARNm puede usarse para buscar el gen a partir del que se produjo, usando la técnica de marcación de secuencia expresada (EST). **Un EST es un marcador de secuencia expresada que se puede usar para identificar genes potenciales**.
- A partir del ARNm los científicos producen un ADNc mediante la transcriptasa inversa, usando este ADNc para sintetizar un EST.

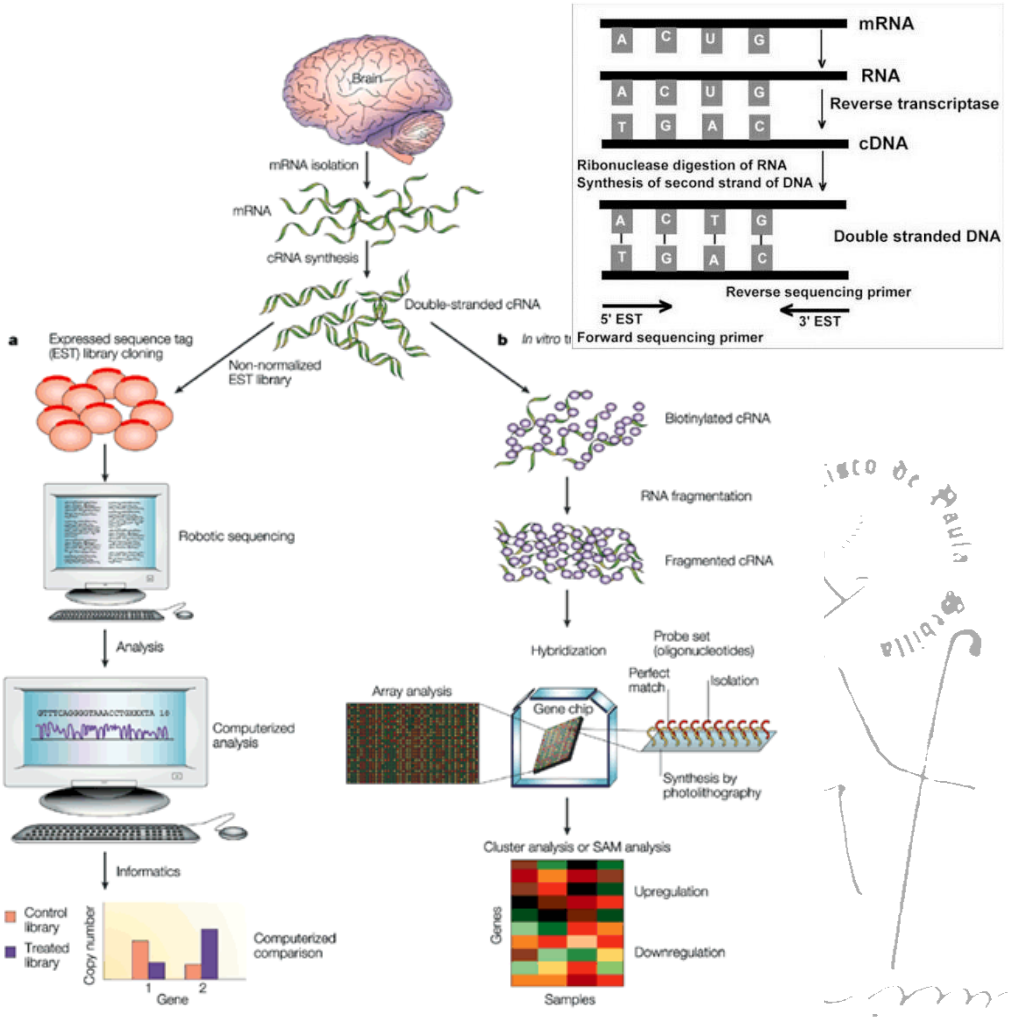


IMAGEN: https://lookfordiagnosis.com

Marcadores de secuencia expresada (EST)

- Los EST son pequeñas secuencias de ADN de 200-500 nucleótidos generadas a partir de los extremos 3’ y 5’ del ADNc.
- Los extremos 5’ poseen una secuencia conservada a lo largo de las especies y dentro de una misma familia génica.
- Sin embargo, en el extremo 3’ es más probable encontrar una secuencia única del gen.
- La localización de un gen en el genoma puede localizarse a través de técnicas de mapeo físico o mediante la búsqueda en bases de datos de EST.

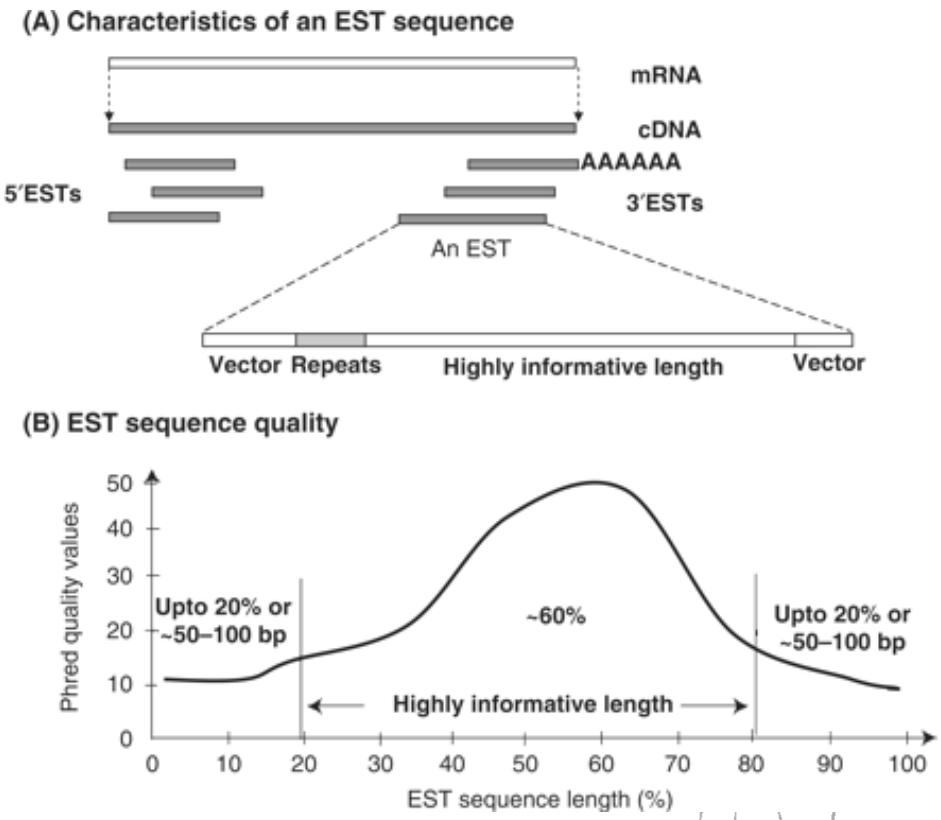


IMAGEN: bib.oxfordjournals.org

APLICACIÓN: Descubrimiento de genes mediante EST

- Debido a la utilidad y la facilidad con la que son generados, se han generado un gran número de EST.
- Las secuencias son depositadas en una base de datos denominada dbEST, que contiene EST de más de 300 organismos. Por tanto, **pueden descubrirse genes mediante prospección de datos EST**.
- Una vez que tienen un EST, los científicos pueden realizar un BLAST para determinar si su secuencia coincide la secuencia de ADN de un gen conocido y de función identificada.

A hitchhiker's guide to expressed sequence tag (EST) analysis

Table 4:
Programs for EST/cDNA sequence to genome DNA alignment

Name#	Website	Category*
BLAT [57]	http://genome.ucsc.edu/cgi-bin/hgBlat	F, W
est2genome [58]	http://bioweb.pasteur.fr/seqanal/interfaces/est2genome.html	F, W
GMAP [56]	http://www.gene.com/share/gmap/	F
MGAlign [59]	http://origin.bic.nus.edu.sg/mgalign	F, W
SSAHA [55]	http://www.sanger.ac.uk/Software/analysis/SSAHA/	F
Sim4 [60]	http://globin.cse.psu.edu/html/docs/sim4.html	F
Splign	http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi	F, W

#Very useful resources are shown in bold.

*F, free for academic users; W, web interface available.

IMAGEN: bib.oxfordjournals.org

Web3

HABILIDAD: Exploración del cromosoma 21

- El proyecto Ensembl recopila información genómica de 75 organismos, permitiendo una exploración detallada de secuencias codificantes y no codificantes de cada cromosoma de estas especies.
- El cromosoma 21 es el cromosoma humano más pequeño y tal vez el más conocido, debido al síndrome de Down o trisomía del par 21, y **puede explorarse mediante esta base de datos**.

IMAGEN: www.ensembl.org/index.html

HABILIDAD: Exploración del cromosoma 21

- Como puede observarse, el brazo p del cromosoma 21 es más corto que el brazo q, por lo que este cromosoma con dos brazos de diferente longitud se clasifica como **submetacéntrico**.

- Como es lógico, el brazo p contiene menos genes al ser de menor longitud que el brazo q.
- Este cromosoma posee más genes que no codifican a proteínas que genes que sí codifican a proteína.

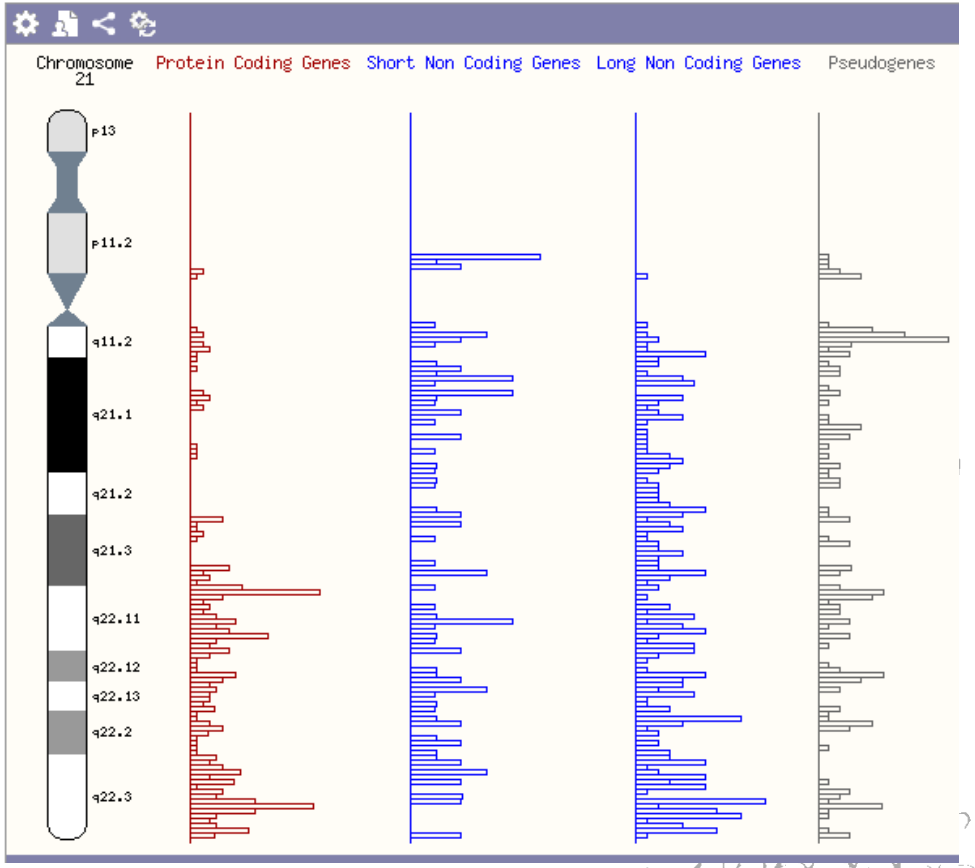


IMAGEN: www.ensembl.org/index.html