Programa Teórico-Práctico de Bioinformática y Genómica Computacional

Gabriel E. Rech

2010

Índice

Prefacio

En respuesta a lo resuelto por el Consejo Académico de la Fundación CAPACIT-AR DEL NOA, en reunión del 18/10/10, Acta N° 88, respecto de los beneficiarios que en el año 2009 hemos recibido ayudas económicas para posgrados y que en la actualidad residimos en el extranjero y según lo propuesto en el Convenio de Retribución suscrito en el Formulario, PROCEDO a enviarles mi DEVOLUCION de interés comunitario.

La devolución consiste en un **Programa Teórico-Práctico de Bioinformática y Genómica Computacional**. Antes que nada, me gustaría dedicar algunas palabras para fundamentar el presente trabajo.

Desde Octubre de 2009, soy beneficiario de una beca doctoral financiada 100% por el Ministerio de Ciencia e Innovación de España, por lo cual me he incorporado al grupo dirigido por el Dr. Michael Thon y la Dra. Serenella Sukno (http://bioinformatica.vil.usal.es/) en el grupo de Genética del CIALE (Centro Hispano-Luso de Investigaciones Agrarias), USAL (Universidad de Salamanca) para trabajar en el proyecto "*A computational and functional approach to understand the evolution of pathogenicity in fungi*" (Análisis bioinformático y funcional de la evolución de genes de patogenicidad en hongos). A su vez, me he incorporado al Programa de Doctorado en Biotecnología Agrícola de la USAL, por lo cual durante el periodo lectivo 2009-2010 he realizado y cumplimentado todos los cursos formativos del doctorado. Incluido dentro de este programa, se encuentra el curso Bioinformática y Genómica Computacional dictado por el Dr. Michael Thon.

Con todo esto, mi trabajo de devolución se fundamenta en lo siguiente:

- La bioinformática es en la actualidad uno de los campos de la ciencia más dinámicos y con más proyección.
- La posibilidad de trabajar diariamente y de haber tomado el curso Bioinformática y Genómica Computacional con el Dr. Michael Thon.
- La bioinformática es un área de vacancia en nuestra región, para la cual no es fácil detectar y promover vocaciones orientadas en esa dirección, por lo cual se torna fundamental fomentar su utilización y dar a conocer su potencial.
- En Salta no existe una carrera o especialización en bioinformática.
- No hay suficiente información disponible en castellano.

El presente **Programa Teórico-Práctico de Bioinformática y Genómica Computacional** no pretende ser exhaustivo. Está basado en el curso dictado por el Dr. Michael Thon[1] en el programa de doctorado en Biotecnología Agrícola de la USAL, con algunas aclaraciones e información complementaria agregados por mi parte y otras fuentes. Cabe aclarar que dicho curso tiene una duración de 15 días y se dicta completamente en inglés, por lo cual he procedido a su traducción al castellano para que, de esta manera, pueda estar disponible para un mayor número de personas interesadas. Además he adaptado los ejercicios prácticos para que pudieran ser

[1] El CV del Dr. Michael Thon se puede consultar en http://bioinformatica.vil.usal.es/?p=6.

desarrollados completamente a través de la web, por lo que sólo es necesaria una computadora con conexión a internet para realizarlos.

Antes de dar paso al desarrollo del trabajo, me gustaría agradecer la ayuda económica brindada por la FUNDACION CAPACIT-AR DEL NOA, ya que cubrieron completamente mis gastos de traslado desde Salta a Salamanca, debido a que la beca con la que he sido beneficiado no cubría estos gastos. Sin su colaboración, hubiera sido muy difícil para mí poder aprovechar esta oportunidad.

Introducción

La era genómica

El fin del siglo XX ha visto una enorme explosión en la cantidad de información biológica disponible debido a los enormes avances en los campos de la biología molecular y la genómica, por esta razón se la ha denominado la era genómica.

La bioinformática es la aplicación de tecnología informática para la gestión y análisis de datos biológicos. Esto es, la utilización de computadoras para recopilar, almacenar, analizar y combinar datos biológicos.

A su vez, la bioinformática es un área de investigación interdisciplinaria, y actúa como interfaz entre las ciencias biológicas y las computacionales. El objetivo final de la bioinformática es descubrir la riqueza de la información biológica escondida dentro de una gran masa de datos y obtener una visión más clara de la biología fundamental de los organismos. Este nuevo conocimiento podría tener un profundo impacto en campos tan variados como la salud humana, la agricultura, el medio ambiente, energía y biotecnología .

Importancia de la bioinformática

El mayor desafío que enfrenta la biología molecular en la actualidad, es el de dar sentido a la abundancia de datos que se están producido gracias a los proyectos de secuenciación de genomas. Tradicionalmente, la investigación en biología molecular se realizaba por completo en el laboratorio experimental, pero el formidable aumento en la cantidad de datos que se producen en la era genómica, ha visto la necesidad de incorporar las computadoras a este proceso de investigación.

La generación de secuencias, y su posterior almacenamiento, interpretación y análisis son tareas totalmente dependientes de las computadoras. Sin embargo, la biología molecular de un organismo es un tema muy complejo, por esto para su comprensión es necesaria la investigación a diferentes niveles: genómico, proteómico, transcriptómico y metabolómico. A raíz de la explosión en el volumen de datos genómicos, se ha observado un incremento similar de datos en los campos de la proteómica, transcriptómica y metabolómica.

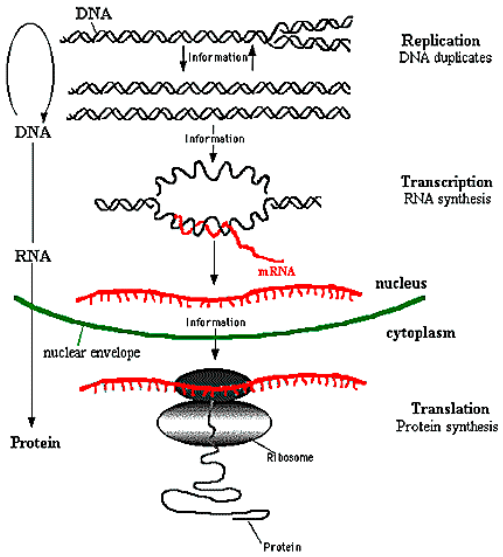
El principal desafío que enfrenta la comunidad bioinformática hoy, es el del almacenamiento inteligente y eficiente de esta masa de datos, de manera de proporcionar un acceso fácil y fiable a ellos. Los datos en sí no tienen sentido si no son analizados, y la cantidad de datos actuales hace que sea imposible, incluso para un biólogo capacitado, empezar a interpretarlos de una forma manual. Por lo tanto, es necesario el desarrollo de herramientas computacionales que permitan la extracción de información biológicamente significativa .

¿Qué significa “Información Biológica”?

Una respuesta muy básica a esta pregunta sería: “conjunto organizado de datos provenientes de los seres vivos”. Pero aquí haremos referencia a los datos derivados

especialmente de la biología molecular: secuenciación de genomas, secuencia y estructura de proteínas y estudios sobre la expresión simultánea de muchos genes bajo muchas condiciones diferentes.

La información biológica se encuentra codificada en los genes y se expresa a partir (o mediante) los genes. Esta idea se refleja en el **Dogma Central de la Biología Molecular** (siempre teniendo en cuenta aquellos elementos que implican la ampliación de este dogma: priones, ribozimas, transcripción inversa, splicing alternativo).



The Central Dogma of Molecular Biology

La biología se enfrenta con el problema de la decodificación del lenguaje biológico, esto es: ¿Cómo se codifica la información en los genes?, ¿Cómo y cuándo se traduce esta información?, ¿Qué determina la estructura de las proteínas?, ¿Cómo se determina la función de las proteínas?, etc…

Existen tres procesos biológicos centrales en torno al los cuales se deben desarrollar las herramientas de la bioinformáticas:

1. La secuencia de ADN determina la secuencia de la proteína.
2. La secuencia de la proteína determina la estructura de la proteína.
3. La estructura de la proteína determina su función.

La integración de la información obtenida a partir de estos procesos biológicos es la clave que permitirá alcanzar el objetivo a largo plazo de la completa comprensión de la biología de los organismos .
Estos procesos están directamente relacionados con el **Paradigma central de la bioinformática:**

Información Genética ⇨ **Estructura Molecular** ⇨ **Función Bioquímica** ⇨ **Fenotipo**

La **Información Genética** de un organismo se encuentra en el ADN o, en el caso de algunos virus, en el ARN. La porción de genoma que codifica una proteína o un ARN se conoce como gen. Esos genes que codifican proteínas están compuestos por unidades de trinucleótidos llamadas codones, cada una de los cuales codifica un aminoácido. Cada subunidad nucleotídica está formada por un fosfato, una desoxirribosa y una de las cuatro posibles bases nitrogenadas. Las bases purínicas adenina (A) y guanina (G) son más grandes y tienen dos anillos aromáticos. Las bases pirimidínicas citosina (C) y timina (T) son más pequeñas y sólo tienen un anillo aromático. En la configuración en doble hélice, dos cadenas de ADN están unidas entre sí por puentes de hidrógeno en una asociación conocida como emparejamiento de bases. Además, estos puentes siempre se forman entre una adenina de una cadena y una timina de la otra y entre una citosina de una cadena y una guanina de la otra. Esto quiere decir que el número de residuos A y T será el mismo en una doble hélice y lo mismo pasará con el número de residuos de G y C. En el ARN, la timina (T) se sustituye por uracilo (U), y la desoxirribosa por una ribosa.

Cada gen codificante de proteína se transcribe en una molécula plantilla, que se conoce como ARN mensajero o ARNm. Éste, a su vez, se traduce en el ribosoma, en una cadena aminoacídica o polipeptídica. En el proceso de traducción se necesita un ARN de transferencia específico para cada aminoácido con el aminoácido unido a él covalentemente, guanosina trifosfato como fuente de energía y ciertos factores de traducción. Los ARNt tienen anticodones complementarios a los codones del ARNm y se pueden “cargar” covalentemente en su extremo 3' terminal CCA con aminoácidos. Los ARNt individuales se cargan con aminoácidos específicos por las enzimas llamadas aminoacil ARNt sintetasas, que tienen alta especificidad tanto por aminoácidos como por ARNt. La alta especificidad de estas enzimas es motivo fundamental del mantenimiento de la fidelidad de la traducción de proteínas.

Hay $4^3 = 64$ combinaciones diferentes de codones que sean posibles con tripletes de tres nucleótidos: los 64 codones están asignados a aminoácido o a señales de parada en la traducción. Si, por ejemplo, tenemos una secuencia de ARN, UUUAAACCC, y la lectura del fragmento empieza en la primera U (convenio 5' a 3'), habría tres codones que serían UUU, AAA y CCC, cada uno de los cuales específica un aminoácido. Esta secuencia de ARN se traducirá en una secuencia aminoacídica de tres aminoácidos de longitud.

La **Estructura Molecular** de las proteínas reúne las propiedades de disposición en el espacio de las moléculas de proteína que provienen de su secuencia de aminoácidos, las características físicas de su entorno y la presencia de compuestos, simples o complejos que las estabilicen y/o conduzcan a un plegamiento específico, distinto del espontáneo. Por ello, deriva de sus componentes, es decir de la propia estructura de los aminoácidos, de cómo interaccionan químicamente éstos, de forma jerarquizada y específica, y evidentemente está en relación con la **función a acometer en el destino celular**, y por lo tanto determinará una (o varias) **característica fenotípica** en el organismo.

Los bloques temáticos de la bioinformática

En la actualidad existen numerosas actividades que los biólogos, agrónomos o médicos realizan mediante la informática. Entre las más importante se destacan: **Ensamblaje de Secuencias** (*Sequence assembly*), búsqueda de secuencias en **Bases de Datos**, determinación de genes en secuencias genómicas o búsqueda de la función de una proteína. De una manera más general, la bioinformática incluye tareas como :

- **Organización de la información**
  - Creación y mantenimiento de bases y bancos de datos.
- **Acceso a la información**
  - Búsqueda de información en bases de datos.
  - Comparación de información con la de las bases de datos.
- **Algoritmia**
  - Algoritmos de búsqueda, alineamientos, etc.
- **Búsqueda de genes**
  - Modelización y análisis estadístico.
  - Programación dinámica (ensamblado)
- **Proteómica**
  - Estudio de la estructura de las proteínas.
  - Predicción de la función de las proteínas.
- **Genómica**
  - Genómica comparativa.
  - Genómica funcional.

Un experimento en la computadora no es distinto de cualquier experimento en la mesada, los resultados deben contestar una pregunta concreta y deben ser reproducibles por otra persona que utilice el mismo método. Para estos es imprescindible antes que nada **identificar el problema** al cuál queremos encontrarle una solución, por ej: – cuál es el mecanismo catalítico de la enzima X?
Luego debemos **identificar las herramientas necesarias** para resolver el problema, como ser búsquedas de secuencias similares, alineamientos múltiples, detección de profiles y motivos, modelado de la estructura tridimensional, evaluación del modelo. También es necesario definir **criterios de satisfacción** (éxito del experimento), para esto es preciso saber que prácticamente todos los métodos computacionales producen resultados. Una búsqueda utilizando BLAST casi siempre produce algún hit, por esto es necesario distinguir resultados significativos del ruido. Por último y para nada menos importante: **hay que entender cómo funcionan los programas**, en qué algoritmos están basados, que puntos débiles tienen, etc.

A continuación abordaremos algunas cuestiones relacionadas con las principales tareas que, como investigadores en el área de la biología, debemos realizar hoy en día, casi de rutina. Además se adjuntan propuestas prácticas asociadas.

Formatos de archivos para representar secuencias

Existen varias formas de guardar las secuencias en un fichero de texto. Si sólo queremos guardar la secuencia se puede crear un fichero se texto y simplemente escribir la secuencia. Esto se conoce como secuencia en texto plano. El fichero sólo puede contener caracteres IUPAC para secuencias (Ver ANEXO 1) y espacios, los números no están permitidos.
Secuencia en texto plano:

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGC
CCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAG
CAGCCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGG
AGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGAC
AGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGT
CACGC
```

Este formato tiene algunas limitaciones, no puede haber más que una secuencia dentro del fichero y no puede incluirse el nombre de la secuencia dentro del fichero.

Formato Fasta. La secuencia comienza con el signo ">", seguido del nombre de la secuencia (*sequence ID*) y luego puede o no contener una descripción. A continuación sigue la secuencia propiamente dicha con el formato en texto plano. Pueden incluirse varias secuencias en un mismo fichero (**multi fasta format**).

Secuencia en formato fasta:

```
>nombre_secuencia1 descripción
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGC
CCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAG
CAGCTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGA
GAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACA
GAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCT
CACGC
>nombre_secuencia2 descripción
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGC
CCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGA
```

Una secuencia puede tener más atributos. Podemos querer incluir el nombre del investigador que la secuenció o un identificador para una base de datos. Los formatos más extendidos con estos atributos extras son el **embl** y el **genbank**.

Secuencia en formato embl:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg  60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg 120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc 180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag 240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga 300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca 360
     gacctgaa                                                         368
//
```

Secuencia en formato genbank:

```
LOCUS       XLU64442                8490 bp    mRNA    linear   VRT 01-FEB-1997
DEFINITION  Xenopus laevis adenomatous polyposis coli mRNA, complete cds.
ACCESSION   U64442
VERSION     U64442.1  GI:1809289
KEYWORDS    .
SOURCE      Xenopus laevis (African clawed frog)
  ORGANISM  Xenopus laevis
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
            Xenopodinae; Xenopus; Xenopus.
REFERENCE   1  (bases 1 to 8490)
  AUTHORS   Vleminckx,K., Wong,E., Guger,K. and Gumbiner,B.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-JUL-1996) Cellular Biochemistry and Biophysics
            Program, Memorial Sloan Kettering Cancer Center, York Avenue 1275,
            New York, NY 10021, USA
COMMENT     On Feb 1, 1997 this sequence version replaced gi:1552959.
FEATURES             Location/Qualifiers
     source          1..8490
                     /organism="Xenopus laevis"
                     /mol_type="mRNA"
                     /db_xref="taxon:8355"
     CDS             1..8490
                     /function="tumor suppressor protein"
                     /note="APC"
                     /codon_start=1
                     /product="adenomatous polyposis coli"
                     /protein_id="AAB41671.1"
                     /db_xref="GI:1809290"
                     /translation="MAAASYDQLVKQVEALTMENTNLRQELEDNSNHLTKLETEATNM
                     KEVLKQLQGSIEDEAMASSGPIDLLERFKDLNLDSSNIPAGKARPKMSMRSYGSREGS
                     V"
ORIGIN
        1 atggctgctg cttcgtatga tcagttggta aagcaagtgg aggcgttgac gatggagaat
       61 acaaaccttc gacaagaatt agaagacaat tccaatcatc ttacaaaact tgagactgag
      121 gcaacaaaca tgaaggaggt tctcaagcag ctgcaaggaa gtattgagga tgaggcaatg
      181 gcttcttctg gcccaattga tcttctggaa cgttttaaag acctaaatct ggacagcagt
     8461 tctggttctt atctggtgac ttcggtttga
//
```

Los campos principales en una secuencia en formato Genbank son:

Locus: El nombre del locus, el tamaño de la secuencia, el tipo de molécula y su topología
Definition: Una breve descripción de la secuencia.
Accession: El identificador único de la secuencia en la genbank.
Keywords: Palabras clave que describen la secuencia
Source: Nombre común del organismo.
Organism: Identificación completa del organismo
Reference: Citas bibliográficas relacionadas con la secuencia.
Features: Detalles adicionales sobre la secuencia.

Existen otros muchos formatos de secuencia, pero estos son los más comunes . Hay varias utilidades para convertir secuencias de un formato a otro (http://mybio.wikia.com/wiki/Sequence_format_conversion).

Uno de los más utilizados es **"seqret" de Emboss.** Emboss es un paquete de software Open Source destinado a la biología molecular. Contiene múltiples aplicaciones para el análisis de secuencias (http://emboss.sourceforge.net/).

Para acceder directamente a la aplicación: http://genome.nci.nih.gov/tools/reformat.html

PRÁCTICA:

0. Ingresar a la página http://genome.nci.nih.gov/tools/reformat.html, en el campo *Input Sequence format* copiar y pegar la secuencia dada como ejemplo anteriormente para el formato genbank, en Input Format debemos colocar GenBank (gb) y en output probar con diferentes formatos (fasta, embl, etc.).

Lecturas recomendadas:

Para aprender más sobre formatos de secuencias, puede acceder a:
http://www.ebi.ac.uk/2can/tutorials/formats.html o
http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

Bases de datos biológicas

Las bases de datos biológicas son conjuntos de datos consistentes, que se almacenan de manera uniforme y eficiente. Estas bases de datos contienen información de un amplio espectro de la biología molecular. Las **bases de datos primarias** contienen anotaciones e información de la estructura de secuencias de ADN y proteínas, como así también perfiles de expresión de proteínas .
Las **bases de datos secundarias** o "derivadas" se denominan así porque contienen resultados del análisis de las primarias, incluyendo información sobre **motivos**, variantes, mutaciones y relaciones evolutivas. Ej: **Refseq** (Colección curada de GenBank en NCBI) o **Unigene** (Clustering de ESTs en NCBI).

La información relacionada con la bibliografía (referencias, autores, etc.) está contenida en **bases de datos bibliográficas** como **PubMed** (http://www.ncbi.nlm.nih.gov/pubmed).

Es esencial que las bases de datos sean fácilmente accesibles y que proporcionen un sistema de consulta intuitivo para permitir a los investigadores obtener información muy específica sobre un tema biológico determinado. Los datos deben ser proporcionados de manera clara, coherente y con la ayuda de herramientas de visualización que faciliten la interpretación biológica.

También existen **bases de datos especializadas** que han sido puestas a punto específicamente para determinadas cuestiones. Algunos ejemplos son:

- Base de datos del EMBL, también conocida como **EMBL-Bank** (European Molecular Biology Laboratory: http://www.ebi.ac.uk/embl/) especializada en secuencias de nucleótidos.
- Las bases de datos **UniProtKB** (Universal Protein Knowledgebase: http://www.uniprot.org/help/uniprotkb) y **Swiss-Prot** especializadas en secuencias de proteínas.
- La **PDBe** (Protein Data Bank in Europe: http://www.ebi.ac.uk/pdbe/), una base de datos de estructura 3D de proteínas.

También es necesario que los investigadores sean capaces de integrar la información obtenida a partir de bases de datos heterogéneas de una manera fiable, con el fin de poder tener una idea clara de la información biológica relacionada con el tema que están buscando. Para esto existen también algunas herramientas como el **SRS** (Sequence Retrieval System: http://srs.ebi.ac.uk/) es una poderosa herramienta, suministrada por el **EBI** (European Bioinformatics Institute) que vincula la información de más de 150 recursos heterogéneos.

El Laboratorio Europeo de Biología Molecular (**EMBL**) fue el primero en generar una base de datos unificada de secuencias biológicas en 1982. Otros dos centros se unieron posteriormente a este esfuerzo: el Centro Nacional de Información Biotecnológica de EEUU (**NCBI**: National Center for Biotechnology Information) y el Banco de Datos de ADN de Japón (**DDBJ**: DNA Data Bank of Japan). (Nota: Hoy en día no se publican secuencias en revistas científicas, más bien se envían directamente a alguna de estas bases de datos y se cita en el artículo el número de acceso de la base de datos).

National Center for Biotechnology Information (NCBI)

El NCBI se fundó en 1988 con la intención de almacenar información biológica, hacer investigación en biología computacional y desarrollar herramientas de análisis de información biológica. En la actualidad, en los servidores del NCBI se encuentra almacenada la mayor cantidad de información primaria en biología molecular.

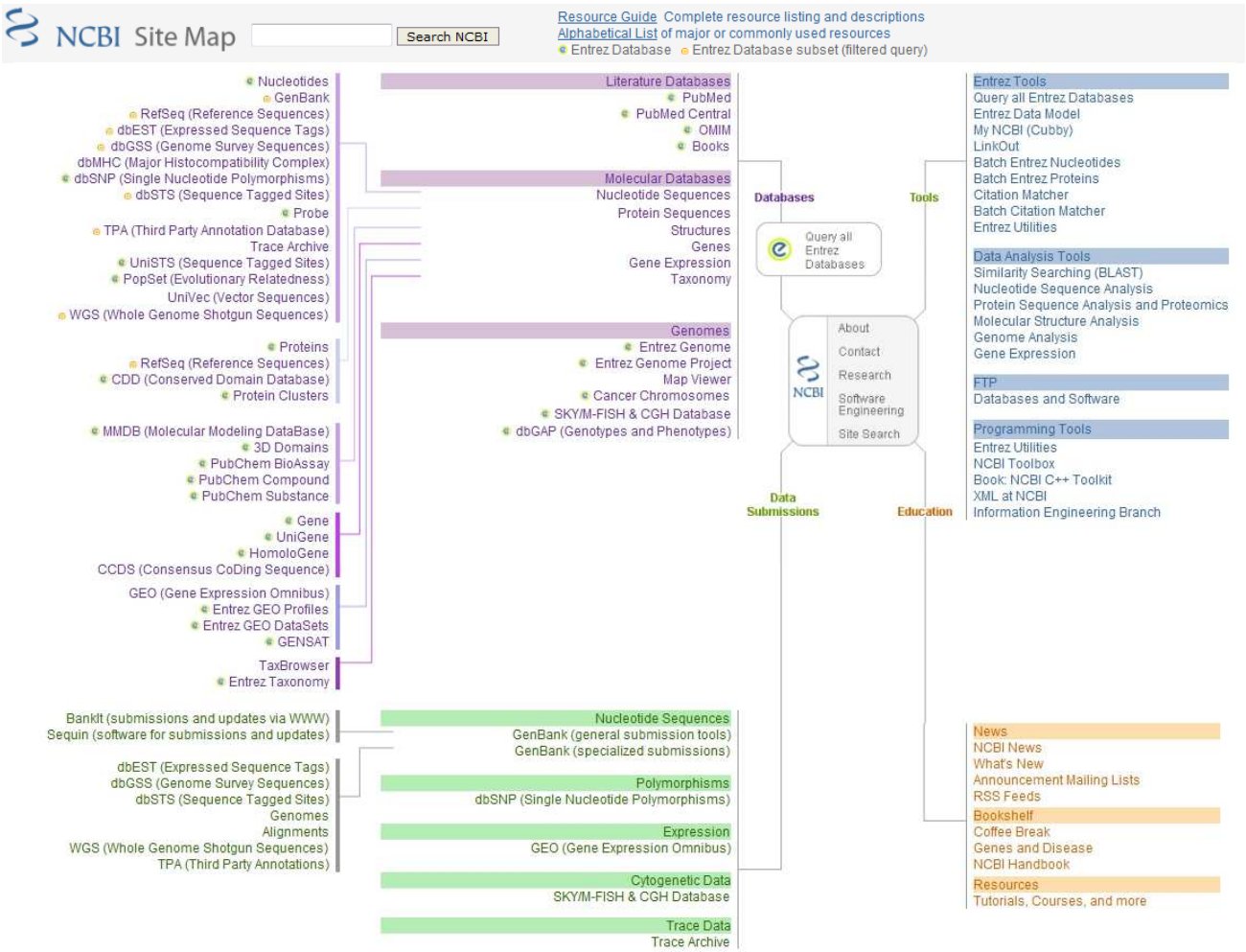
Página Web: http://www.ncbi.nlm.nih.gov/

El NCBI almacena y constantemente actualiza la información referente a las secuencias genómicas depositadas en **GenBank**, contiene el índice de artículos científicos referentes a biomedicina, biotecnología, bioquímica, genética y genómica **PubMed** del cuál hablamos previamente, además de una recopilación de enfermedades genéticas humanas en **OMIM**, base de datos taxonómica y otros datos biotecnológicos de relevancia en diversas bases de datos.

Todas las bases de datos del NCBI están disponibles en línea de manera gratuita, y son accesibles usando el buscador **Entrez**.

El NCBI ofrece además algunas herramientas bioinformáticas para el análisis de secuencias de ADN, ARN y proteínas, siendo **BLAST** una de las más usadas.

Para tener una idea general de la cantidad de recursos disponibles en la web del NCBI, la imagen a continuación muestra un **mapa del sitio**, al cual también se puede acceder mediante: http://www.ncbi.nlm.nih.gov/Sitemap/index.html.



GenBank

GenBank es la base de datos de secuencias genéticas del NIH (National Institutes of Health de Estados Unidos), una colección de secuencias de ADN de disponibilidad pública. Realiza una puesta al día cada dos meses.

GenBank es parte de **International Nucleotide Sequence Database Collaboration**, que está integrada por: la base de datos de ADN de Japón (**DDBJ**), el **EMBL** y el GenBank en el **NCBI**. Estas organizaciones intercambian datos diariamente. GenBank y sus colaboradores reciben secuencias genéticas producidas en laboratorios de todo el mundo, procedentes de más de 100.000 organismos distintos. GenBank continúa creciendo a ritmo exponencial, **doblando** la cantidad de información contenida cada 10 meses.

Los envíos directos (*direct submissions*) a GenBank se hacen utilizando **BankIt**, que es un formato basado en la Web, o el programa (*stand-alone)* **Sequin**. Tras la recepción de una secuencia, el personal de GenBank asigna un número de acceso a la secuencia y realiza controles de calidad. Luego, las secuencias son publicadas en la base de datos pública, desde donde las entradas son recuperables por **Entrez** o se pueden descargar por FTP. La mayoría de las secuencias de **Expressed Sequence Tag** (EST), Sequence Tagged Site (STB), Genome Survey Sequence (SSG) y High-Throughput Genome

Sequence (HTGS) son enviadas por los grandes centros de secuenciación. El *GenBank direct submissions group* también procesa secuencias completas de genomas microbianos.

GenBank Division (divisiones o particiones dentro de GenBank)

Existen dos maneras principales por medio de las cuales las secuencias están organizadas dentro de Genbank: por tipo de **Organismo** y por tipo de tecnología utilizada para la secuenciación ("**Tecnológica**"). La división de GenBank a la que pertenece un registro, se indica con una abreviatura de tres letras.

Organismo:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences

Tecnológica:

12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences

Nota: Si observamos en el archivo de ejemplo para el formato genbank (página 12) podemos observar que en la primera línea (LOCUS), contiene la abreviatura VRT, indicando que la secuencia pertenece a un vertebrado.

Entrez

El **Entrez** (Global Query Cross-Database Search System) es un potente motor de búsqueda que por medio de su web (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi) permite a los usuarios realizar consultas integradas, proporcionando acceso a todas las bases de datos simultáneamente a partir de un texto simple y con una única interfaz de usuario.

Entrez encontrará eficientemente secuencias (ADN y proteínas), estructuras, referencias y libros de texto relacionadas con la búsqueda realizada. El sistema Entrez proporciona gráficos de las secuencias de genes y proteínas y mapas de cromosomas.

Páginas de ayuda:

Documentos generales de ayuda:
http://www.ncbi.nlm.nih.gov:80/entrez/query/static/help/helpdoc.html
Tabla de campos de búsqueda:
http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers

Algunos trucos para las búsquedas en Entrez:

Los campos más comunes de búsqueda son:

[ORGN] → Organismo o taxonomía
[AUTHOR] → Autor/es del artículo o personas que han subido la secuencia a GenBank
[TITL] → Título del artículo o descripción de la secuencia de AND o proteína.

Operadores Booleanos:

Utilizando los operadores **AND** (unión)**, OR** (intersección)**, NOT** (subconjunto) podemos filtrar o especificar mejor los resultados de la búsqueda (deben estar en mayúsculas). Es importante tener en cuenta que estos operadores son evaluados de izquierda a derecha, por lo que podemos usar paréntesis () para cambiar el orden de evaluación.

Ejemplo: Búsqueda de transposones en hongos

• (transposon[TITL] OR transposase[TITL]) AND Fungi[ORGN]

– 867 results in **Nucleotide:** Core subset of nucleotide sequence records
– 357 results in **Protein:** sequence database

• (transposon[TITL] OR transposase[TITL] OR "transposable element"[TITL]) AND Fungi[ORGN]

– 1055 results in **Nucleotide:** Core subset of nucleotide sequence records
– 556 results in **Protein:** sequence database

• transpos*[TITL] AND Fungi[ORGN]
(El asterisco * indica que deben buscarse todas las palabras que contengan al inicio "transpos" y luego cualquier otro complemento: transpos**on**, transpos**ase**, transpos**able**)

– 1096 results in **Nucleotide:** Core subset of nucleotide sequence records
– 633 results in **Protein:** sequence database

PRÁCTICA:

1. Usar Entrez para buscar todas las G protein-coupled receptor (GPCR, G-protein coupled receptor) de Arabidopsis, Medicago, u Hongos (fungi).
2. Realizar la búsqueda con y sin el uso de la búsqueda por campos ([ORGN], etc) y con y sin operadores booleanos (AND, OR, NOT).
3. Comparar los resultados.

UniProt

UniProt es la fuente universal de proteínas, un repositorio central de datos de proteínas creado por la combinación de: UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef), y la UniProt Archive (UniParc).

UniProt es una colaboración entre el European Bioinformatics Institute (EBI), el Swiss Institute of Bioinformatics (SIB) y el Protein Information Resource (PIR). Entre las tres instituciones, se encuentran trabajando cerca de 150 personas en distintas tareas como la **curación de las bases de datos**, el desarrollo de software y el soporte técnico. Esto la ha convertido en el líder mundial en el almacenamiento de información sobre proteínas.

Página web: http://www.uniprot.org/

Ventajas y desventajas de Uniprot:

Entre las ventajas podemos destacar que Uniprot presenta una alta calidad de las anotaciones (*high quality annotations*) referidas a la función de las proteínas, que han estandarizado el uso de palabras clave (*keywords*) para describir las funciones, que tiene una herramienta interactiva de búsqueda y que tiene la posibilidad de exportar los resultados en forma de tabla.

Las desventajas son que, sólo está destinada a secuencias de proteínas y que su base de datos es más chica que la del **NCBI**.

Páginas de ayuda:

Documentación general de ayuda:
http://www.uniprot.org/help/
Tabla de campos de búsqueda:
http://www.uniprot.org/help/query-fields

Algunos trucos para las búsquedas en UniProt:

Los campos más comunes de búsqueda son:

organism → Organismo
taxonomy → Todos los organismos incluidos dentro de un grupo taxonómico.
author → Autor del artículo o personas que han subido la secuencia o han participado en su obtención.
keyword → Todos los registros asociados a determinada palabra clave. Todos los registros están etiquetados con palabras claves que pueden ser utilizadas para obtener un particular set de datos. Más información en http://www.uniprot.org/keywords/.

Ejemplo: Buscaremos nuevamente transposones en hongos como lo hicimos en Entrez, pero utilizando los campos adecuados para UniProt:

• (keyword:"transposon" OR keyword:"transposase") AND taxonomy:"Fungi"
– 91 results in UniProtKB

• (keyword:"transposon" OR keyword:"transposase" OR keyword:"transposable element") AND taxonomy:"Fungi"
– 91 results in UniProtKB

• keyword:transpos* AND taxonomy:"Fungi"
– 139 results in UniProtKB

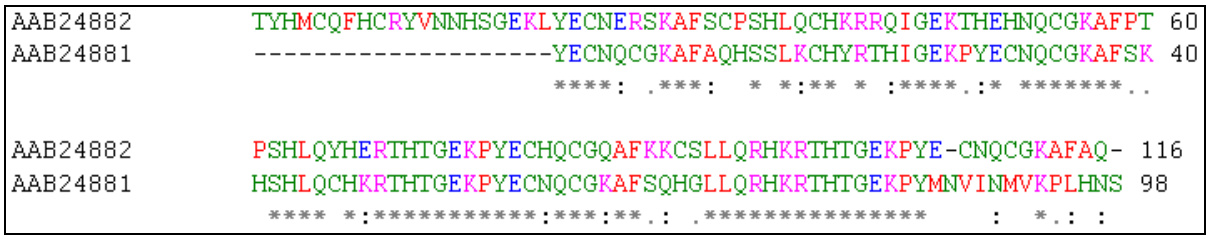
Si comparamos con los resultados obtenidos en la búsqueda Entrez, comprobaremos una de las desventajas que mencionamos anteriormente: La base de datos de UniProt tiene un menor número de registros que la del NCBI.

PRÁCTICA:

4. Repetir la búsqueda realizada para las G protein-coupled receptor usando la base de datos UniProt.
5. Comparar los resultados obtenidos en esta búsqueda con los resultados en la base de datos de proteínas (no de nucleótidos) del NCBI (results in **Protein: sequence database**)

Alineamiento de secuencias BLAST

Un **alineamiento de secuencias** es una forma de representar y comparar dos o más secuencias de ADN, ARN, o estructuras primarias proteicas para resaltar sus zonas de similitud, que podrían indicar relaciones funcionales o evolutivas entre los genes o proteínas consultados. Las secuencias alineadas se escriben en filas de una matriz en las que, si es necesario, se insertan espacios (gaps) para que las zonas con idéntica o similar estructura se alineen.
A continuación se muestra un alineamiento de secuencias, generado por **ClustalW** entre dos proteínas identificadas por el número de acceso en GenBank.



Si dos secuencias en un alineamiento comparten un ancestro común, las no coincidencias pueden interpretarse como **mutaciones puntuales** (sustituciones), y los huecos como **indels** (mutaciones de inserción o deleción) introducidas en uno o ambos linajes en el tiempo que transcurrió desde que divergieron.

Secuencias muy cortas o muy similares pueden alinearse manualmente. Aun así, los problemas más interesantes necesitan alinear secuencias largas, muy variables y extremadamente numerosas que no pueden ser alineadas por humanos. El conocimiento humano se aplica principalmente en la **construcción de algoritmos** que produzcan alineamientos de alta calidad, y ocasionalmente ajustando el resultado final para representar patrones que son difíciles de introducir en algoritmos (especialmente en el caso de secuencias de nucleótidos).

Terminología

Homología vs. Similitud: “Similitud es la observación o medición de parecido y diferencia, independiente del origen de ese parecido. Homología significa, específicamente, que las secuencias y los organismos en los que están presentes, descienden de un ancestro común” .

En sentido estricto, la homología se refiere únicamente a un origen común entre dos caracteres. Por tanto, dos secuencias son homólogas o no homólogas y no hay ninguna gradación intermedia. Una situación similar del mundo real es el embarazo: una mujer no puede estar 50% embarazada: o está o no está, o no se sabe. Similitud, en cambio, es una medida del parecido entre dos secuencias que puede cuantificarse (por ejemplo, mediante el porcentaje de identidad) .

Alineamiento Global vs Alineamiento Local

Las aproximaciones computacionales al alineamiento de secuencias se dividen en dos categorías: **alineamiento global** y **alineamiento local**. Calcular un alineamiento global

es una forma de optimización que "fuerza" al alineamiento a ocupar la longitud total de todas las secuencias introducidas (secuencias problema). Comparativamente, los alineamientos locales identifican regiones similares dentro de largas secuencias que normalmente son muy divergentes entre sí. A menudo se prefieren los alineamientos locales, pero pueden ser más difíciles de calcular porque se añade el desafío de identificar las regiones de mayor similitud.

Los **alineamientos globales**, que intentan alinear cada residuo de cada secuencia, son más útiles cuando las secuencias problema son similares y aproximadamente del mismo tamaño. Una estrategia general de alineamiento global es el algoritmo **Needleman-Wunsch** basado en programación dinámica. Los alineamientos locales son más útiles para secuencias diferenciadas en las que se sospecha que existen regiones muy similares o motivos de secuencias similares dentro de un contexto mayor. El algoritmo **Smith-Waterman** es un método general de alineamiento local basado en programación dinámica. Con secuencias suficientemente similares, no existe diferencia entre alineamientos globales y locales.

Los métodos híbridos, conocidos como **semiglobales** o métodos "glocales" intentan encontrar el mejor alineamiento posible que incluya el inicio y el final de una u otra secuencia. Puede ser especialmente útil cuando la parte "corriente arriba" de una secuencia se solapa con la parte "corriente abajo" de la otra. En este caso, ni el alineamiento global ni el local son completamente adecuados: un alineamiento global intentará forzar a la alineación a extenderse más allá de la región de solapamiento, mientras que el alineamiento local no cubrirá totalmente la región solapada.

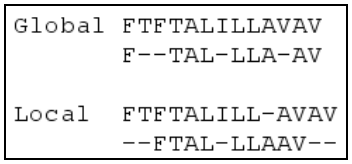


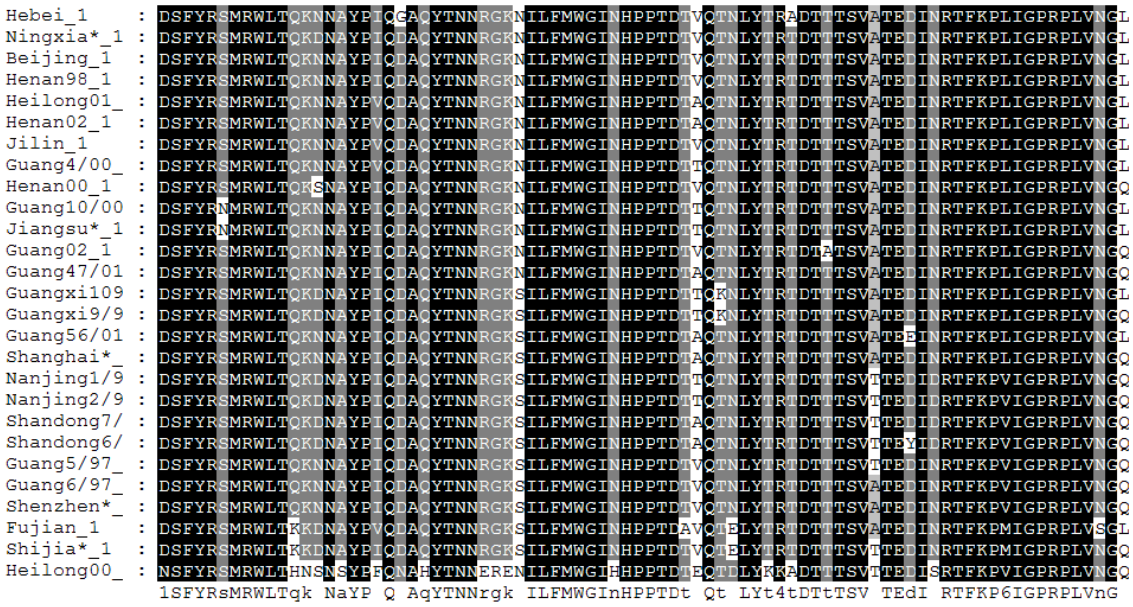
Imagen de un alineamiento local y uno global demostrando la tendencia a poner huecos de los alineamientos globales si las secuencias no son muy similares.

Alineamiento de pares vs Alineamientos Múltiples

Los métodos de **alineamiento de pares**, o emparejamientos, se utilizan para encontrar la mejor coincidencia en bloque (local) o alineamiento global de dos secuencias. Los alineamientos de pares sólo pueden utilizarse con dos secuencias a la vez, pero son fáciles de calcular, y son utilizados a menudo en métodos que no requieren precisión extrema, como **la búsqueda en bases de datos de secuencias con alta similitud con respecto a una petición**. Los tres métodos principales de generar alineamientos de pares son los de matriz de puntos, los de programación dinámica y los de búsqueda de palabra, aunque la mayoría de métodos de alineación múltiple de secuencias pueden funcionar con sólo dos secuencias. Aunque cada método tiene sus propios puntos fuertes y débiles, todos ellos tienen problemas para alinear secuencias repetitivas con bajo contenido en información, especialmente cuando el número de repeticiones puede ser diferente en las dos secuencias que se alinean.

El **alineamiento múltiple de secuencias** es una extensión del alineamiento de pares que incorpora más de dos secuencias al mismo tiempo. Los métodos de alineamiento múltiple intentan alinear todas las secuencias de un conjunto dado. Los alineamientos

múltiples son usados a menudo en la identificación de **regiones conservadas** en un grupo de secuencias que hipotéticamente están relacionadas evolutivamente. Estos motivos conservados pueden ser usados para localizar sitios catalíticos activos en las enzimas. Los alineamientos múltiples son también utilizados para ayudar al establecimiento de **relaciones evolutivas** mediante la construcción de árboles filogenéticos. Los alineamientos múltiples de secuencias son computacionalmente más difíciles de producir, sin embargo la utilidad de estos alineamientos en la bioinformática ha dado lugar al desarrollo de una variedad de métodos adecuados para la alineación de tres o más secuencias.



Alineamiento de 27 secuencias de la proteína hemaglutinina de la gripe aviaria. Coloreado según la conservación de residuos (más oscuro cuanta mayor conservación).

Existen numerosos programas que realizan alineamientos de secuencias (de todos los tipos). Una muy buena recopilación puede obtenerse en http://es.wikipedia.org/wiki/Anexo:Software_para_alineamiento_de_secuencias. De todos estos, podemos destacar los más importantes: **BLAST** que hace alineamientos locales contra una base de datos, y **ClustalW**, **MUSCLE y T-Coffee** que hacen alineamientos globales múltiples. Volveremos a hablar de los programas para alineamientos múltiples en el siguiente Capítulo de este curso, ahora nos centraremos en BLAST.

BLAST (Basic Local Alignment Search Tool)

Como ya mencionamos anteriormente, **BLAST** es un programa informático de alineamiento de secuencias de tipo local, ya sea de ADN o de proteínas. El programa es capaz de comparar una **secuencia problema** (*query sequence*) contra una gran cantidad de secuencias que se encuentren en una **base de datos**. El algoritmo encuentra las secuencias de la base de datos que tienen mayor parecido a la secuencia problema. Es importante mencionar que BLAST usa un algoritmo heurístico por lo que no nos puede

garantizar que ha encontrado la solución correcta. Sin embargo, BLAST es capaz de calcular la significación de sus resultados, por lo que nos provee de un parámetro para juzgar los resultados que se obtienen.

Normalmente el BLAST es usado para encontrar **probables genes homólogos**. Por lo general, cuando una nueva secuencia es obtenida, se usa el BLAST para compararla con otras secuencias que han sido previamente caracterizadas, para así poder **inferir su función**. El BLAST es la herramienta más usada para la anotación y predicción funcional de genes o secuencias proteicas. Muchas variantes han sido creadas para resolver algunos problemas específicos de búsqueda.

BLAST es de dominio público y puede usarse gratuitamente desde el servidor del NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi). También está disponible para ser instalado localmente. Algunas ventajas de usar el servidor del NCBI son que el usuario no tiene que mantener ni actualizar las bases de datos y que la búsqueda se hace en un cluster de computadoras, lo que otorga rapidez. Las desventajas son: no se permiten hacer búsquedas masivas dado que es un recurso compartido, no se puede personalizar las bases de datos contra la que busca el programa, y las secuencias son enviadas al servidor del NCBI sin ningún tipo de cifrado, lo que puede ser un problema para quienes quieran mantener sus secuencias privadas. La aplicación local de BLAST tiene la ventaja de que permite manejar varios parámetros que en las búsquedas de NCBI están estandarizados, por lo que provee una mayor flexibilidad para los usuarios avanzados.

Algoritmo del BLAST

BLAST usa el algoritmo Smith-Waterman para realizar sus alineamientos. BLAST usa una **matriz de sustitución de aminoácidos o nucleótidos** para calificar sus alineamientos. Dicha matriz contiene la puntuación (*score*) que se le da al alinear un nucleótido o un aminoácido X de la secuencia A con otro aminoácido Y de la secuencia B. Las matrices más usadas para calificar alineamientos de proteínas son:

BLOSUM (http://es.wikipedia.org/wiki/BLOSUM) y
PAM (http://es.wikipedia.org/wiki/PAM_%28bioinform%C3%A1tica%29)

Ambas fueron obtenidas midiendo la frecuencia de los aminoácidos en una gran muestra de proteínas. También se permite al usuario definir su propia matriz. El tipo de matriz usada es determinante para los resultados que se obtendrán, el uso de una matriz incorrecta puede llevar a calificar erróneamente los alineamientos y por lo tanto obtener resultados equivocados.

El algoritmo de BLAST tiene tres etapas principales: **ensemillado** (*seeding*), extensión (*extension*) y evaluación (*evaluation*). A continuación se describen brevemente cada una de ellas:

Primera etapa: ensemillado o seeding.
En esta etapa se buscan "palabras" (*word hits*) pequeñas en las secuencias de la base de datos, que corresponden a fragmentos de la secuencia problema. BLAST asume que los alineamientos significativos deben contener estas palabras. Sólo se consideran significativas las palabras que tengan una puntuación mayor a T (T es un parámetro que se puede modificar al usar el programa) y que se encuentren al menos a una distancia A de otra palabra. W es otro parámetro usado por BLAST y se refiere al tamaño de las palabras a buscar. Ajustando los parámetros T, A y W se puede escoger entre hacer un alineamiento sensible pero lento, o uno más rápido pero con menor sensibilidad.

Segunda etapa: extensión.
Una vez obtenidas las palabras que cumplen con los criterios dados, se pasa a la etapa de extensión. En esta etapa el alineamiento se va extendiendo a ambos lados de las palabras. La extensión realizada en este punto se realiza haciendo uso del algoritmo de Smith-Waterman. BLAST va extendiendo el alineamiento hasta que la puntuación del alineamiento descienda X o más puntos con respecto a la puntuación más alta obtenida anteriormente. **Aquí reside el factor heurístico** del BLAST, ya que al imponer el límite X, evita extender a lo largo de toda la secuencia todos los alineamientos (proceso que llevaría demasiado tiempo). El peligro que esto conlleva es que el programa se puede quedar atorado en un máximo local. Es por ello que la definición de X es determinante para el resultado.

Tercera etapa: evaluación
Una vez terminada la extensión de todas las palabras, cada uno de los alineamientos realizados es evaluado para determinar su **significación estadística**. Para ello, el programa elimina los alineamientos inconsistentes (alineamientos que junten la misma parte de la secuencia problema con distintas partes de una secuencia en la base de datos). Los alineamientos resultantes son llamados **pares de alta puntuación** (*High Score Pairs o HSPs*). Una vez realizado esto, se calcula la puntuación final de los alineamientos resultantes y se determina su significación tomando en cuenta la probabilidad que tiene dicho alineamiento de haber sido obtenido por azar de acuerdo al tamaño de la base de datos. Al final se reportan sólo los alineamientos que hayan obtenido una probabilidad mayor a E. El parámetro E es conocido como **e-valor** (*e-value*) de corte, y nos permite definir qué alineamientos queremos obtener de acuerdo a su significación estadística. El *e-value* representa el número de alineamientos esperados por azar, teniendo en cuenta el tamaño de la base de datos, la secuencia problema, los gaps y la matriz de puntuación. **Cuanto menor sea el valor de E, más significativo es un alineamiento**.

Tipos de búsquedas BLAST
Hay muchas formas diferentes de hacer BLAST. Las grandes divisiones son: Nucleótidos, proteínas, traducciones, BLAST genómico y “BLASTs” especiales.

Tipo de Búsqueda	Secuencia Problema	Tipo de base de datos
Blastn, megablast	Nucleotídica	Nucleótidos
Blastp	Proteínica	Proteínas
BlastX	Nucleotídica (será traducida)	Proteínas
TBlastn	Proteínica	Nucleotídica (será traducida)
TBlastX	Nucleotídica (será traducida)	Nucleotídica (será traducida)
Bl2seq	Secuencia A	Secuencia B

El tipo de BLAST a seleccionar depende de varios factores, entre ellos:

a) la naturaleza de nuestra secuencia (¿es ADN o proteína?)
b) la base de datos que queremos sondear (¿queremos buscar en toda la base de datos, o restringirnos a un tipo de molécula especial u organismo particular?)
c) la hipótesis que queremos comprobar (¿estamos buscando secuencias potencialmente homólogas a la nuestra o más bien la posición de nuestra secuencia en un genoma particular?)
d) los supuestos acerca de nuestros resultados (si buscamos secuencias homólogas, ¿esperamos encontrar alta o baja conservación?)

Algunos Parámetros del NCBI BLAST

PROGRAM: Aquí se elige el programa Blast a usar ej.: blastp

DESCRIPTIONS: Restringe el número de descripciones cortas de secuencias que aparean, por default es 100

ALIGMENTS: número de alineamientos mostrados en pantalla, por default es 50 y está restringido a los de mayor score

EXPECT: significancia estadística de los apareamientos contra las secuencias de la base. Por default es 10, esto implica que se requieren 10 apareamientos para ser considerado probable.

CUTOFF: valor de corte del score. Se calcula del EXPEXCT value. Cuanto mayor es el CUTOFF mayor rigurosidad de apareamiento se le esta pidiendo al programa.

MATRIX: Matriz empleada para hacer los apareamientos. Estas matrices están construidas a partir de algoritmos que surgen de valores de comparación entre distintos aminoácidos a lo largo de la evolución

Bases de Datos en BLAST:

- nr – Non redundant proteins
- nt – Non redundant nucleotides
- refseq_rna
- refseq_genomic
- refseq_protein
- est
- swissprot – Curated proteins

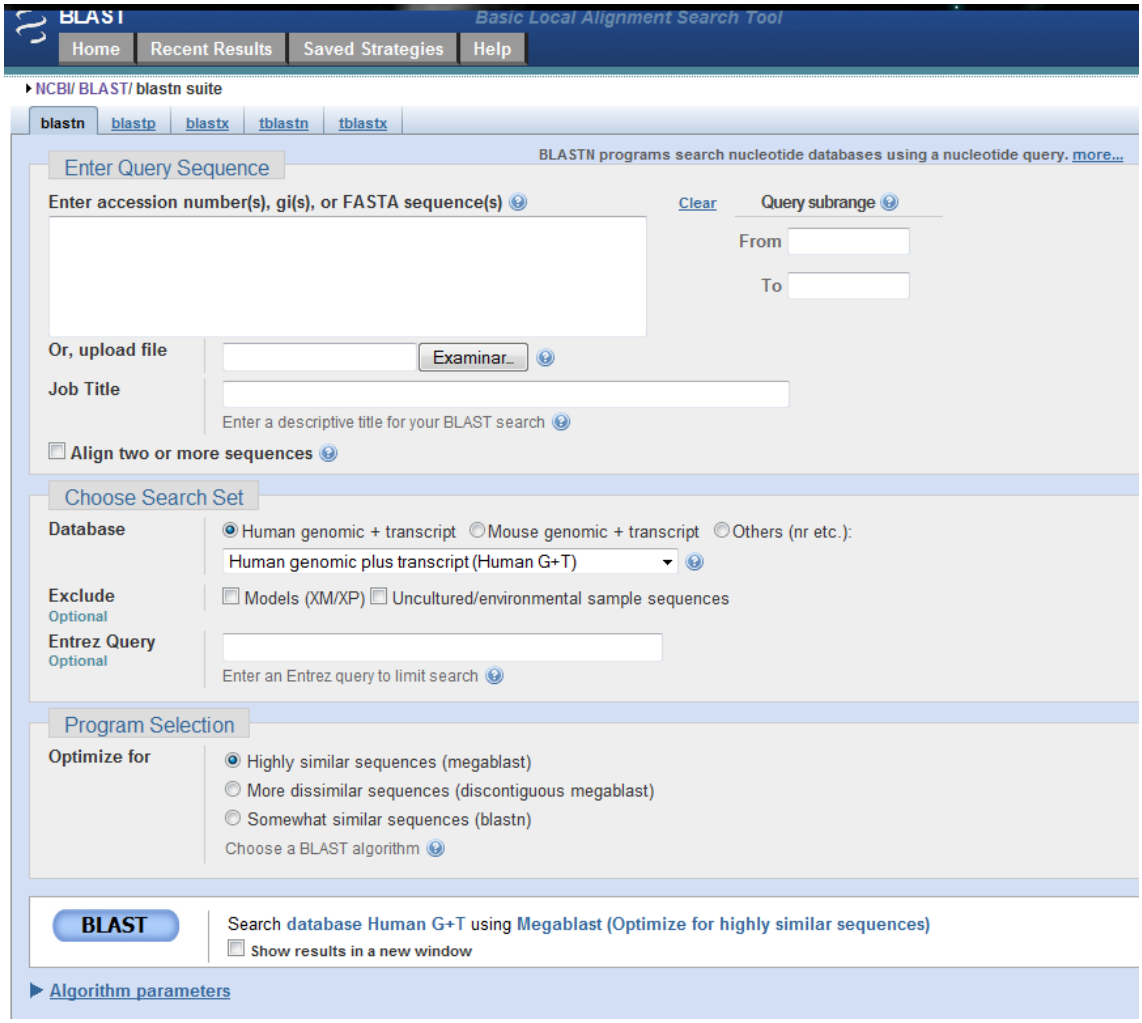
Para más detalles ver ANEXO 2 - Contenido de bases de datos disponibles en BLAST

Consideraciones al usar BLAST

- A pesar de que BLAST es un programa muy poderoso y casi siempre podemos confiar en sus resultados, se debe recordar que el **programa es heurístico** y por lo tanto puede que no encuentre la solución óptima. En la actualidad, el abuso y la pobre interpretación de los resultados de BLAST ha llevado a múltiples errores de anotación. Una cosa a tener en cuenta al usar BLAST es que cuanta más evidencia externa se pueda obtener para corroborar un alineamiento (fisiológica, filogenética, genética, etc.) es mejor.
- El programa de BLAST **NO garantiza que las secuencias que alinea sean homólogas** y mucho menos que tengan la misma función, simplemente provee posibles candidatos. Se necesitan más análisis para anotar correctamente una secuencia.
- La puntuación del BLAST depende del **largo de la secuencia**, una secuencia muy corta tendrá una puntuación menor que una grande simplemente por la cantidad de caracteres que tiene. Así que siempre se debe interpretar la puntuación con respecto al largo de la secuencia.
- **El e-value depende del tamaño de la base de datos**. Para bases de datos muy pequeñas, e-valores altos son más significativos que para bases de datos muy grandes. Para la base de datos no redundante (NR) de NCBI por lo general **e-values de 0.01** o menos son considerados como significativos, pero esto puede depender de la secuencia que se esté analizando.
- Se debe tener cuidado con los **errores de anotación**; es común que alguna secuencia que se anotó mal (ya sea porque se anotó automáticamente o por error humano) sea utilizada como referencia para anotar otras secuencias similares, por lo que los errores de anotación se pueden propagar rápidamente. Siempre debemos especificar que la función de nuestra secuencia **es posible o probable si fue asignada usando identidad con otras secuencias**. Asimismo debemos tener en cuenta que la gran mayoría de las funciones asignadas en la actualidad son putativas y que pueden no ser una buena referencia para una asignación funcional.
- A pesar de lo que comúnmente se piensa, las secuencias con la mejor puntuación o el mejor e-value **NO necesariamente son los mejores candidatos a ser genes homólogos**. Es importante analizar todos los alineamientos que encuentra el programa y sacar conclusiones en base al resultado global.
- BLAST tiene varios parámetros por defecto que en general funcionan bien para la mayoría de los casos, pero habrá situaciones en las que es necesario cambiarlos para obtener mejores resultados. **No hay forma de saber exactamente qué parámetro es el óptimo**, y se tienen que realizar múltiples pruebas hasta encontrar las mejores condiciones.

Realizar una búsqueda con BLAST:

La página principal del BLAST en el NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) nos permite elegir directamente distintos organismos (Human, Mouse, etc.), distintos programas (nucleotide blast, protein blast, blastx, blastn, etc.) y otras búsquedas más especializadas. Por ejemplo podemos realizar un BLAST de ADN (nucleotide blast):



Copiar y pegar la siguiente secuencia en el campo: *Enter accession number, gi or FASTA sequence*.

>cl2ct2cn2
CTCGAGCCTATTCTGGAAGGACGCATTCTGGGAAACGTCCAGCTTCGGGCA
CCTCCCTCATTCCACCAATTCCACCACGCATTGGATTTTGCAGAAAATGACA
GCATGAACTCGCGGGACCGAGCACTGTCTCCGGAATACACCGAATCATCGA
CGTCAACCACGACGACGACAACATAAAATTTCATCCTCTCCCCCTTGAGATT
CCGGCCGTCGCAACCATGAAGCGCCAAGTCGTCCGCCCCCACCGAAATGCG
CTCTTCTCCTCGACCTCCCTCGCAACCACCCTCCTCGTCCTGACGCTCGCCTT
TGCGTCCCTAGTAAATGCCTACACTGAAATGTCCGAGGCCTTTCTCAAGGCG
ATCCCACTCGACCATGGCGATTTCGACATACTAAACGAGCGCGGCCTGCTG
GCGCCCATCCTGATCCCCCGAGTGCCAGGCACGCCGGGACAGGTGCAAGCC
CAACAACACATCACCTCCTTCTTTGCGCGCGAGCTGCCCAAATGGAACGTTT
CGTGGCAGAACTCGACCGGCACGACCCCCCGTGACGGGCAACAAGCAGGTG
CCCGTTCCAAAACATCATTTTGCGCCCGCGAGCCCGCCCTGGACAAAGCCC
GGGCAGGCAAAACTACCTGGAGGCGGGTGGCACACTACGACTCCAAAAAA
TTCGCCAGAGGGCTTCAATCGGAGCCACC

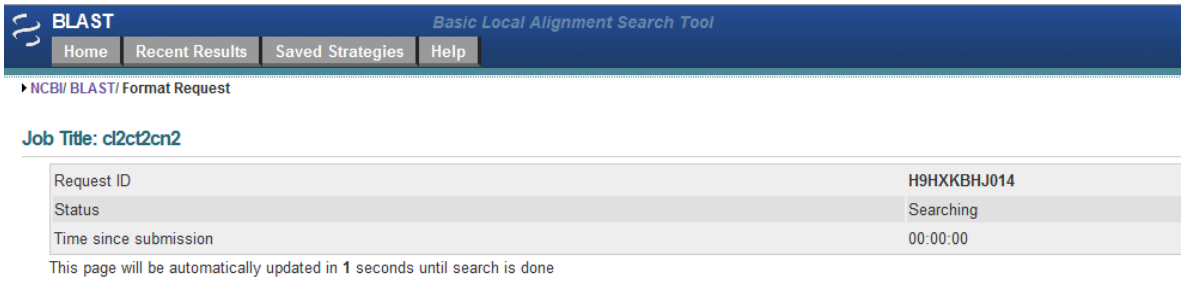
En la página de búsqueda del BLAST podemos modificar numerosos parámetros. El más importante de ellos es la secuencia que queremos utilizar en la búsqueda (Enter Query Sequence). Podemos poner una secuencia en formato fasta (como acabamos de hacer) o un número de acceso de Genbank. Además podemos limitar la búsqueda a una región concreta de la secuencia (Query subrange). El formulario nos permite también escoger un archivo que contenga la secuencia.

La segunda decisión importante es la base de datos con la que vamos a comparar nuestra secuencia (Choose Search Set). Podemos elegir una de las numerosas bases de datos ofrecidas por el NBCI (humano, ratón, nr, refseq, etc.) o podemos escribir una expresión de búsqueda para el entrez. Si elegimos esta última opción la búsqueda se realizará en comparando nuestra secuencia con las secuencias resultantes de esta búsqueda.

Por último podemos seleccionar el programa a utilizar: megablast (para encontrar secuencias muy similares), discontiguous megablast (para secuencias algo diferentes) y blast para secuencias algo más distintas. Cuanto más sensible sea el algoritmo más tiempo tardará la búsqueda .

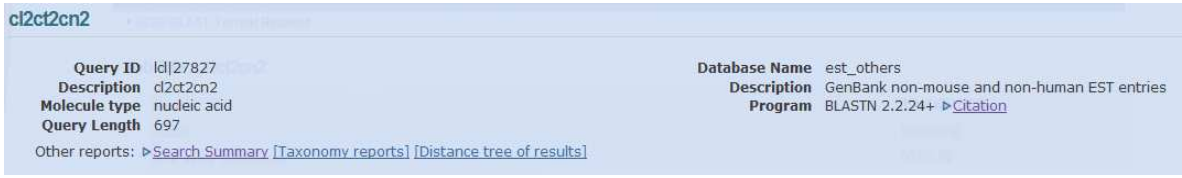
Como ejemplo realizar una búsqueda con megablast sobre la base de *datos Non-human, non-mouse ESTs (est_others).*

Mientras el BLAST se está ejecutando veremos una página en la que se nos informa sobre el tiempo estimado que requerirá la búsqueda.

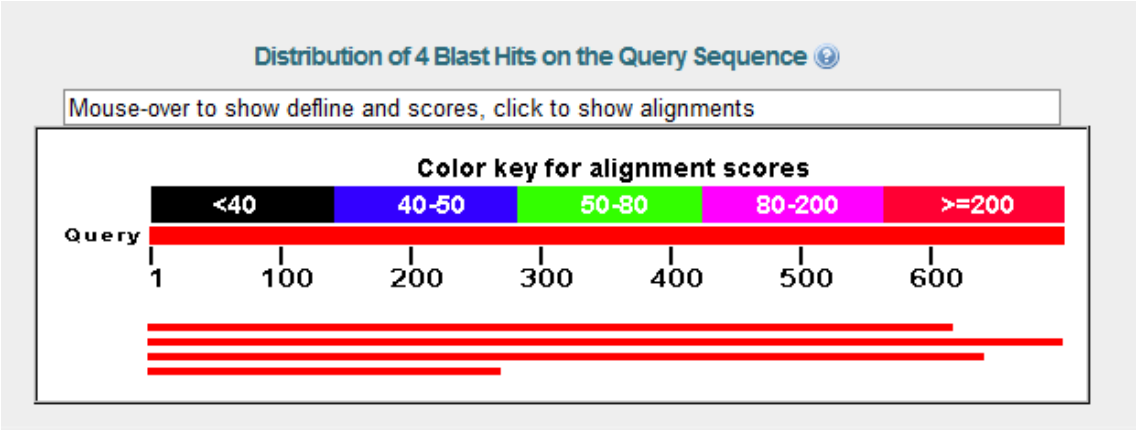


Explicación del reporte de BLAST:

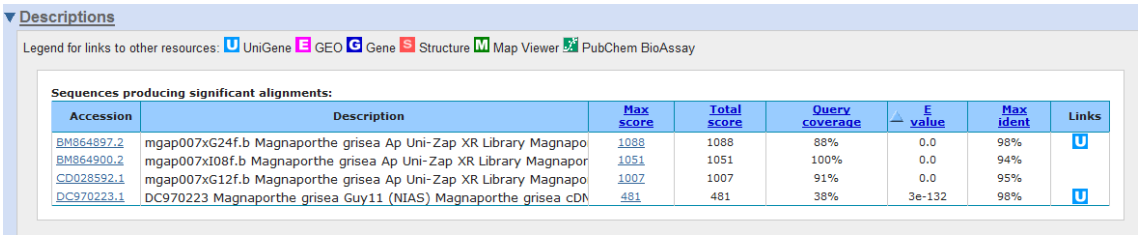
El resultado del BLAST se divide en varias secciones. En la primera se nos informa sobre la versión del programa utilizado, el nombre de la base de datos y el número de secuencias buscadas.



A continuación aparecen una figura y una tabla con un resumen de los resultados. En la figura se observa una fila para cada una de las secuencias similares obtenidas (hit). En cada fila se representa la región cubierta por los distintos HSP (High Scoring Segment Pairs). Estas regiones están pintadas de un color u otro dependiendo del e-value correspondiente.



A continuación se observa una tabla, donde también se encuentran los distintos resultados (hits). Para cada uno se puede observar la definición de la secuencia (description) y el e-value. En la tabla los hits están ordenados por el e-value, de menor a mayor:



Descriptions

Legend for links to other resources: UniGene GEO Gene Structure Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
BM864897.2	mgap007xG24f.b Magnaporthe grisea Ap Uni-Zap XR Library Magnapo	1088	1088	88%	0.0	98%	U
BM864900.2	mgap007xI08f.b Magnaporthe grisea Ap Uni-Zap XR Library Magnapor	1051	1051	100%	0.0	94%	
CD028592.1	mgap007xG12f.b Magnaporthe grisea Ap Uni-Zap XR Library Magnapo	1007	1007	91%	0.0	95%	
DC970223.1	DC970223 Magnaporthe grisea Guy11 (NIAS) Magnaporthe grisea cDN	481	481	38%	3e-132	98%	U

En la siguiente sección de la página de resultados se muestran los alineamientos de la secuencia problema (query) con las distintas secuencias de la base de datos (Sbjct). Para cada una de ellas se muestra la descripción y los alineamientos correspondientes a los distintos HSPs.

>gb|BM864897.2| mgap007xG24f.b Magnaporthe grisea Ap Uni-Zap XR Library Magnaporthe grisea cDNA clone mgap007xG24 5', mRNA sequence.
Length=614

HSP info
score = "1088"
expect = "0.0"
identity = "99%"

```
 Score = 1088 bits (589),  Expect = 0.0
 Identities = 607/616 (99%), Gaps = 4/616 (0%)
 Strand=Plus/Plus

Query  1    CTCGAGCCTATTCTGGAAGGACGCATTCTGGGAAACGTCCAGCTTCGGGCACCTCCCTCA  60
            ||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    CTCGAGCCTATTCTGGAAGGACGCATTCTGGGAAACGTCAAGCTTCGGGCACCTCCCTCA  60

Query  61   TTCCACCAATTCCACCACGCATTGGATTTTGCAGAAAATGACAGCATGAACTCGCGGGAC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TTCCACCAATTCCACCACGCATTGGATTTTGCAGAAAATGACAGCATGAACTCGCGGGAC  120

Query  121  CGAGCACTGTCTCCGGAATACACCGAATCATCGACGTCAACCACGACGACGACAACATAA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  CGAGCACTGTCTCCGGAATACACCGAATCATCGACGTCAACCACGACGACGACAACATAA  180

Query  181  AATTTCATCCTCTCCCCCTTGAGATTCCGGCCGTCGCAACCATGAAGCGCCAAGTCGTCC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  AATTTCATCCTCTCCCCCTTGAGATTCCGGCCGTCGCAACCATGAAGCGCCAAGTCGTCG  240

Query  241  GCCCCCACCGAAATGCGCTCTTCTCCTCGACCTCCCTCGCAACCACCCTCCTCGTCCTGA  300
            |||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  GCCCCCACCGCAATGCGCTCTTCTCCTCGACCTCCCTCGCAACCACCCTCCTCGTCCTGA  300

Query  301  CGCTCGCCTTTGCGTCCCTAGTAAATGCCTACACTGAAATGTCCGAGGCCTTTCTCAAGG  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  CGCTCGCCTTTGCGTCCCTAGTAAATGCCTACACTGAAATGTCCGAGGCCTTTCTCAAGG  360

Query  361  CGATCCCACTCGACCATGGCGATTTCGACATACTAAACGAGCGCGGCCTGCTGGCGCCCA  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  CGATCCCACTCGACCATGGCGATTTCGACATACTAAACGAGCGCGGCCTGCTGGCGCCCA  420

Query  421  TCCTGATCCCCCGAGTGCCAGGCACGCCGGGACAGGTGCAAGCCCAACAACACATCACCT  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  TCCTGATCCCCCGAGTGCCAGGCACGCCGGGACAGGTGCAAGCCCAACAACACATCACCT  480

Query  481  CCTTCTTTGCGCGCGAGCTGCCCAAATGGAACGTTTCGTGGCAGAACTCGACCGGCACGA  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  CCTTCTTTGCGCGCGAGCTGCCCAAATGGAACGTTTCGTGGCAGAACTCGACCGGCACGA  540

Query  541  CCCCCCGTGACGGG-CAACAAGCAGGTGCCCGTTCCAAAACATCATTT-TGCGCCCGCGA  598
            ||| |||||||||| || |||||||||||| || |||||||||||||| ||||||||||||
Sbjct  541  CCCNCCGTGACGGGNCA-CAAGCAGGTGCCNGT-CCAAAACATCATTTGTGCGCCCGCGA  598

Query  599  GCCCGCCCTGGACAAA  614
            ||||||||||||||||
Sbjct  599  GCCCGCCCTGGACAAA  614
```

PRÁCTICA:

6. Encontrar homólogos de la secuencia de ADN que se encuentra a continuación (cl25ct25cn25) usando BLASTX. Comparar los resultados obtenidos usando las bases de datos de swissprot vs nr. ¿Cuál es el e-value de las secuencias más similares en cada base de datos? ¿Se puede determinar una probable función para la secuencia?
7. Utilizar la secuencia de proteína (CPR1) para realizar un blast de proteínas (BLASTP – protein blast). Realizar la búsqueda usando las bases de datos de swissprot y nr. ¿Cuál es el e-value de las secuencias más similares en cada base de datos? ¿Se puede determinar una probable función para la secuencia?
8. En unas células cancerosas se aisló un mRNA que estaba expresado en cantidades anormales y se secuenció. La secuencia obtenida es (mARN_aislado_y_secuenciado) que se encuentra a continuación. ¿Hay alguna razón para creer que el mRNA aislado esté vinculado con el hecho de que la célula sea cancerosa?

>cl25ct25cn25
TTTTTAATGATAACTGTTTATTGTGACTCTATAGGATGCTTCTATGATGACTC
TATAAGATGAGATGAGATACGAAAATGTATCGAGATGTATATGTGTACACG
GGCATGCCAAAAAATTCCCTACATTCCAGCTCGTACGGCCTCAACCTCGGAT
GTGTTTCATTGCCGGACCTTGTTCAACTTGACCCAGAGGTCCGGCTTGGTCA
TGGAGTGAAAGACCTTCATGCGCTTCATGTTGCCATCCAAGCCCCAGACCG
AATTCACGTCCAAGTATTTTTCTCTATCCAGCTCGAGGTTCATAATTCCACA
ACAGAGTGGCAGTGATGAGTTGCATCTCCATGTAGGCCAGGTTCCTGCCAA
TGCAGACGCGAGGCCCGGTCCCAAAGGGCAGCGAGGCGTTGAAGTCGTCCC
TGGCATACTTGTCATCCTCGCCCAGCCACCTCTCTGGCCTAAAGGACTCGGG
GTCGGAAAAGTGCATCTCCGAGTGCGTCGACACCCACGGCGGAACGGCGAC
GGTCGTCCCCTCGGGTACAAAATACCCGTCGATGGTGGCTCCGCCCCGGGG
GACGACGCGCTGCTGGCTTGCCGGTGAGGGCGGGTAGAGCCTGAGGGCTTC
GTGTATGGTCGCTTCGAGGTAGCTCATCTTGCGGACGTTTTCCCACCGGATG
TGGTCCTCTGTCGGCAGCCCACCCCTGACCTCGTCAATCAGCCTTTGGTACA
CCTCTGGGTGCGTGCAGACGAAATAGGTCCAGCCGGTCAGGGCCGTCGCCG
TCGTGTCGGTGCCGGCGCTGATCATGAGGGCCACGTTGAGTATGACCTCGTT
GTGGC
>CPR1
MYNTMTRAQNTFGFFTSVAFFVAAIIALSDLVAPRAPSVGSLKTTNVQVVKGRP
HYYSSKKEEYAIIKFSLDADLSDLFTWNTKQVFVYVTAEWPDSSKAAAGTNAT
NQAVIWDQIITSPSADHLANIGPAAMRKLKKSAEGKSIDPNRGKLRIKNQRPKY
QITHPSSKIAETDKVVLKLHYNVQPWVGILTWNQNRDIGGWKALKGGVSKAF
KLPALKVKENKDKKKP
>mARN_aislado_y_secuenciado
CAGGAAACATTTTCAGACCTATGGAAACTACTTCCTGAAAACAACGTTCTGT
CCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCCGGACGATAT
TGAACAATGGTTCACTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCC
AGAGGCTGCTCCCCGCGTGGCCCCTGCACCAGCAGCTCCTACACCGGCGGC
CCCTGCACCAGCCCCCTCCTGG

Alineamiento Múltiple de Secuencias

Ya nos hemos introducido brevemente en el Alineamiento Múltiples de Secuencias (*MSA – Multiple sequences alignment*) en la sección anterior (página 24). Vimos que dos de las aplicaciones principales de este tipo de alineamiento son: Identificación de regiones conservadas (dominios/residuos) y el Establecimiento de relaciones evolutivas (Análisis filogenéticos).

Normalmente es complicado alinear a mano tres o más secuencias de longitud biológicamente relevante, y casi siempre consume mucho tiempo. Por esto se utilizan algoritmos computacionales para producir y analizar los alineamientos. Los MSA requieren metodologías más sofisticadas que los alineamientos de pares porque son computacionalmente más complejos de producir. La mayor parte de los programas de alineamiento múltiple de secuencias usan métodos heurísticos en lugar de optimización global, porque identificar el alineamiento óptimo entre más de unas pocas secuencias de longitud moderada es prohibitivamente costoso computacionalmente.

El principal problema que nos plantean los alineamientos múltiples es la posición de los gaps en el alineamiento. En el alineamiento, todas las secuencias tienen que tener la misma longitud y eso se consigue introduciendo gaps.

```
IMPRESIONABLE     IMPRES-IONABLE        IMPRES--O-----
IMPRES-O----      INCUESTIONABLE        INCUESTIONABLE

IMPRESIONABLE-          IMPRES-IONABLE
IMPRES-O------          IMPRES--O-----
INCUESTIONABLE          INCUESTIONABLE
```

La posición de los gaps es diferente en cada comparación de 2 secuencias y es necesario encontrar una solución óptima del alineamiento.

Alineamientos de construcción progresiva (Clustal – T-Coffee)

Los programas más utilizados se basan en esta estrategia. No nos garantizan que el alineamiento obtenido sea el mejor posible; pero son capaces de encontrar una solución óptima de forma muy eficaz. Ejemplos de programas basados en este método heurístico:
Clustal (http://www.ebi.ac.uk/Tools/msa/clustalw2/)
T-coffee (http://www.tcoffee.org/)
El método consiste en primero realizar alineamientos de dos en dos con lo que construye un árbol guía (usando el método de Neighbor-joining), basado en la distancia genética, para detectar la pareja de secuencias más similar. Al alineamiento del par de secuencias más similar va añadiendo el resto de secuencias o alineamientos en el orden determinado por el árbol.

En nuestro ejemplo anterior:
Distancias:

IMPRESIONANTE x IMPRESO: 7/13
IMPRESIONANTE x INCUESTIONABLE: 10/14
INCUESTIONABLE X IMPRESO 4/14

El primer alineamiento seria: INCUESTIONABLE x IMPRESIONABLE
IMPRES–IONABLE
INCUESTIONABLE

Posteriormente uniría IMPRESO
IMPRES–IONABLE
INCUESTIONABLE
IMPRES—O– – – – –

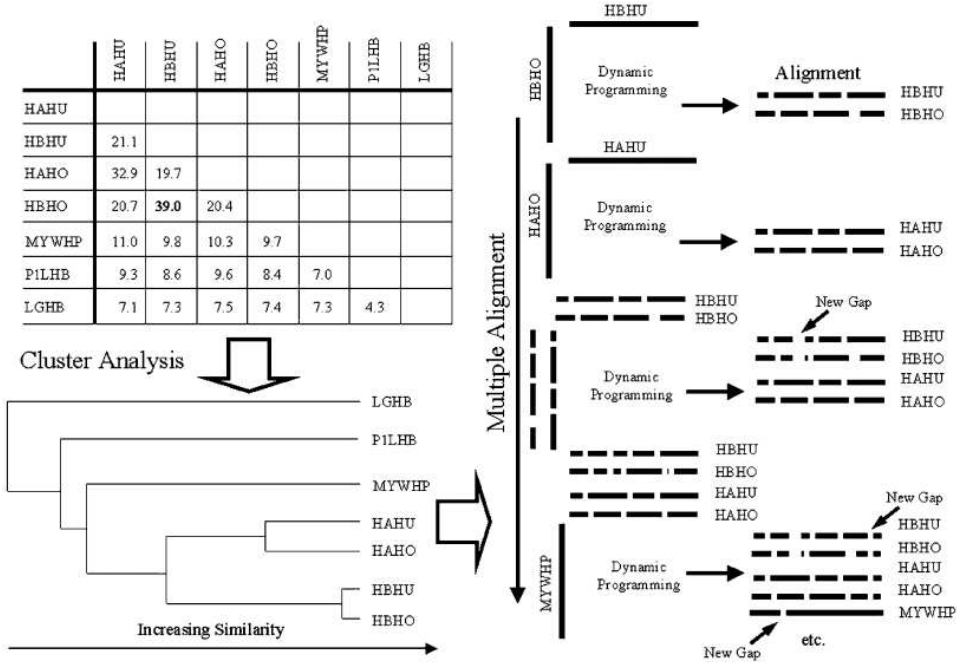


Figure 9.2. Illustration of the stages in hierarchical multiple alignment of seven sequences. The codes for these sequences are HAHU, HBHU, HAHO, HBHO, MYWHP, P1LHB, and LGHB. The table at the top left shows the pairwise *Z*-scores for comparison of each sequence pair. Higher numbers mean greater similarity (see text). Hierarchical cluster analysis of the *Z*-score table generates the dendrogram or tree shown at the bottom left. Items joined toward the right of the tree are more similar than those linked toward the left. Based on the tree, LGHB is least similar to the other sequences in the set, whereas HBHU and HBHO are the most similar pair (most similar to each other). The first four steps in building the multiple alignment are shown on the right. The first two steps are pairwise alignments. The third step is a comparison of profiles from the two alignments generated in *steps 1* and *2*. The fourth step adds a single sequence (MYWHP) to the alignment generated at *step 3*. Further sequences are added in a similar manner.

Baxevanis, Andreas D., y B. F. Francis Ouellette. 2001. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Second Edition*. 2º ed. Wiley-Interscience.

Los programas derivados de Clustal incorporan un sistema de puntuación del alineamiento. La puntuación está basada en la distancia genética entre cada secuencia y la raíz del árbol, teniendo en cuenta la puntuación por los cambios de aminoácidos ó nucleótidos. Además de esta puntuación, los gaps y la extensión de los gaps están penalizados. La puntuación del alineamiento global será la suma de la puntuación de los alineamientos de parejas de secuencias. Clustal tiende a situar los gaps entre las zonas altamente conservadas en vez de separar estas regiones.

Problemas de los alineamientos de construcción progresiva

Los mayores problemas de este tipo de alineamiento derivan del primer alineamiento. Si las dos primeras secuencias son cercanas (muy similares) el alineamiento base probablemente contenga pocos errores. En cambio, si estas dos secuencias son muy divergentes el alineamiento obtenido no será muy adecuado y los gaps y errores se propagarán al resto de secuencias añadidas. Puesto que este primer alineamiento ya no es modificado al unirse el resto de alineamientos. Para paliar este problema es mejor alinear secuencias del mismo tamaño es decir incluir únicamente en el alineamiento aquellas regiones presentes en todas las secuencias; ya que el programa intentara alinear toda la longitud de la secuencia introduciendo gaps en el resto.
Este tipo de alineamientos funciona correctamente para secuencias con cierto grado de conservación y que varíen de forma más o menos continua. Pero a pesar de estos inconvenientes, este tipo de algoritmos encuentran una solución óptima con pocos recursos; permitiendo el análisis de muchas secuencias.

T-Coffee: Alineamiento de construcción progresiva y comparación entre alineamientos

El otro método común de alineamiento progresivo denominado T-Coffee es más lento que Clustal y sus derivados, pero generalmente produce alineamientos más precisos para conjuntos de secuencias lejanamente emparentadas. T-Coffee calcula alineamientos de pares combinando el alineamiento directo del par con alineamientos indirectos que alinean cada secuencia del par con una tercera. Realiza alineamientos de dos en dos usando Clustal y Lalign, que encuentra regiones múltiples de alineamiento local entre dos secuencias. Los alineamientos y el árbol filogenético resultantes se usan como guía para producir nuevos y más precisos factores de ponderación.

El funcionamiento de T-Coffee es el siguiente:

Para puntuar los alineamientos no utiliza la matriz de sustitución, si no un "extended weight". Esta puntuación no se basa únicamente en la comparación de las dos secuencias, tiene en cuenta el resto de secuencias y alineamientos de la librería para obtenerla.

1- Alineamiento de dos en dos utilizando ClustalW (alineación global) y Lalign (alineación local). Utiliza la librería en vez de la matriz de sustitución para puntuarlos
2 -Cálculo de la matriz de distancias
3- Creación del árbol mediante neighbor-joining
4- Construir el alineamiento múltiple siguiendo el árbol. Los gaps no son penalizados porque ya han sido tenidos en cuenta en el "extended weight".

T-Coffee usualmente obtiene mejores resultados que Clustal y facilita el alineamiento de secuencias alejadas; sobre todo cuando se dispone de información adicional

(dominios, estructura secundaria, etc). Debido a las mejoras en la precisión, T-coffee no puede compararse con Clustal en la velocidad.

Puesto que los métodos progresivos son heurísticos y, por lo tanto, no garantizan la convergencia a un óptimo global, la calidad del alineamiento puede ser difícil de evaluar, y su verdadera significación biológica puede ser oscura. Un muy reciente método semi-progresivo que mejora la calidad del alineamiento y que no utiliza una heurística "con pérdidas" se ha implementado en el programa PSAlign (http://faculty.cs.tamu.edu/shsze/psalign/).

Métodos iterativos (Muscle)

Un conjunto de métodos para producir alineamientos múltiples de secuencias que reducen los errores inherentes en los métodos progresivos son los clasificados como “iterativos”, ya que trabajan de forma similar a los métodos progresivos, pero realinean repetidamente las secuencias iniciales además de añadir nuevas secuencias al MSA en crecimiento. Una razón por la que los métodos progresivos son tan fuertemente dependientes de la alta calidad del alineamiento inicial es el hecho de que estos alineamientos se incorporan siempre al resultado final; esto es, una vez que una secuencia ha sido alineada dentro del MSA, su alineamiento no vuelve a ser considerado. Este enfoque mejora la eficiencia a costa de la precisión. En contraste, los métodos iterativos pueden volver a alineamientos de pares previamente calculados (o sub-MSAs) incorporando subconjuntos de la secuencia problema como un medio de optimización de una función objetivo general, tal como encontrar una puntuación de alineamiento de alta calidad.

Se ha implementado una variedad de métodos de iteración sutilmente diferentes, que han sido dispuestos en diferentes paquetes de software. Existen revisiones y comparaciones útiles, pero evitan, generalmente, elegir la "mejor" técnica.

Uno de los métodos más populares basado en la iteración, llamado MUSCLE (http://www.drive5.com/muscle/) (de *multiple sequence alignment by log-expectation*, o alineamiento múltiple de secuencias por log-esperanza; este último término corresponde a una función de puntuación no común basada en la esperanza matemática, y resultado de modificar la función log-average o log-promedio), mejora en relación a los métodos progresivos con una medida más precisa de la distancia para valorar el parentesco de dos secuencias. La medición de la distancia se actualiza entre las etapas de la iteración (sin embargo, en su forma original, MUSCLE contenía sólo dos o tres iteraciones, dependiendo de si se activaba o no el refinamiento).

Funcionamiento de Muscle:

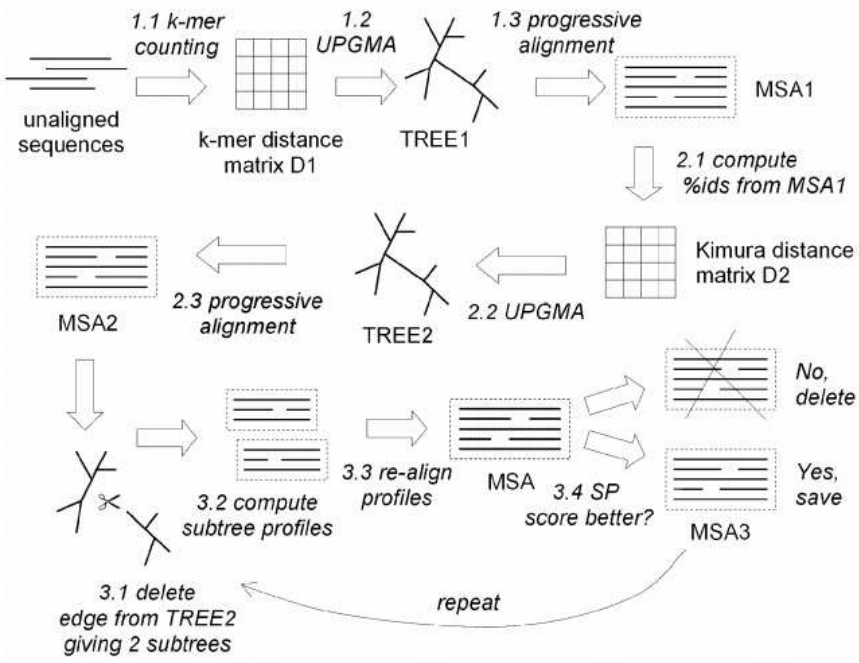


Figure 2. This diagram summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple alignment is available at the completion of each stage, at which point the algorithm may terminate.

(Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput.Nucleic Acids Res. 32(5):1792-1797.)

Como acabamos de ver, existen una serie de algoritmos diferentes para la construcción de las alineaciones de secuencias múltiples. La tabla a continuación describe brevemente tres de los algoritmos de alineación más utilizados sobre la base de algunas de las consideraciones antes mencionadas.

Programa	Ventajas	Precauciones
ClustalW	Utiliza menos memoria que otros programas.	Menos precisión y escalabilidad.
MUSCLE	Más rápido y preciso que ClustalW. Buena relación Precisión/Costo computacional. Opciones para correrlo aún más rápido con una precisión promedio mas baja en aplicaciones a gran escala.	Para conjuntos de datos muy grandes (más de 1000 secuencias) se deben modificar las opciones de tiempo y memoria a ser utilizados.
T-Coffee	Alta precisión y con posibilidad de incorporar fuentes de información heterogéneas.	El uso de memoria y el tiempo computacional son factores limitantes para grandes alineamientos (más de 100 secuencias)

Adaptado de: Edgar, R. C. and Batzoglou, S. (2006). Multiple sequence alignment. Curr Opin Struct Biol, 16(3):368--373.

Formatos comunes para Alineamientos Múltiples de Secuencias:

Los formatos más comunes son: msf (.msf), fasta (.fa .fasta), clustal (.aln) y phylip (.phy). la mayoría de los programas de MSA utilizan el formato FASTA como archivos de entrada. También podemos utilizar la aplicación **"seqret" de Emboss** como vimos anteriormente para formatear los archivos al formato deseado.

El paquete Emboss contiene además numerosos programas relacionados con el análisis de alineamiento múltiple de secuencias:
http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/alignment_multiple_group.html

Program name	Description
emma	Multiple alignment program - interface to ClustalW program
infoalign	Information on a multiple sequence alignment
plotcon	Plot quality of conservation of a sequence alignment
prettyplot	Displays aligned sequences, with colouring and boxing
showalign	Displays a multiple sequence alignment
tranalign	Align nucleic coding regions given the aligned proteins
mse	Multiple Sequence Editor

Visualización y Edición de MSA:

SeaView (http://pbil.univ-lyon1.fr/software/seaview.html)

- Lee y escribe varios formatos de secuencias de ADN y proteínas (NEXUS, MSF, CLUSTAL, FASTA, PHYLIP, MASE, Newick) y de árboles filogenéticos.
- Corre MUSCLE o ClustalW para la alineación de secuencias múltiples, y también permite utilizar cualquier algoritmo de alineamiento externo capaz de leer y escribir archivos en formato FASTA.
- Realiza Arboles Filogenéticos por:
    - Parsimonia, utilizando el algoritmo de PHYLIP (dnapars/protpars)
    - Distancia, con el algoritmo NJ (neighbor-joining) o BIONJ
    - Máxima Verosimilitud (maximum likelihood), mediante el programa PhyML.
- Dibuja árboles filogenéticos en la pantalla, exporta archivos en PDF o PostScript
- Permite descargar secuencias de EMBL / GenBank / UniProt usando Internet.

Jalview (http://www.jalview.org/)

- Visualización
  - Lee y escribe alineamientos en varios formatos (Fasta, PFAM, MSF, Clustal, BLC, PIR)
  - Guarda alineamientos y árboles asociados en formato XML.
- Edición:
  - Insertar o remover gaps usando el mouse o el teclado.
  - Edición de un grupo de secuencias (eliminación o inserción de gaps en grupos de secuencias).
- Análisis:
  - Alinear secuencias utilizando Clustal o MUSCLE.
  - Análisis de conservación de aminoácidos.
  - Opciones de ordenación de alineamientos (por su nombre, por el orden del árbol, por porcentaje de identidad)
  - Cálculo de árboles UPGMA y NJ.
  - Agrupamiento de secuencias utilizando el análisis de componentes principales.
  - Eliminación de secuencias redundantes.
  - Alineamiento de a pares con Smith Waterman de las secuencias seleccionadas.
- Anotación:
  - Uso de programas de predicción de estructura secundaria basados en la web (JNET).
  - El usuario puede elegir las combinaciones de colores para visualizar los alineamientos.
- Publicación
  - Imprimir el alineamiento con colores y anotaciones.
  - Crear páginas HTML.
  - Diversas opciones para exportar los alineamientos.

BAliBASE (http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/)

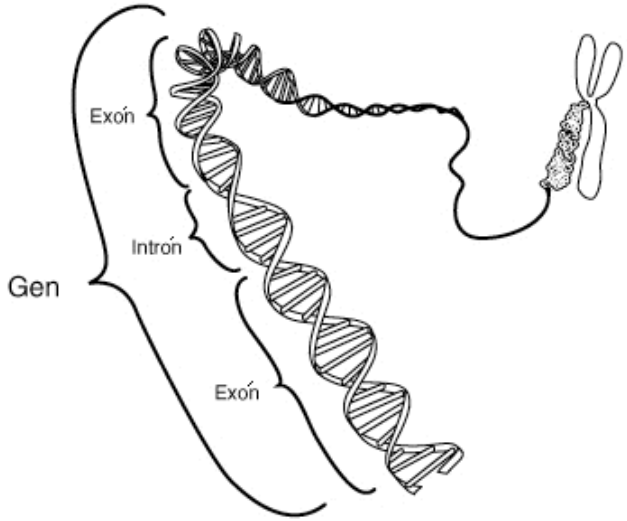
Se trata de una base de datos de referencia de Alineamientos Múltiples. Sirve para evaluar los resultados de los alineamientos generados por distintos programas, ya que los alineamientos que en ella se encuentran han sido editados manualmente. Veremos su utilidad a continuación en la práctica.

PRÁCTICA:

10. En este ejercicio vamos a utilizar un alineamiento múltiple de secuencias de la base BALiBASE para evaluar el desempeño de varios programas de MSA. Para esto dirigirse a la página web http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/ y seleccionar un MSA de la base de datos (Ejemplo: *Look at a list of all the alignments* → Reference 1 → 1aboA)
11. Al final del alineamiento tenemos la opción de guardarlo en formato MSF. Descargaremos el archivo en nuestra computadora.
12. A continuación utilizaremos el programa **Emboss – degapseq** (http://inn-temp.weizmann.ac.il/cgi-bin/emboss/degapseq) para remover los gaps de la secuencia descargada y de esta manera tener las secuencias listas para correrlas con otro programa de alineamientos. En la opción "Select an input sequence" seleccionamos la forma 2 (Upload a sequence from your local computer) y subimos la secuencia descargada en el paso anterior. En Output sequence format seleccionamos FASTA y corremos el programa.
13. Realizaremos un alinemiento con ClustalW a partir de la secuencia FASTA obtenieda en el paso anterior. Para esto utilizaremos el programa **Emboss – emma** http://emboss.bioinformatics.nl/cgi-bin/emboss/emma que utiliza el algoritmo de Clustal para realizar alineamientos. Copiamos la secuencia FASTA obtenieda en el paso anterior y la pegamos en el campo 3. Corremos emma con todas las opciones siguientes sin modificar. Se puede ver el alineamiento generado al final de la página.
14. Ahora realizaremos la evaluación del alineamiento generado con emma en el paso anterior. Para esto accedemos a la página de T-Coffee (http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi) y vamos a la sección EVALUATION (CORE) y hacemos click en REGULAR. Copiamos el alinemiento generado con emma y lo pegamos en el campo habilitado para esto. Mandamos el trabajo (Submit). Cuando el análisis haya terminado hacemos click en el link score_html para ver la evaluación. Tener en cuenta la evaluación total del alineamiento (SCORE=).
15. Ahora repetiremos los 2 pasos anteriores, pero reemplazaremos el alineamiento generado por emma/ClustalW por los generados con MUSCLE y con T-Coffee. Para esto copiamos nuevamente la secuencia obtenida con **Emboss – degapseq** y la utilizamos para generar alineamientos con MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/) y con T-Coffee (http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi). En ambos casos debemos seleccionar como FORMATO de salida *ClustalW* o *ClustalW strict*.
16. Una vez finalizados los análisis deberíamos tener 3 CORE SCORES. ¿Qué programa produjo el valor más alto?
17. Puedes probar con más alineamientos de la BALiBASE.

Predicción de genes

Los mecanismos o procesos de predicción de genes (*gene prediction* o *gene finding*) son aquellos que, dentro del área de la biología computacional, se utilizan para la identificación algorítmica de trozos de secuencia, usualmente ADN genómico, y que son biológicamente funcionales. Esto, especialmente, incluye los genes codificantes de proteínas, pero también podría incluir otros elementos funcionales tales como genes ARN y secuencias reguladoras. La identificación de genes es uno de los primeros y más importantes pasos para entender el genoma de una especie una vez ha sido secuenciado.



Representación de un gen en una cadena de ADN. En general, la predicción de genes trata de localizar en las largas secuencias de ADN, y de forma automatizada, las subsecuencias de nucleótidos que conforman los diferentes genes.

Existen varios métodos para identificar genes: Métodos basados en el contenido (Content-based Methods), Métodos Comparativos y Métodos Ab initio (desde el principio). Los más populares en la actualidad son los últimos dos.

Métodos Comparativos

Son sistemas de predicción de genes basados en comparaciones. En el genoma objetivo se buscan secuencias que sean similares a la evidencia externa, que toma la forma de una secuencia conocida de un ARN mensajero (ARNm) o producto proteico. Dada una secuencia de ARNm, es trivial derivar una única secuencia genómica de ADN desde la cual haya tenido que ser transcrita. Dada una secuencia de proteína, se puede derivar por traducción reversa del código genético una familia de posibles secuencias de ADN codificante. Una vez que las secuencias de ADN candidatas han sido determinadas, es un problema algorítmico relativamente sencillo el buscar eficientemente un genoma objetivo para las coincidencias, totales o parciales, exactas o inexactas. BLAST es un sistema ampliamente utilizado para este propósito.

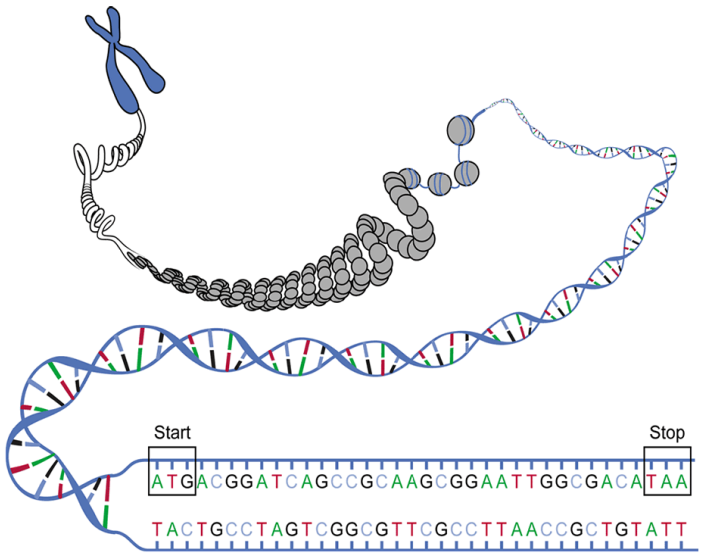
Un alto grado de similitud con un ARN mensajero conocido, o con un producto proteico, es una fuerte evidencia de que una región del genoma en cuestión es un gen codificante de proteína. Sin embargo, aplicar esta aproximación sistemáticamente requiere una exhaustiva secuenciación de ARNm y productos proteicos. No sólo esto resulta caro, sino que en organismos complejos sólo un subconjunto de todos los genes

del genoma del organismo se expresan en un determinado momento, lo que significa que la evidencia extrínseca para muchos genes no está accesible fácilmente en cualquier cultivo de una única célula. Así, para recoger esta evidencia para la mayoría o para todos los genes en un organismo complejo, deben ser estudiadas varios centenares o miles de tipos de células diferentes, lo que representa en sí dificultades añadidas. Algunos genes humanos, por ejemplo, podrían sólo expresarse durante su desarrollo como embrión o feto, lo que dificultaría su estudio por razones éticas.

A pesar de estas dificultades, se han generado unas exhaustivas bases de datos de transcripciones y secuencias de proteínas tanto para el ser humano como para otros organismos modelo importantes en biología, como los ratones o la levadura. Por ejemplo la base de datos **RefSeq** contiene transcripciones y secuencias proteicas de muchas especies diferentes, y el sistema **Ensembl** proyecta intensivamente esta evidencia al ser humano y a bastantes otros genomas. Sin embargo, es probable que ambas bases de datos estén incompletas, y que contengan pequeñas, pero significativas, cantidades de datos erróneos.

Métodos ab initio

Dado el gasto y la dificultad inherentes a la obtención de evidencias extrínsecas para muchos genes, es también necesario recurrir a la predicción de genes ab initio, en la cual se busca, sistemáticamente y de forma exclusiva en la secuencia genómica de ADN, ciertos signos reveladores de genes codificantes de proteínas. Estos signos pueden ser categorizados, en líneas generales, bien como señales (secuencias específicas que indican la presencia cercana de un gen), bien como contenido (propiedades estadísticas de la propia secuencia codificante). El término predicción de la expresión "predicción de genes ab initio" queda precisamente caracterizado como tal puesto que la evidencia externa es generalmente necesaria para establecer de forma concluyente que un supuesto gen es funcional.



En los genomas de los organismos procariotas, los genes tienen secuencias promotoras (señales) específicas y relativamente bien conocidas, como la **caja Pribnow** (Pribnow box) y los sitios de unión de los factores de transcripción, que son fácilmente identificables de forma sistemática. Además, la secuencia codificante para una

proteína se presenta como un marco abierto de lectura (open reading frame, ORF) contiguo, que típicamente mide varios centenares o miles de pares de bases. Las estadísticas de los codones de parada son tales que encontrar un marco abierto de lectura de esa longitud es prácticamente un signo informativo: puesto que 3 de los 64 posibles codones en el código genético son codones stop, podría esperarse un codón stop, aproximadamente, por cada 20-25 codones, o 60-75 pares de bases, en una secuencia aleatoria. Además, el ADN codificante tiene ciertas periodicidades y otras propiedades estadísticas que son fáciles de detectar en una secuencia de esta longitud. Estas características convierten la predicción de genes en procariotas en algo relativamente sencillo, y los sistemas bien diseñados son capaces de alcanzar altos niveles de precisión.

La predicción de genes en organismos eucariotas, especialmente en organismos tan complejos como el ser humano, es considerablemente más desafiante por varias razones. Primero, el promotor y otras señales regulatorias en estos genomas son más complicadas y menos comprendidas que en los procariotas, haciéndolas más complicadas de reconocer fidedignamente. Dos ejemplos clásicos de señales identificadas por los descubridores de genes eucariotas son las **islas CpG** y los sitios de unión para una **cola poli-A**.

Segundo, los mecanismos de splicing (empalme o ayuste) empleado por las células eucarióticas suponen que una determinada secuencia codificante (a proteínas) en el genoma es dividida en diversas partes (exones), separadas por secuencias no codificantes (intrones). (Los sitios de empalme son, en sí mismos, otra señal para cuya identificación están diseñados a menudo los descubridores de genes eucariotas.) Un gen codificante en los humanos puede dividirse en una docena de exones, cada uno de ellos menor de doscientos pares de bases de longitud, y algunos tan cortos como veinte o treinta pares. Es, por lo tanto, mucho más difícil detectar periodicidades u otras propiedades conocidas del ADN codificante en los eucariotas.

Los predictores de genes avanzados para genomas tanto procariotas como eucariotas, usan típicamente complejos modelos probabilísticos, como los **modelos ocultos de Márkov** (Hidden Markov Model - HMM), para combinar información conseguida de una variedad de diferentes medidas de señal y contenido. El sistema **GLIMMER** es un identificador de genes ampliamente usado y muy preciso para organismos procariotas. **GeneMark** es otra aproximación popular. Los predictores de genes “ab initio”, en comparación, han conseguido sólo éxitos limitados. Ejemplos notables de estos son los programas **GENSCAN** y **Geneid**. Unos pocos programas, como **CONTRAST** usan aproximaciones de aprendizaje automático, como máquinas de soporte vectorial, para una eficaz predicción de genes.

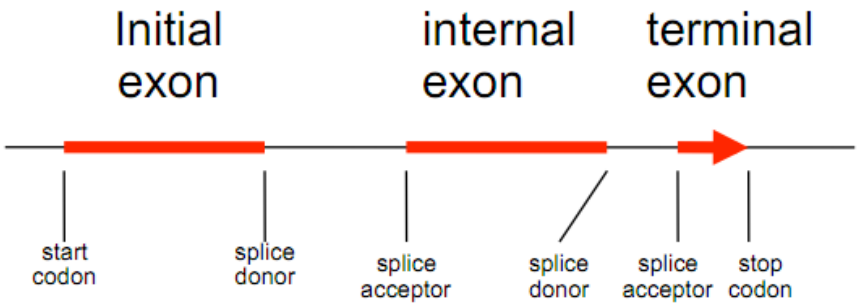
Información utilizada para encontrar genes:

1. Búsqueda de señales. La maquinaria celular reconoce secuencias más o menos conservadas en el DNA genómico.
2. Estadísticos codificantes. Las regiones codificantes tienen propiedades estadísticamente diferentes a las regiones no codificantes.
3. Uso de homología. La similitud con secuencias conocidas es un indicativo de que esa región pueda contener un gen homólogo.

1) Búsqueda de señales:

Estructura del Gen:

Tipos de Exones:



Las señales conocidas son alineadas y se generan patrones con las regiones conservadas.

2) Estadísticos codificantes:

El DNA codificante tiene una composición de nucleótidos diferente al resto de DNA genómico, debido a que ha de codificar para proteínas (es menos aleatorio). Un Estadístico codificante es una función que dada una secuencia de DNA, nos devuelve un número relacionado con la probabilidad de que esa secuencia corresponda a una región codificante. Por ejemplo el "codon bias usage".

3) Uso de Homología:

Algunos programas de predicción de genes permiten el uso de homologías con secuencias conocidas para mejorar las predicciones.
Estas homologías las podemos encontrar en:

- Proteínas de otras especies.

- Fragmentos genómicos que sabemos que se transcriben (ESTs o cDNAs)
- Comparación de genomas completos.

Codon usage bias ("*Sesgo en el uso de codones*")

Codon usage bias se refiere a las diferencias que pueden existir en la frecuencia de ocurrencia de los codones sinónimos en la codificación del ADN.
Hay 64 codones diferentes (61 codones que codifican para aminoácidos más 3 codones de stop), pero sólo 20 aminoácidos diferentes. El exceso en el número de codones permite muchos aminoácidos se codifiquen por más de un codón. Debido a la redundancia se dice que el código genético es **degenerado**. Los diferentes organismos a menudo muestran preferencias particulares de uno de los varios codones que codifican el mismo aminoácido.

Existen numerosos métodos estadísticos para analizar este sesgo en el uso de codones (codon usage bias). Métodos como el "Frequency of Optimal Codons" (Fop) (Frecuencia óptima de codones) y el "Codon Adaptation Index" (CAI) (Índice de adaptación codón) son utilizados para predecir los niveles de expresión génica, mientras que los métodos de "Effective Number of Codons" (Nc) (número efectivo de codones) y el método de "Shannon entropy from information theory" se utilizan para medir la uniformidad en el uso de codones.

Hay muchos programas que aplican los análisis estadísticos mencionados anteriormente, entre ellos:
CodonW: http://codonw.sourceforge.net/
GCUA: General Codon Usage Analysis: http://bioinf.may.ie/GCUA/
INCA: Interactive Codon Analysis software: http://bioinfo.hr/research/inca/
Codon Usage Database: http://www.kazusa.or.jp/codon/

Herramientas para utilizar los Métodos Comparativos

Como ya mencionamos anteriormente, este tipo de métodos de predicción de genes están basados en búsqueda de similitudes con secuencias ya existentes (de ahí que también se denominen Métodos Extrínesecos). Estos métodos usan herramientas de alineamiento local para comparar contra secuencias anotadas (proteínas, cDNAs, Est)

Alineamiento de cDNA con DNA genómico:
Blastn
est2genome (Parte de EMBOSS)
spidey (from NCBI, for Drosophila, vertebrates, C. elegans, plants)
SIBsim4 (sim4, http://sibsim4.sourceforge.net/)

Alineamiento de Proteínas con DNA genómico:

Blastx
genewise (wise2)
AAT (www.tigr.org)

Herramientas para utilizar los Métodos ab initio:

Como ya mencionamos anteriormente, los más utilizados se basan en el uso de Modelos Ocultos de Marcov (Hidden Markov Models - HMMs).

Genscan: http://genes.mit.edu/GENSCAN.html
Fgenesh: www.softberry.com
Glimmer: http://www.cbcb.umd.edu/software/glimmer/
GlimmerHmm: http://www.cbcb.umd.edu/software/GlimmerHMM/
GeneMark: http://exon.biology.gatech.edu/

Más información en castellano puede encontrarse en:
DETECCION Y MODELADO DE GENES:
http://novacripta.cbm.uam.es/bioweb/courses/UFV0506/tema05/Det_Mod_Genes/1_2_Estrat.html

Para una mejor comprensión de cómo se aplica este método para la predicción de genes es muy recomendable leer el artículo: *Eddy, 2004 Nature Biotechnology. What is a Hidden Markov Model?*

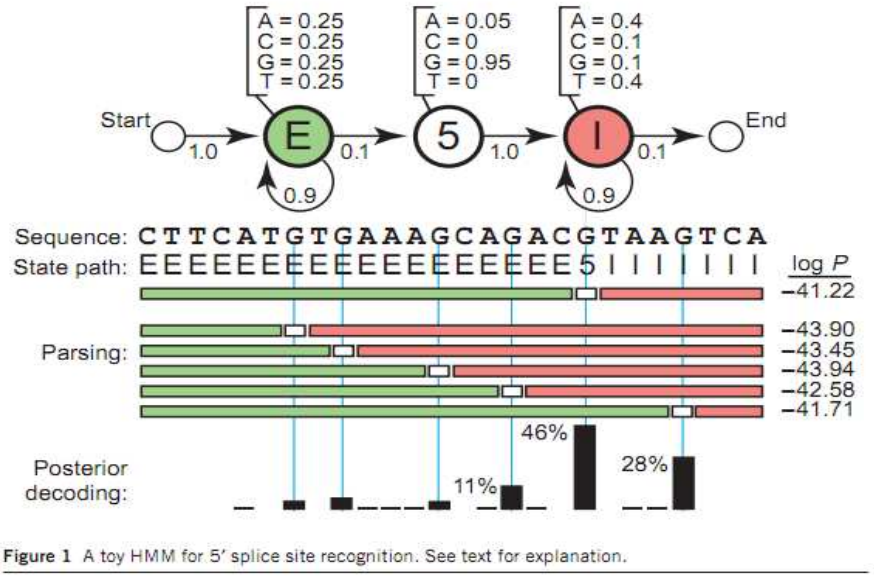
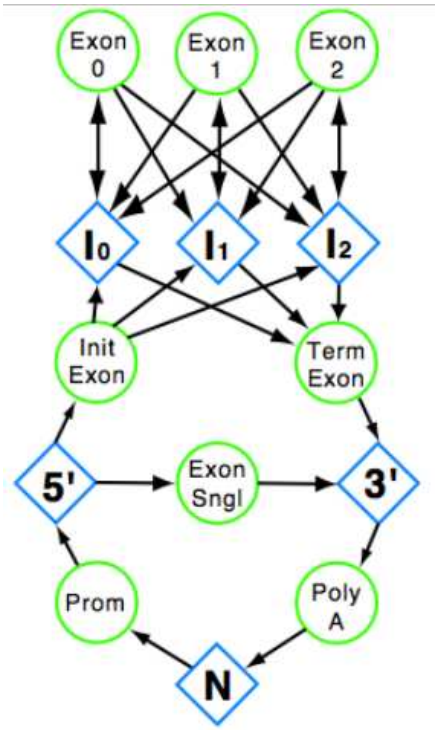


Figure 1 A toy HMM for 5′ splice site recognition. See text for explanation.

En un Modelo Oculto de Markov, se asigna un estado a cada nucleótido (codón de inicio, exón, intrón, región intergénica, etc.). Existe una probabilidad de **transición** (la probabilidad de pasar de un estado a otro nuevo) y una probabilidad de **emisión** - probabilidad de emisión u observación de un nucleótido en particular.

En la figura que se muestra arriba se destacan 3 Estados: E (exón), 5 (sitio de empalme 5') e I (intrón). Para identificar el sitio de empalme 5' sabemos que: el estado "5" es casi siempre G (95%) y raramente A (5%), el estado "5" siempre está seguido ("t*ransiciona*") al estado "I" y siempre se empieza con el estado "E". El primer nucleótido tiene una probabilidad de 0,1 de cambiar al estado "5", pero el segundo nucleótido sólo puede ser G o A. Por lo tanto el segundo nucleótido debe tener pertenecer al estado "E". P es el producto de todas las probabilidades de transición y emisión. Se calcula el logaritmo de P (log P) para cada estado posible y se determina cual es el modelo que mejor se ajusta (mayor valor de log P).

Genscan: http://genes.mit.edu/GENSCAN.html



Los estados en el modelo de GENESCAN. Las flechas indican las transiciones con probabilidad diferente de cero de los genes de la cadena "forward". Los estados **Exón 0**, **Exón 1** y **Exón 2** representan los exones internos con diferentes marcos de lectura; los estados $\mathbf{I_0}$, $\mathbf{I_1}$ e $\mathbf{I_2}$ representan intrones que siguen a los exones con marcos de lectura diferentes; **5'** representa a la región "río arriba" del primer exón codificante de un gen; **3'** representa la región "río abajo" del último exón codificante de un gen; **Prom** representa la región del promotor, y **PoliA** representa la señal de poliadenilación. Un modelo similar se utiliza para los genes en cadena "reverse". (Korf, *et al*. 2001. Bioinformatics 1:1 S1-S9).
Por ahora, principalmente para Vertebrados, Arabidopsis y Maíz.

Buscadores de Genes Combinados (intrínsecos y extrínsecos):

Geneid: http://genome.crg.es/geneid.html

- Puede combinar información de HSPs generados con Blast con genes anotados previamente y generar una nueva anotación con toda esta información.
- Identifica los sitios de empalme del intrón (intron splice sites) y codones de inicio y stop usando positional weight array (PSSM).
- Identifica exones mediante HMM.
- Identificar los genes utilizando un conjunto fijo de normas (los exones deben estar *in frame*, etc.)

sgp2: http://genome.imim.es/software/sgp2/index.html

- Combina búsquedas con tBlastx con geneid.

- Cuando los exones están definidos se incluyen los alineamientos de tBlastx

Twinscan: http://mblab.wustl.edu/software/twinscan/

- Se basa en Genscan, pero incorpora la homología de secuencia en las predicciones de genes.

Formatos de Archivo para describir las características de una secuencia:

Numerosos formatos de archivos se han establecido con el objetivo de "facilitar" la anotación de secuencias genómicas. El facilitar entre comillas se debe a que a primera vista estos archivos no parecen ser muy fáciles de comprender, pero uno tiende a acostumbrarse.

En la actualidad existen dos tipos principales de archivos: El formato **GFF** (**general feature format**, **gene-finding format**, **generic feature format**, **GFF**) y el formato **GTF** (**Gene transfer format**). Por ahora nos centraremos solo en el formato GFF.

El formato de archivo GFF se utiliza para describir los genes y otras características de una secuencia de DNA, RNA o proteína. El archivo consta de 9 columnas (campos) separados por tabulaciones. Aquí una breve descripción de cada campo:

1. **seqname** - El nombre de la secuencia. Debe ser un cromosoma o un scaffold*****.
2. **source** - El programa con el cual se determina la característica.
3. **feature** – El nombre de este tipo de característica. Por ejemplo: "CDS", "start_codon", "stop_codon", "HSP" o "exon".
4. **start** – La posición inicial de esa característica en la secuencia. La primera base es la número 1.
5. **end** – La posición final de la característica (inclusive).
6. **score** – Una puntuación de entre 0 y 1000.
7. **strand** – Dirección de la característica. Las entradas pueden ser "+", "-", o "." (para no se sabe o no importa).
8. **frame** – Marco: Si la característica es un exón codificante, el marco (frame) debe ser un número entre 0 a 2 que representa el marco de lectura de la primera base. Si la característica no es un exón codificante, el valor debe ser ".".
9. **group** or **attributes** - Todas las líneas con el mismo grupo se unen en un solo ítem. Aquí también se pueden incluir otros atributos de la secuencia.

*****scaffold: Al elaborar el mapa genómico, un scaffold se refiere a una serie de contigs que están en el orden correcto, pero no necesariamente vinculado en un tramo continuo de la secuencia.

seq	source	feature	start	end	score	strand	frame	attributes
Chr5	BLAST	similarity	2475	5120	.	.	.	Target "dbj|BAF16701.1|"
Chr5	BLAST	HSP	3142	3603	499	+	.	Target "dbj|BAF16701.1|"
Chr5	BLAST	HSP	4827	5120	438	+	.	Target "dbj|BAF16701.1|"
Chr5	BLAST	HSP	2475	2603	221	+	.	Target "dbj|BAF16701.1|"
Chr5	BLAST	HSP	2916	3017	182	+	.	Target "dbj|BAF16701.1|"

Para complementar esta parte del curso, y comprender mejor cómo realizar predicción de genes en eucariotas, daremos paso directamente a la parte PRÁCTICA:

PRÁCTICA:

Para esta práctica utilizaremos la secuencia "oryza_chr5" que se encuentra a continuación. Esta secuencia corresponde a un segmento de ADN genómico de *Oryza sativa* (arroz).

19. Primero utilizaremos BLASTN para alinear el ADN genómico a ETSs (expressed sequence tags). Para esto copiamos la secuencia oryza_chr5 (Copiarla completamente. Tener en cuenta que debido a la longitud de la secuencia ocupa 2 páginas). Pegar la secuencia en el servidor de BLAST disponible en la página web del NCBI. Realizamos un Blast de nucleótidos usando la base de datos "est_others database".
20. Examinar la ubicación de los alineamientos en la secuencia problema (la que hemos introducido). Estos alineamientos debería corresponder a las partes del gen que son transcriptas (exones y UTRs)
21. Seleccionar las 5 primeras ETS con los mejores alineamientos. Hay una casilla junto a cada nombre de la secuencia en la parte de los alineamientos en la salida del BLAST. Marcamos las 5 primeras y nos vamos a la parte inferior de la página donde hay un botón para recuperar las secuencias seleccionadas (Get selected sequences). Sobre la siguiente página, es necesario seleccionar las secuencias de nuevo con las casillas de verificación. Luego presionamos en "Send to" (en la parte superior derecha), pedimos que nos la envíe a un archivo (File) y en Format seleccionamos FASTA. Guardamos el archivo con las 5 EST seleccionadas en formato fasta.
22. Ahora utilizaremos el programa **Emboss - est2genome** (http://inn-temp.weizmann.ac.il/cgi-bin/emboss/est2genome) para alinear las ESTs con la secuencia de ADN genómico (oryza_chr5). Accedemos mediante el enlace anterior y en el campo "Spliced EST nucleotide sequence(s)" (2) subimos el archivo con las ESTs que hemos descargado en el paso anterior. Luego en el campo "Unspliced genomic nucleotide sequence" pegamos la secuencia oryza_chr5. Dejamos todas las otras opciones por default y corremos est2genome.
23. Examinar el resultado. Poner atención en la ubicación de los exones en el reporte. Debido a que cada EST tiene diferente longitud, estas pueden alinearse con distintas partes de la secuencia genómica y resultar en la identificación de diferentes exones.
24. Regresar al sitio del NCBI y realizar una búsqueda con BLASTX utilizando la secuencia de ADN genómico. Probar con diferentes bases de datos de proteínas y comparar los resultados. ¿Se corresponden los hits del BLAST y los HPSs con los exones encontrados con los ESTs?
25. Identificar los sitios de inicio y fin de los exones en el reporte obtenido con est2genome y compararlos con los resultados obtenidos usando BLASTX.
    a. ¿Cuántos genes hay en el segmento de ADN genómico?
    b. ¿Cuántos exones hay en los genes?
    c. Se puede determinar el "*reading frame*" de cada exón?
26. Usar el servidor de Geneid (http://genome.crg.es/software/geneid/geneid.html) para buscar genes en la secuencia de ADN genómico. ¿Coinciden los exones predichos por Geneid con los detectados mediante Blast?

>oryza_chr5
tgggatggtgatggagaattgagtaaggaactccctccttttttaatacctgggtaggggc
aggtaactgggaggaaattcctcctttacttttttctaagcaaatctcagctatccgtttt
tattaatgacctaatctctaactaaactccctatcaatttctatcatctctaccaaacaa
gatattaggtttaaaaatcaaattcccatcttaatctcatcatcaattccctcgtgtaaa
ttcccaatccccttctctcgagttatcaaacaagccgtcaggaaaaggacggcggcgcgg
cgagggctttgttgtggtgatcggccatacggtgtggggaatttcatctaatcccgattt
ttacggggaatcatattatgcgaaagaaaaggcgaaacaaggtatttacaaataaaaaaa
taatttataaataaatttttatataaatggtaacgattttaaaaaaggctgaaaaataaa
ctacgatgaaaaaattccaaaatcaactctacatttaacgttagaaatttaaatgtttta
ctaataagtataaacataattgaaaagatgaggagttttttttttattctgcactttagc
cctttggttatggggttttaccgaaacagataatatcactattaacaattatcgtacatg
ccatgagtttgatcaatggacaaagaaaccataaacccttccatgcgcacgcttcctata
ctattaaaatgtgtatttttttctgaaaagtttcgatataaaagttgttttaaaaaataa
tagcaatctatttttaaatttaaaataacaatatttctctcaaacactccaacttactt
cctcgttttccgcgcgcacgtttctcaaactactagatgatgagttttttgcgaaaaaa
aatttctatacgaaagttgttttaaaatattatgttgactcatttttaaaaaatagttaa
tatttaattaatcacgcactaatggattatttcgtgttttgtgtatggggagttggggat
aggcacctcccgaacacgagaaatttggcattatgccatctttggaatggggtttcgctg
aaatgtcattccggaacgtggcttcgttaaaataccactttcagacgatgccaacttact
ggaatgccatttttctccattttacacgaattgtaaatgttttggcacatacactaatctt
tttgcgccttacctaagccgatgatgcctttgagctctgccgccgattgtagccgagtga
ggaggacgggttgccatcggaataaagatggcgcatccggccgatgtcacctttgagcta
ctccgctgtcgcggcgccggtttccgctgagcgagaaggagtggtatgggtcatgggaga
tgggtggtgttatggaggaggaaggtgaggagctgggacaatagccgccatcctgaccat
agagcgaaccgaggggagactcgagttcgtcgggcttcctgtttctctcgccgccgttc
tgctttcccaacgctagagcggctcaacctttcactcatctcaccgtgggtgcacccgct
cctctcctccatcccagtcctcgctcgccatctattgctgcgaccctaccaagcttgtca
acctcgccgcactgggagggctccctctgtcgcgtcaagctcgtgcgatggcaccagcgc
tgcctgagcctctaggcgtggatctagagccactgctataggcctgccattttctccaag
gcacggggaggaggataatggcagtggtggctcgccggtgacggttcacataggaacgaa
agggaagagatggaatctagagggtattttggttagcttgacaagaaaaatagggtaaaa
ttcatgtgatttgataaaaaatggaagaaatagtagtagcatgttagccttgtttgaaga
tgacattctagggaagccaagcaatgtcaattttctcccccggagacagcctagtgcaca
gctgagagcatcctatttatcaacacaccctagccgtccattttccgtccctctcatcaa
cggccgtgatcaaacaccaacgctagtctccctctataaatcttcctccccccttccccat
ccaaaccctaacccccctccaaaaccctaaacctctccccctcacctgcggcggcggcggc
ggcggcgaagcacctcgtcggcacgacaccatgaggccaccgcgaggtaagcaaccaccc
gaaaccctaggattcatctcgcgctcgattccacttccacccttacacgatgctgttctg
cgtcaggcaggggattcgggaggggaggaggagggagaggcgatggcggcggccgcggcg
gcggcggtgggaggggattcgggagggtcggggacagcggcgggagaggaggccgtggtg
ggaggggaggcaggacgccgaggggccgtggcggcggccgcggtggcggcgggaggggtg
gcatgaagggcgggagcaaggtggtcgtggtgccgcacaagcacgacggcgtgttcatcg
ccaaggccaaggaggacgcgctctgcaccaagaacatggtccccggcgagtccgtctatg
gcgagaagcgcatctccgtccaggtcctttgctcttactccctccatttttttttgcgtt
ctacttatttagtgttttgctagaactgtagtatgattgtattgtactgtagcaagctgt
aagatgttgcaatgttgattctatgcagtgtacacagattatgttcaaatcgtgctagaa
ctgttaggataatttagtgcgctgtagaaagctgtaagatgttgcgcaatattgcttctt
tgtagtgtgatacaaagattgtgtacatgacgcttgtttctgttcatattgatctgtttt
attgtacatgactgttgtttctgcttttctgctatgtacagaatgaggatggaaccaagg
ttgaatacagggtgtggaatcccttccgttcaaagttggctgctgctgtgcttggtggtg
ttgacaacatctggattgtaagtttcttggttgaatccacgtgtgcaactttatgcttaa
gtggtttaaaatcactgtggtttgttaatgctgtagtataaaaaaatgcatcaagtgtta
atgtttgcatcatgattttaggctcctggtacccgtgtgctgtatctcggtggtgcctca
ggaacaaccgtgtctcatgtgtctgatattgttggaccggtgagatctcatttactgttc
tcatttattattcagtcacctaggtatgcaattttgccattcgctaatcaagtcattttg
cagaccggtttggtctatgctgttgagttctcgcacaggagtggtagggatcttgtcaac
atggccaaaaagaggaccaatgtaattcccatcattgaagatgctaggcacccggcgagg
taccggatgttggttggcatggttgatgttatcttctctgatgttgcccagccagatcag
gtatacctttctcacaatgatttcatgatcttgaaaacgaccttagataatttattatat
ggttactagttagtattcattaagtgtgtgacaatggttttggtgacattctttgattac
ttgcgggctgcaaatgatgagcttactcggattccccttcaggacatcaaattggatgat
gcagattcgagtttctgatgcattaaagcccattaccacctgtttttttctttgctgtttt
tgcctgctttttaaactgtctttgttgtcagttttcttatgtctagctgtttatattcca
gtttatgtttaagcaacctccaacccagaagttctattcctacacacggagtagtattct
tgtatgcttctgaagtacttaaattgtaacacgaagatgcattctttcttttgaaacact
tctgttatatatttttttgaatctatccttttttgaaattggttttcaattgttatgagtta

accatggctatggatttcttttggaagcgtttggtggccttgatgtcttttactaattta
tagaattatttacagcacgagtatacctgcagttttgcttctatatggctgtgttatgca
tatgaggactttgcatttcagtgccagcactacttactatgtccctatatccccttattat
ttaatattttgattgggcttagcatatgatgatcttactcggattccccttcaggacatc
aaattggatgatgcaaattcgagtttccgatgcgctgtgcctccattatttcattccctt
ctctttgcattttatgagtattctgtacatcatatttggattgtgctagcaatttgccat
atgcttggttgtttatactctctttagtatggtacaaccttcagcgatatgtgtgctttg
ttggtggctaatagtcatgtttttgacttttgcatattaattttcttgatattgtcttg
ttctttttctgtaatatgtcagctttcctggtatagctttctctcctattgttattttg
ttagtctagtgctaattagaactattccagtacttgccgtactacatttcttaattctgt
ggcagtgtcggacattgatatgtttttgccttgtaacattgcttacaataattcatgtat
gtttgctgctagagtctcggataaattgtctgcatataaaattccctttatggaatcttc
ttccatggccatgtcctttgttgaagatcttgaagattctcactgatttttgaaaccttt
cctatggaatcctctgtcctgcaggctaggatcttagccctcaatgcatcctacttcctg
aagaatggtggacattttgtcatttcaatcaaggtacttttgtcttttctctcccctgaa
tcctgaagataatcaccttgattgaggttgagattggggagatattagatgtaatactag
gtattggactgaaattgtttgaacaggcgaattgtatcgactccaccatgccagctgagg
ctgtgtttgctagtgaagttgagaagctgaaggcagatcagttcaagccctctgagcagg
tgacgctggagcctttcgagcgtgaccatgcctgcgtcgttggtggttacaggatgccca
agaagcaaaagggtacatcttaagaagaataaaacttgtggttgtgagtgactggtgcta
aaatgtagtgctgcacaaaactttggtctaattgtttttcacttgcaccctggtttgta
aggctgaaatttttggacaaatgatgctaggagagaagtaaaaagtcactcaggttttgc
actgccctttactgcaagttgcatcattattggtgctactaggggtgatacgctaatcaa
tagggatgaaattggtcaccgaaattcctcacagttcctgtttccgacagttagaccgtt
atcgtttttggccacatatatacaattttgatttttttttatgcgaacagtgaatatttt
aattgataaatatggtgaccgaccgagtgtctttatgacgacccccttcctaccatcaaa
ggttttcattttcgtcataaatcttgcgatgacttgccttttcatttatcagaattctt
gcctcatcttctacttatcagctaaaatttaaattttcaaccttaaatttagctaaaatt
taaatttttaccttaaatttagagctaattttaggggttttttttattaaagattattt
ttatcctttgcttttatattgttaaaaacatgtatataaaagttttattcataaattatt
atttttttacaaatacgccgttagcgaagcgaagtttctaattttcttgctctcaagtat
tataataagctaatgttataggtaacatttccatgcattgagagtataatgtggacgtta
aaattatcgtgacatgtccatgctcaacgcattgagagtatagatcaacgctaatattac
cggtaacatgttcatgcattgaaacgggcccttctgatttggatggtttgggagtattga
tggaacctaaagcacgggattgagagcaatgtggtgggaggagggagcggaaggagaatg
acttcacttttacgaaaaatattcaaccaacctttcgagaaaaataaaaaaaaatatcaa
ctacgaacttagttcagttaggccttattttgttcccacatcaaaatttttcacccaat
cacatgaatgtttggtcaaatatgtagtattaaatatgaaaaaaactaaatacatagttt
gtgtgtaaattgcgagacaataagtctaactgcgccatgatttgacaatgtgttgttaca
gtaaacatttacaaatgatggattagattaattaggtttaataaattcgtctcgcagttt
tctggcgaaatctgtaatttgttttgttattagattacgtttaatacttcaaatatatgt
ccgtatatccgatgtgacaaacaaacccaaaaatttaccccaactaaacaaaccctaagt
gaaatggttgctccaaaaaggaaacgggattcgctttttacattaaaaaactgctctgat
ccctgaaaaaaaactgcactgaactaatactccctacatccccaaacgtacgaaacagag
gctgtgtttagtttctgaaattggggagaagtttggggaaagttggtagtttggaaaaaa
aattgagagtttacgtgtgtaggaaagttttgaatgtaatgtgatgtgatggaaagttgg
gagttaggggaagtttagtgtgaactaaacacacccagaaaatgaaatctgaaaaataat
ttagcactcgagcaattttgttatgcgttttatttacttttattattagcaacgtggattc
cacgtgcaagcatactatttgtgaagaaagtagtggatttgttaacaacatacttcctct
agttttttatgacgttggctagtttaaatttatactggtgaacgttacacatatatgca
tggaggtagtatctttctgcctctacatgtgtcaacaatctttaatcaaacaaagtggat
tttgcttgctaactaaagattaaactttgggtcactcattgcttatgaaccaggccttgc
tacctcaactttggtcggttgggcttgtatattttgcttgtgatttggtgtttcagcatc
ggaaagaatgtatgttcttcaatgttggtatgtgatcgagttaaaatgcttttatagctg
taaatattgttgatgtatagcacgtgcattgttgttgtaaatcttaagagtggccagtgt
accgaagaagatccggcggtcacggcggcgaggcaaggaatttctggttgcggcgatgcg
ccgcacctatcgacgacaagcaaccatgcggtgcgggggactgagaatttgggtctgagc
tgaatttgttttaaaaagtctcaatttaggctcctcaaacaggggttgagtttggatagt
atccctcatcaacgaaactggaatagagctgattagttgactgattatgacctgccagag

Filogenia Molecular

La filogenia molecular consiste en la obtención de información filogenética a partir del estudio de secuencias moleculares, es decir, ácidos nucléicos (DNA, RNA) o proteínas. La información filogenética se utiliza para la clasificación de las especies basada únicamente en las relaciones de proximidad evolutiva entre ellas, reconstruyendo la historia de su diversificación (filogénesis) desde el origen de la vida en la Tierra hasta la actualidad.

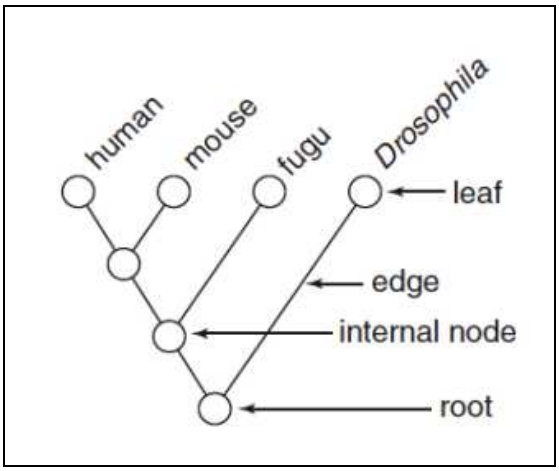
Algo de Terminología

La mayoría de los métodos filogenéticos nos muestran la información con una estructura tipo **árbol** en el cual las ramas y los nudos nos unen las diferentes especies. Recorrer las ramas desde su extremo hasta el nudo original nos describe la historia evolutiva de ese gen u organismo. Se basa en que un grupo comparte unas características comunes, que están mas relacionados entre si que con miembros de otros grupos y que provienen de un ancestro común. A este tipo de análisis filogenéticos se le denomina cladístico, que proviene del griego klados (rama). Los análisis cladísticos se suelen usar con conjuntos de caracteres, por ejemplo secuencias de nucleótidos o aminoácidos.

En estos análisis cladísticos hay que asumir ciertas hipótesis:

1- Cada grupo de organismos está relacionado por descendencia de un ancestro común
2- El patrón es de bifurcación.
3- Los cambios en las características ocurren a lo largo del tiempo

En la descripción de un árbol utilizamos los siguientes términos:
Taxón: es un grupo de organismos
Clade: es un taxón monofilético, todos sus miembros y sus descendientes provienen de un ancestro común
Nodo: es la bifurcación de una rama y corresponde a la existencia de un ancestro común de ambas ramas
Ramas: en algunos análisis su longitud corresponde a las divergencias entre los organismos o los nodos que une.



Nos vamos a centrar en las filogenias basadas en datos de secuencias. Aunque la mayoría de los conceptos son aplicables a filogenias basadas en otros tipos de caracteres.

Los análisis filogenéticos basados en árbol parten de ciertas premisas que se tienen que cumplir en casi todos ellos:

1- La secuencia es correcta y pertenece al organismo.
2- Las secuencias son homólogas y se han originado a partir de un ancestro común.
3- Cada posición de una secuencia del alineamiento es homóloga a la del resto de secuencias
4- Todas las secuencias incluidas en el análisis tienen una historia evolutiva común.
5- La representación de taxas es el adecuado para resolver el problema.
6- La secuencia utilizada es representativa del grupo de interés
7- Deben tener la suficiente variabilidad para poder resolver el problema
8- Las secuencias evolucionan con un patrón al azar
9- Las posiciones del alineamiento evolucionan independientemente unas de otras.
10- No existe intercambio de información genética entre los diferentes organismos (recombinación)

También hay que distinguir las relaciones de **homología** que pueden tener secuencias homólogas:

Ortólogas: Secuencias homólogas de dos especies diferentes y que provienen de una secuencia común

Parálogas: Secuencias homólogas de un mismo organismo. Provenientes de duplicaciones génicas.

Xenólogas: Secuencias homólogas de dos especies diferentes pero provienen de una transferencia horizontal.

Evolución génica y evolución de las especies

En principio cuando reconstruimos la filogenia de un grupo de secuencias esperamos que concuerde con la filogenia de las especies portadoras de esas secuencias; pero no siempre es así. Hay muchos factores que pueden hacer que las filogenias de secuencias no concuerden con la de las especies o que diferentes secuencias del mismo organismo den filogenias diferentes.

La mayoría de las ocasiones están causadas por que nuestras secuencias o análisis no cumplen las premisas necesarias. Por ejemplo: Escasa variabilidad y el resultado no es informativo, casos de **transferencia horizontal**, identificación errónea de las especies, etc. Pero hay casos en que lo que estamos viendo son fenómenos de especiación muy recientes o en curso, de forma que no ha habido el tiempo suficiente para que se acumulen suficientes cambios y lo que nos muestran los árboles es las relaciones de las diferentes poblaciones que las originaron y por lo tanto la distribución de cada gen podría ser diferente en esas poblaciones. Además en

estos casos, lo más probable es que estemos viendo recombinaciones todavía muy recientes entre las distintas especies.

Pero a pesar de que estemos utilizando correctamente las técnicas filogenéticas y las especies hayan divergido completamente; podemos encontrarnos que tanto la topología como las longitudes de las ramas no coincidan entre los árboles de las especies de las secuencias. Esto es debido a que las mutaciones no se tienen que producir en el momento de la especiación. Si una mutación se ha producido antes que la especiación, la longitud de las ramas será mayor en el árbol del gen. Si se han producido varias mutaciones antes de la especiación estás pueden perderse en las especies, mostrándonos una topología del gen debida a los efectos de la selección o al azar. Esto provoca que las relaciones de las secuencias sean diferentes que las relaciones entre las especies.

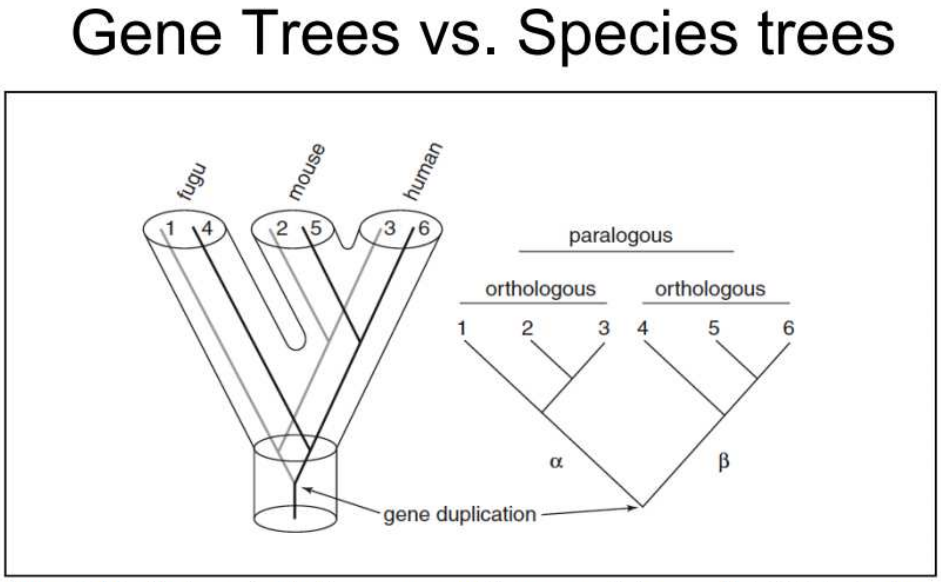


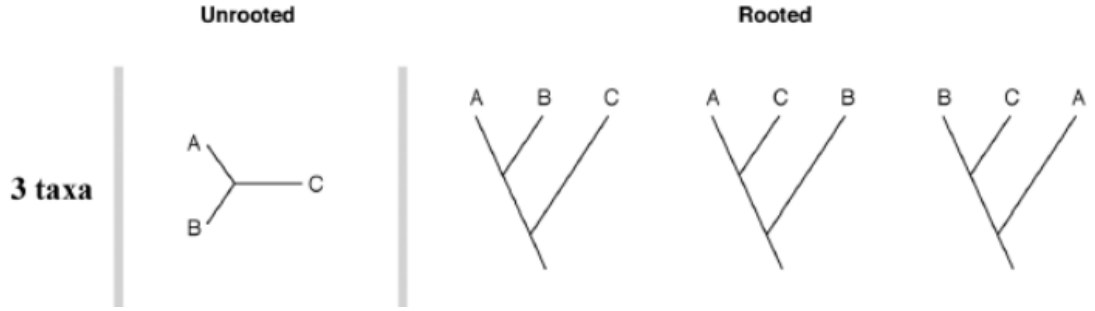
Figure 6.1.5 Phylogeny for the fugu, mouse, and humans, and six genes (1 to 6) that stem from a gene duplication resulting in two paralogous sets of genes, α and β. The α genes 1 to 3 are orthologous with each other, as are the β genes 4 to 6. However, each α gene is paralogous with each β gene, as they are separated by a gene duplication event, not a speciation event.

La evolución de las especies se ha dado en un tiempo determinado y las mutaciones se producen al azar, por lo tanto esperaríamos que las secuencias de dos organismos diferentes mostraran la misma variabilidad; es decir hayan sufrido el mismo número de cambios. En realidad esto no sucede así. La tasa de cambio, número de cambios en la secuencia en un determinado periodo de tiempo, puede ser diferente para cada secuencia determinada. Aunque las mutaciones se produzcan al azar estás mutaciones se podrán conservar en la especie o desaparecer; dependiendo de la acción de las fuerza evolutivas (migración, selección, deriva, etc.). Nos fijaremos en el ejemplo más fácil la selección, aquellas mutaciones deletéreas o que resulten menos eficaces serán eliminadas por la selección. Dependiendo la estructura de una proteína o secuencia las mutaciones que puede soportar sin que se vea afectada su funcionalidad dependerá de la propia secuencia y de su función. Hay proteínas altamente conservadas y proteínas que presentan una mayor variabilidad, debido a que los cambios en la secuencia no están afectando a su función. Por lo tanto, arboles realizados con diferentes secuencias pueden tener longitud de ramas distintas, ya que están evolucionando a diferente

velocidad. Si analizamos regiones en las que las cuales las mutaciones no tienen ningún efecto, son neutras frente a la selección; podemos inferir el tiempo trascurrido desde la separación de las secuencias. Esto se denomina **reloj molecular**: si asumimos que las mutaciones son neutras, se producen y fijan al azar y conocemos la tasa de mutación podemos calcular el tiempo de divergencia entre las dos especies.

Topología del árbol: con raíz o sin raíz

Los análisis filogenéticos nos van a dar la información de las relaciones evolutivas de las especies. La longitud de las ramas es la distancia que separa a las diferentes especies o secuencias y es aditiva, para conocer la distancia entre dos especies hay que sumar las distancias de todas las ramas que las unen. Los árboles producidos pueden ser con raíz (nos indica la posición del ancestro común a todas las especies del árbol) o sin raíz. La posición de la raíz se puede determinar o estar indeterminada; pero un árbol sin raíz puede tener diferentes topologías dependiendo donde se sitúe el ancestro común. Pero la información de las relaciones evolutivas que nos muestran los diferentes árboles es la misma.



Alineamientos de secuencias como base de los árboles filogenéticos

Los árboles filogenéticos se construyen a partir de **alineamientos múltiples**, ya que como hemos visto suponemos que cada posición evoluciona independientemente y suponemos una relación de homología en cada posición entre las especies del estudio. Por lo tanto, construir un buen alineamiento es esencial para la resolución del árbol.

La edición manual de los alineamientos es habitual para conseguir el mejor posible. Como vimos la mayoría de los programas de alineamiento nos dan una solución sub-óptima (la mejor posible utilizando pocos recursos del sistema); por lo tanto, a veces, se pueden corregir posiciones de gaps para optimizar el alineamiento. Muchos programas se basan en modelos de sustitución que no tienen en cuanta los **indels** (inserciones y deleciones), por lo que no cuentan los gaps en la secuencia para construir el mapa. El único método que trabaja con gaps es la máxima parsimonia, en este método los indels son tomados como un carácter más que pueden ser puntuados en el alineamiento (solo se tomará el gap como una posición única aunque cubra varias posiciones) o como un carácter independiente no sujeto a los modelos de sustitución.

Esto hace que sea necesario revisar el alineamiento para corregir las posiciones de los gaps y para eliminar aquellas regiones muy ambiguas que no nos van a proporcionar información fiable en todas las secuencias. Es decir, tenemos que incluir solo regiones conservadas que sean informativas para la mayoría de las secuencias analizadas.

Modelos de sustitución

Tanto en los alineamientos como en la reconstrucción de árboles vamos a puntuar los diferentes cambios nucleotídicos o aminoacídicos. Existen diferentes sistemas de puntuación para cada tipo y para cada método de reconstrucción filogenética. El método utilizado depende en muchos casos de las propias secuencias utilizadas; por ejemplo si utilizamos aminoácidos el modelo **BLOSUM** se suele utilizar para secuencias divergidas y el modelo PAM para secuencias mucho más conservadas.

Métodos de construcción de árboles

Existen muchos métodos de construcción de árboles pero se pueden agrupar en dos tipos: métodos basados en distancia y métodos basados en caracteres. Vamos a ver los cuatro métodos más usuales.

Métodos basados en distancias

La distancia es una medida del grado de divergencia entre dos secuencias; por lo tanto lo primero que se realiza es una matriz de distancias en la cual se calcula la distancia entre cada par de secuencias del alineamiento. Existen diferentes estadísticos para calcular la distancia entre dos secuencias, pero pude ser tan básico como el porcentaje de posiciones cambiadas entre dos secuencias. La mayoría de los estadísticos tienen en cuenta que las posiciones pueden estar saturadas; es decir en el alineamiento vemos los cambios definitivos pero no aquellos que han afectado a la misma posición varias veces. El mayor inconveniente de estos métodos es que no tienen en cuenta la información del carácter, solo la distancia que separa las secuencias.

UPGMA

Es un método de clustering (clasificación) que se basa en la identificación de las parejas más similares y en el cálculo de la media de las distancias entre ellas y el resto de las secuencias para reconstruir el árbol. Únicamente funcionará si se cumple la hipótesis del reloj molecular (todas las secuencias han evolucionado a la misma velocidad). ASUME QUE LAS SECUENCIAS EVOLUCIONAN A UNA TASA CONSTANTE

Neighbor-joining

El método de neighbor-joining es un método muy utilizado en las filogenias de secuencias. Presenta muchas limitaciones, pero es un método rápido y permite el manejo de muchas secuencias. Se basa en el diseño de una estrella en que cada punta

corresponde a cada secuencia. Selecciona las parejas al azar, las une y recalcula la longitud de las ramas. Repite el proceso para todas las combinaciones. Detecta la pareja que produce ramas más cortas (más cercana) y las fija, construyendo otra estrella y continuará hasta que solo quede un extremo. NO ASUME TASAS CONSTANTES DE EVOLUCION.

star-decomposition methods

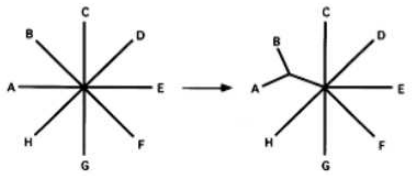


Figure 14.8. Star decomposition. This is how tree-building algorithms such as neighbor-joining work. The most similar terminals are joined, and a branch is inserted between them and the remainder of the star. Subsequently, the new branch is consolidated so that its value is a mean of the two original values, yielding a star tree with n-1 terminals. The process is repeated until only one terminal remains.

- UPGMA
- Neighbor-joining

Métodos basados en caracteres (Discretos)

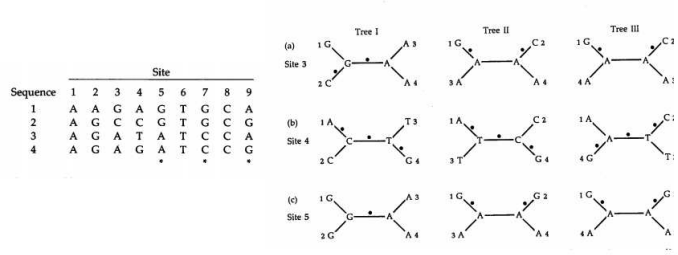
Estos métodos tienen en cuenta los caracteres durante todo el proceso, no se fijan solo en las divergencias.

Máxima parsimonia (Maximum Parsimony (PAUP, phylip))

Se basa en la filosofía de que la explicación más simple es la menos compleja, la que requiere menos cambios. Va a construir los árboles intentando minimizar los procesos de mutación. Como existen múltiples soluciones este método produce más árboles posibles que otros. Se pueden puntuar los cambios respecto a una matriz de sustituciones; por ejemplo puntuando diferente las transversiones y las transiciones. Pero muchos de ellos, nos van a dar puntuaciones muy similares ya que se puntúan respecto al número de cambios necesarios para explicar el árbol.Otro problema es la solución de los sitios muy heterogéneos; se puede solucionar eliminado estos sitios del análisis o revalidándolos respecto a su tasa de cambios en los diferentes árboles construidos. Este método funciona peor cuando la divergencia entre los nodos finales es mucho mayor que la divergencia entre los nodos intermedios,; es decir cuando las especies más cercanas se han diferenciado mucho más que los ancestros comunes.

parsimony informative sites

- sites (or columns) in a MSA that favor one tree topology over another



maximum parsimony

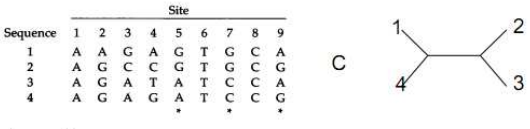
- assumes the tree with the least number of changes is the correct tree.
- No assumption of equal rates of evolution
- Usually, several trees with the same number of changes are found (many equally parsimonious trees)
- Time consuming to compute - there are too many possible trees

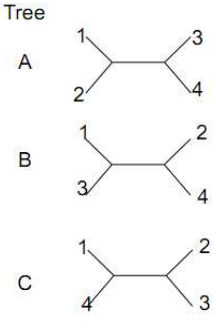
Maximun likelihood (phylip, PAML)

Busca el modelo evolutivo que tiene la mayor verosimilitud de explicar los datos. Va a analizar cada posición del alineamiento independientemente, calculando la verosimilitud de que un patrón de variación explique los datos con un particular modelo de sustitución y el árbol asociado. El programa asigna todos los posibles caracteres a cada nodo y luego evalúa las diferentes probabilidades de que ese carácter sea el del nodo. Posteriormente va a ponderar todas las posiciones del alineamiento dándonos el árbol más verosímil. El método calcula el modelo de sustitución que mejor se ajusta a los datos a la vez que realiza el árbol; por lo que necesita mucho tiempo de cómputo. La mejora de los algoritmos y de los ordenadores ha hecho posible que se comience a utilizar este tipo de métodos.

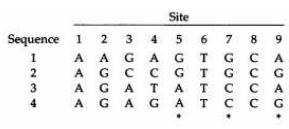
ML trees

Which tree has the greatest likelihood for generating the data in the alignment?



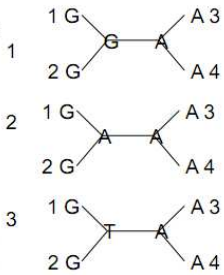


ML trees



Tree A, Site 5

Reconstruction Number:



4 x 4 = 16 possible reconstructions of tree A

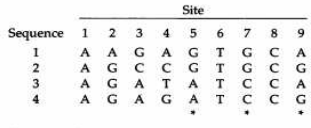
ML trees

Probability of a branch - compute using model of evolution (PAM, BLOSUM, JTT, Kimura's Method, etc).

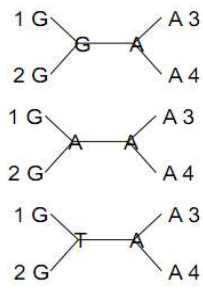
Probability of reconstruction 1 for tree A = product of probabilities of each branch.

Likelihood of site 5 in tree A = sum of the probability of each of the 16 trees.

Likelihood for tree A = Product of the likelihoods of each site.



Tree A, Site 5



4 x 4 = 16 possible reconstructions of tree A

ML Trees

- Uses all of the sequence data (unlike Neighbor Joining)
- Slow - many tree topologies must be examined. Impractical for large datasets.
- Phylip - use application proml

Métodos Bayesianos (Bayesian Methods (MrBayes))

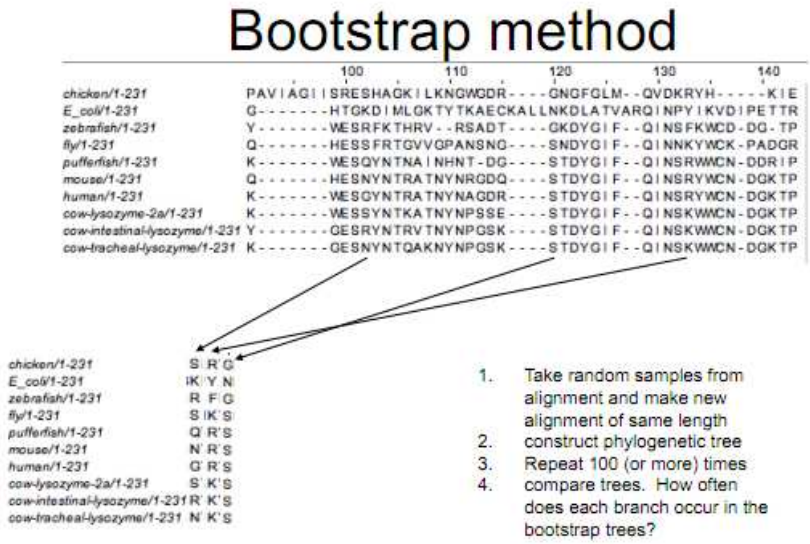
La inferencia bayesiana se basa la interrelación cuantitativa entre la función de verosimilitud y las distribuciones anteriores y posteriores de probabilidad. Esta interrelación viene dada por el teorema de Bayes, el cual permitecalcular la probabilidad posterior a partir de la verosimilitud y probabilidad anterior de los datos; está basada en la definición de probabilidades conjuntas.

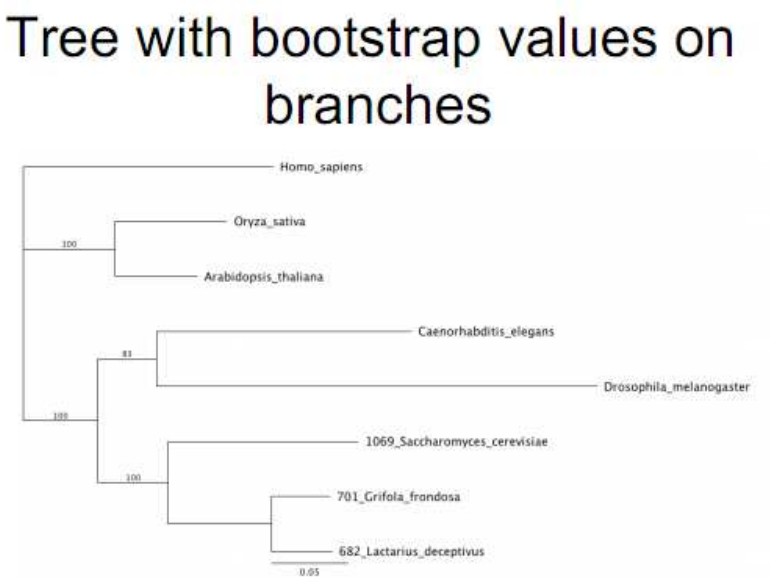
Quartets Methods (tree puzzle)

Basado en cuartetos (estimación de parámetros con ML). Diversos modelos. ADN y aminoácidos. Mecanismo de variación de tasas.

Métodos de validación de árboles

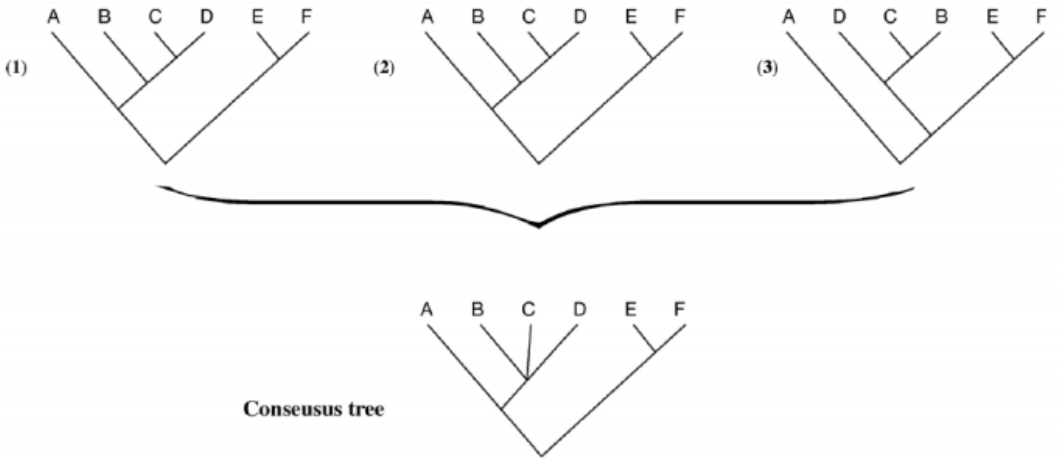
Existen diferentes métodos para validad los árboles filogenéticos. El más utilizado es el **bootstrap**. Este método se puede aplicar a todos los métodos y consiste en crear replicas de los alineamientos a partir del original, eliminando cierto número de posiciones al azar en cada replica. El número final de posiciones se mantiene constante añadiendo duplicaciones de los sitios que han permanecido Los análisis los realizaremos a todas estas replicas de los alineamientos y nos producirán tantos árboles como alineamientos derivados hayamos realizado. Posteriormente se calcula el árbol consenso indicando la frecuencia en que ese nodo está presente en el conjunto de los árboles. Se ha visto que valores mayores del 0,7 tienen un significado biológico alto. Por el contrario, valores menores del 0,5 nos indican ese nodo no tiene consistencia en los replicados; por lo que probablemente no se significativo.
Normalmente los árboles consenso obtenidos mediante bootstrap no muestran la distancia, únicamente la topología. Esto es debido a que la distancia calculada en cada replica no es real, puesto que hemos eliminado datos para obtenerla. Lo que podemos hacer es superponer al árbol sin bootstrap que nos muestras las distancias reales los valores de los nodos calculados con el análisis **bootstrap**.





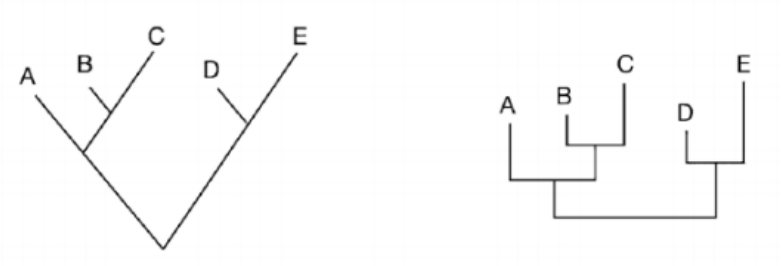
Árbol de consenso

Los árboles de consenso son un tipo de árbol que pueden ser considerados como árboles derivados. Estos árboles se construyen a partir de otros árboles y resumen a un conjunto de árboles. Suele tener dos aplicaciones básicas. Una es la de unir la información obtenida para la misma filogenia pero realizada con datos diferentes, por ejemplo, unir un árbol formado a partir de datos morfológicos y otro de datos moleculares. La otra aplicación que es la más frecuente es para reunir en un solo árbol la presencia de varios árboles resultado de ser los más parsimoniosos en el análisis cladista. Por ejemplo el método de máxima parsimonia nos daría como resultado múltiples árboles, ¿Cómo decidir cuál es el correcto? Generando un árbol consenso (por majority rule o strict consensus).

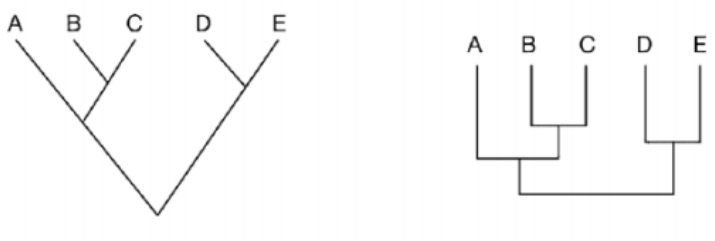


Formas de representación de árboles

En un **filograma**, las longitudes de las ramas representan la cantidad de divergencia evolutiva (están a escala). Tienen la ventaja de mostrar tanto las relaciones evolutivas como la información sobre el tiempo relativo de divergencia de las ramas.



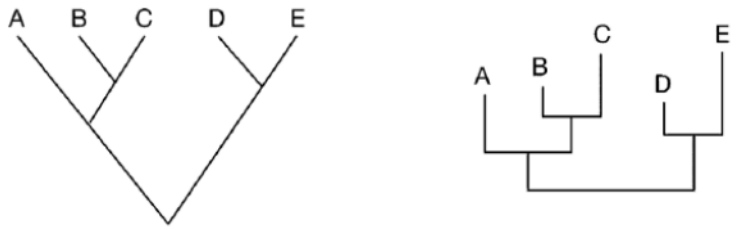
En un **cladograma**, sin embargo, los taxones externos se alinean claramente. Las longitudes de sus ramas no son proporcionales al número de cambios evolutivos y, por tanto, no tienen ningún significado filogenético.



Para proporcionar información de topología de árbol a los programas de computadora sin tener que dibujar el árbol en sí, se desarrolló una representación especial en texto conocida como el **formato *Newick*.**

En este formato, los árboles están representados por los taxones incluidos entre paréntesis anidados. En esta representación lineal, cada nodo interno está representado por un par de paréntesis que encierran todos los miembros de un grupo monofilético separados por una coma.

Para un filograma (ramas a escala), las longitudes de las ramas en unidades arbitrarias se colocan inmediatamente después del nombre del taxón separado por dos puntos.



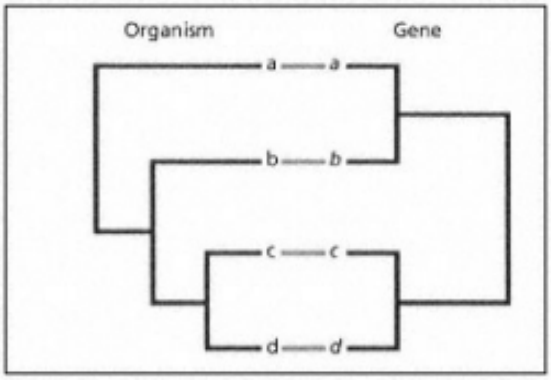
Formato *Newick* (((B,C),A),(D,E)) (((B:1,C:2),A:2),(D:1.2,E:2.5))

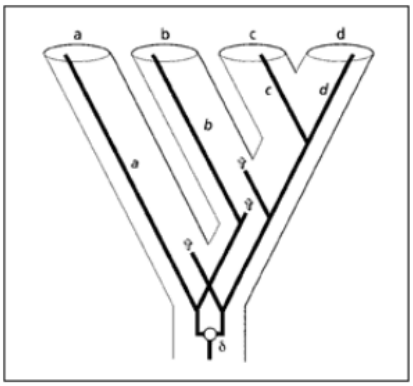
Árboles de Genes vs Árboles de Especies

Volvemos a tocar este tema, ya que estará directamente relacionado con la actividad práctica que realizaremos a continuación.

Como decíamos anteriormente, un árbol de GENES se construye a partir del análisis filogenético de la secuencia de uno o varios genes. En cambio un árbol de ESPECIES se infiere a partir de un árbol de genes además de otras evidencias como el registro fósil.

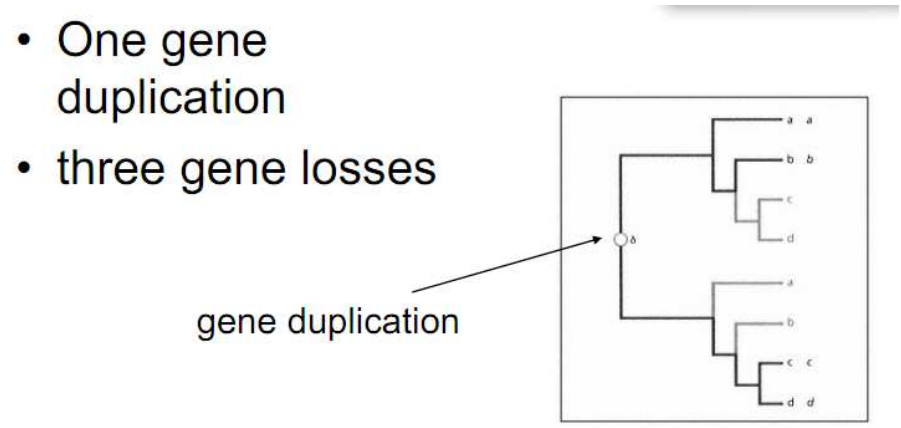
Muchas veces podemos encontrarnos con INCONGRUENCIAS entre el árbol de genes y el árbol de especies:





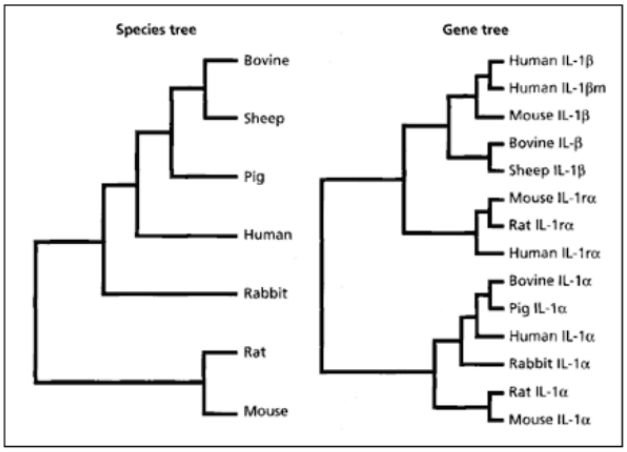
¿Cuál es la correcta?
La duplicación del gen dio lugar a dos conjuntos de genes.
Sólo cuatro genes se incluyeron en la filogenia – los otros se perdieron o no fueron secuenciados.

Reconciliación de los árboles:



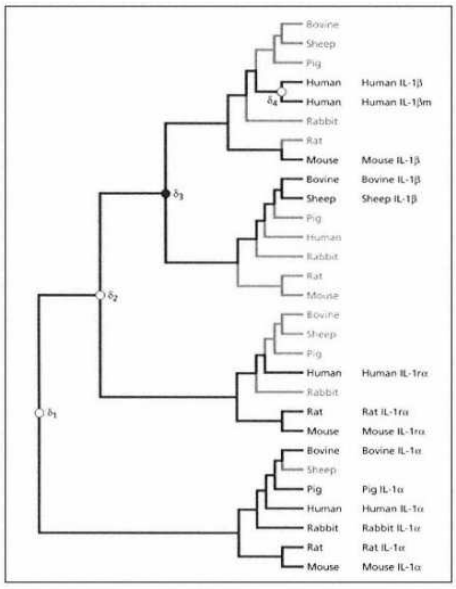
Otro ejemplo de “reconciliación” entre los árboles:

Mammalian Interleukin-1 genes



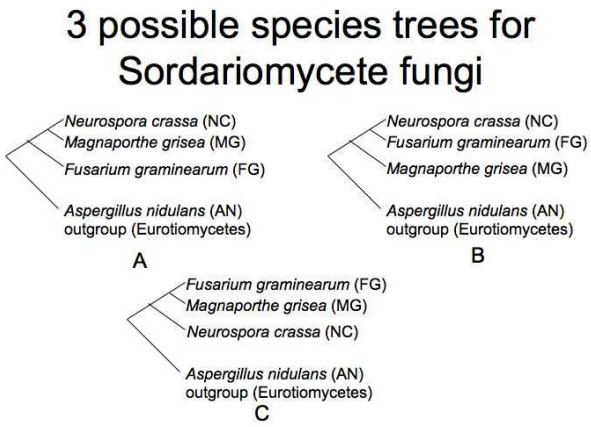
Reconciled Interleukin-1 tree

- Evidence for 4 duplications



PRÁCTICA:

Entre las cuatro especies de hongos de la clase Sordariomycetes, hay tres de árboles de ESPECIES posibles. En esta práctica, vamos a construir tres árboles de GENES y compararlos con los árboles de especies. Luego utilizaremos los árboles de genes para determinar cuál de los árboles de especies es el correcto.



En las páginas siguientes tenemos tres sets de datos:

a. Familia 915 (Chromatin Associated Protein)
b. Familia 979 (Acid Phosphatase)
c. Familia 838 (Unknown Function)

28. Abrimos la página de Phylogeny.fr (http://www.phylogeny.fr/) seleccionamos **"One Click"** dentro del menú **Phylogeny Analysis**.
29. Copiar y pegar cada uno de los sets de datos de secuencias (uno a la vez) en la página web y click en "submit".

30. Comparar los árboles de genes generados con los tres sets de datos con los árboles de especies que aparecen en la figura. ¿Cuál de los tres árboles de especies se ajusta mejor a lo obtenido con las familias de genes?

31. Ahora construir un árbol filogenético para cada una de las tres familias de genes **254, 268 y 279**. Cada familia contiene una o más duplicaciones de genes. Comparar cada árbol generado con el árbol de especies correcto elegido en el punto anterior e intentar identificar cual de los nodos internos representa un evento de duplicación génica.

>NCU04738.1
MSRSNPALAARYTDNEDAVMIDVDDVEDDDDEDEEMVQLKLQEIQARMKLKKIQNAKAQR
RANSSMDARSNGGRPESASANNNNNNNSMQPPAGLPIQSRLAAARDRMEQQSNSVQIPAS
PVRRGRSESAPQPPQSPTRLMLGIDKGRKAADMSLKRAPTLRKVESERPSQNLNSQQSGY
LRRSMTTSFSEDQPSSSSRRPVSFDGRASSFSQAEERRPLSFNERLACARNEEQERREKA
QRIQKIRSNAFSIGREEMEEYKTKAVDVPEVPYKAPEYSRDDILSATGRSSSRPTTMSAS
SSFTQPSSTQVNSDNEPGFEPYSGLRLSKRILPHNVVTRAITGKKTYGLKDLLRQVKAPD
WSLPDVESDVVVFAIVATKSEPRSHRSDGKPLQERGKYMVISLCDLQYEVELFLFNSGFD
RFWKLTPGTILAILNPTIMAPKQGQQDTGRFSLVINSDEDTILEIGNARDLGYCKSIKKD
GMLCKSWVNLRRTEYCEFHTNEAVTKSRQGRLELNNSFGFGDKYKQNSRYAKSEEAKRKE
EEIRLRGQYDRSTGSHFFMNRTAAELIDGGGMADQAEKKEAIKRSLARKEKEAEIARKLG
EVGGGAGRDYMSRAAGGLRASFSSTNSVPMSGSTSFSSATASTPMSSLSGHSQSNSTNDP
LGFGGRPRYDLQALGLLRKKGEEQPKIDLGPIKRKRPESAQSSENSFSKSNSTSSATITP
SNYTTNTKAPSLGWGSSLRDKLSRMKEGERLNLSHRPSNSSMRDSTPALTNSSSTTASTG
LTPAINRLSTSARDFNFTNSSSNTQTNLPNRPSPAVRFSAAPNNSSGTTTTTNGPTTSHH
HHHSSSLSSSRLLGVSASARLAAARAAAAAQAAAAADGHSQGSQPPVRKKTRFVTDKGIR
EAGRDSLGDKGFLMSAAQKNRRQVVLSEDDDDLIVLK*
>MG02482.1
MPSPPAQPSDEPQWPPRSPHEVLLSTPGGREKLRRLAERTSPSPSPARGRMSRTGSSGLP
LRSTNSDVMAGMILDGDDDEDEDEEILQLKLQEIQAKLKLKKLQSARAANGAPSSRPDSG
ADSNARRPQARTAAAAAAARAAKVEDARPRSVEVPASPVKKAIAPPKSPMKVQLGIDKGL
RAKDVSLRRSSSTLQQRMQDASERSRSLNDGAEAPVRPLSFSERLALKRNEEAAQVEKQK
RIQQIRSSAFAVDREEIERYKAVAQDLPDLADEPQVYSRAEIVGSGKLQKDGGYLRRSST
APSIRGGQAGSGTGSDSGQAKDATTTTAKSRGKDEASFESYSGFHLKKRILPHEIVTRSI
TGKQVYKLKDLLRKVKAPTWELPDDECDVVAFAILAAKSEPRSHKARTDSEGNTVNQGNR
GMYMVMTLVDLQYEVELFLFNSGFDRFWKLTPGTVLAILNPVIMKPPKGREATGKFSLII
NSDEDTILEIGTARDLGFCQATKKDGKACPSWINSKRTEFCEFHSNEAISKARAGRMELN
GAGFGGGGGNGKKSKKEWEPFSDPWKYKNSRPVSIDHRTEEQKLADAKRGSYDRDTQSRV
FVSRSASSASVIEREGEYLESLQRKEALKRKLLQDEHEKTVAKKLGNLGAGAGREYMQRR
AGHKTNDQPHALASNFSLEQQQPSTSAAALDLLNRAKNPPKIDLGPVKRKRPDSATSNST
AGGSSSLGWGSGLRDRLSRMKGQATLSFDQPQKTDNAAAANANSSSGSIASSKADRSPVR
KKTRFVTEKGIREAGRESLGTELLPSAAAAAADDALEPRRGKPRRVVVQDDDDDDDLVIV
*
>AN4959.1
MRRNRTLLASPLKNSATTPNSRTKATQLLSDGLGEDEGDDTEEDEETLQLKLAAIQARLK
LKQLQKNRGKAGTPHAERQQGADALSRPASSVSFSSFTQNQAPKSKFLQETKNQDGSGDI
QVPLSPTRRSRPQLEPTSPGRYRLGIDKGLKGSDVSLRRPPSSRGVETRPTSRLGSRDGP
ISHLRDTLAPQSHSFEDRRPKSFSERMAEGRAAEKARRERDVKVEQIRAKRSSAFEYDKA
EMEALKAAAAERRPNSPSKPTTRSGRTDSFSRDDIMRSINNIKPSLKRSQTTPSLRRTEE
TETKSFLQRRSQKPEAQSFPPQPGSSRTNSFTDEASNEAERVLEKPPDASKYEPFSQLHL
SNRILPHSFMDRTLKDKKVLRIPDLLRIIKGPPFELPEDIDTDYVVFGIVASKSDTKQVK
EAKSVTAQAADPFDDGLNNRNQYMAITLTDLKWTIDLFLFDTAFPRYYRLSEGTLIAILN
PTILPPPKGKLDTNRFSLSISSSDDTILEVGSARDIGYCKAVRKDGKICQVWVDGRKTEF
CDFHVDLQLRRAQADRMGVNNGTGLFGPGGRSGYRTGFYGGQKRSTQKGSGEGLLRADGA
HYDHGTQSLYYVAPSLKPNNTNKSLFHHLHPGQSAASLIDADVDDPFLAAGMMGRGAENK
GERFRRRLVEKQKEKEITQMITSTRAGGVAGEYLRAQNNENTRITDSTRSGAQHGRTDSR
DLKPQLSSLAGDKPLGMNFRRAEAVRLSPKKRAHDGDRPYGSGVKKTRFITSKGIKEAGR
DSLGGVSNATPRTDDYDDDDELEIV*
>FG09520.1
MAPASPAKQGNDSQWPPRSPHEALLSTPRGRERYRQMMTSPTSSPSKRGRNLPSMNLMAE
IENDEDDDDEETLQLKLQEIQARLKLKKLQSAKSKKTDDNVENLYATTSELAISSQPPSR
SRRATTPTRAELPTQENNIQVPASPVRRIQEPQVPISPSRVLLGIDKGLKARDISLKRAP
SYRDSQTTADIGHSGYLRRSRTTGSANSSFESRPLSFNERLASARKDDEARAQRQKEIQK
LRSNAFGISQDEMEQYKKKAIDIPDEPIKTPSFSRDEIMGIAKPAERQRSYTVPNIQSPS
INSPATTSMALTRRNKTPPGEVSDEQAAGIEPYSTFHLSKRILPHSVLARHVSGKKVHMI
KDLLRVVKAPDFALPDIEQDVVVFGILAKKSEPRAHKPTQKNGKTEDRGKYMVMTLVDLE
WELDLFLFNSGFTRYWKLTEGTVIGILNPTIMPPPPGRHDTGKFSLVINSDDDSIVEIGT
ARDLGGCQSVKKDGDRCGTWINKKRTHYCEFHSNEAIRKQRSTRLEVNGSSFGARKNNSR
EVYHFGAQKKEEPKKYDWETKTHWFASRTMSAGDLIDGKDRTPHDRKEKAEYLKRNMEAK
ERETEMMKKLGQVGNAAGREYMRHGGTRTSNMPISGSSIQSSSSSANIADEVFRPDARSL
GLLGKSAIHLSPVKRKRPDSSSAGSQAGSAVTNGPSAFGWGSSLKDKLSKMKDGEKLRKD
GQPPVRKKTRFVTDKGIREAGRESLGSELLNRQVMLDDDDDLVIV*

>MG02486.1
MHILVVNDDGPPSSQSSPYVHTLVRALQGAGHTVSVCLPHTQRSWIGKAHLIGQTVRPTY
YRPPPLPANPAGLIFDKDVPESEGSTHKRPTTTADEEWVLVDGTPASCVQIGLHHMFTDR
GKVDLVLSGPNYGRNTTALFALSSGTLGGALEGAACKVPSIALSYAFFTRNHDPEIVAGA
SRHAVRVVEALVKQWPQDGSVDLYSVNIPLVEGVEKARTLWTPMLQNYWADGGCFTAVEG
AGDEDEDETEERIRQGAGGEAALSSGAQNGKAAADGGEGTAHTRTHKHFKWQPRFTDVYK
SVEDAPPGNDGWAVKEGYTSVTPLKANFWQAATHLHGQDLKLPVLESHEGTDTLPIRSAA
GIGAAATEQSINKTSPAGDKGAYDSQDTGKLHALIAYDDPYVQPLIRDALGSLLPSDKYC
LVDAPEQQPEDKDNPVSLASLIPGDGTKALQIASYEAIDFDYAAEHQNSIVINSYKFRKA
LIRKHFLATTVENWVIKHPNSILARHVKPSTSFEVDYAEFLDDALVEAWDLRASLENDEK
REWWILKPSMSDRGQGIRLFSTMEELQGIFDEWDVDTDDEDEEADGIMAAHLRHFVAQPY
IDPPLLLPGDRRKFHIRTYVLCVGAMKVYVYREMLALFAGREYRAPSSDEGDLDAHLTNT
CLQNETTAASATATAADIATSDAKDGEAPPPPLVRKFWDLPEVLVEGRSDVSKGSIFTQI
CEATGAVFEAAARGMPMHFSPLANAFEVFGVDFLVDAQGVAWLLEVNSFPDFKQTGDELR
DDVVAELWRQVLRLAVVEGGPLRLGGGKGDRDDEGHGDATGATTKTLSAAAAAGDGKMVL
VKDVDLGKEGLLGFLG*
>NCU04743.1
MHILIVNDDGPPSAHSSPYVHSLVRDLQAAGHTVSVCLPHTQRSWIGKAHMIGQTVKPLY
YRPPPASSPAAGLTTALVPSSEKPEQTVNVTDHGSVHLRPSTVPGTEEWILVDGTPASCV
QIGLYHFFQDRGPVDLVVSGPNYGRNTTAVFALSSGTLGGALEAAVCKRRAIALSYAFFN
RNHDPAIITKASRQSVRVIEALWKQWPTDGSVDLYSVNVPLLEGLEEGKVLYTPMLQNYW
GAGSCFEEVEGSVDGEEVDEERIREGGGADAETGDGGGGLEVGDGKRDGREGLHTHKHFK
WSPRFTDVYKSVEEAPPGNDGWAVKEGHTSVTPLKANFWNTAENLHGKELQLPPLETPSA
IKTESTTTLPIRNSTTTANSKDHLYALIDYQDAYVQPLILSAIEKLLPSSSYTLLPSPFT
SENKEPEIHLSTLLPSEDAKVLQITPYETIDFDHAMSHPATTLINSYIIRKALIRKHFLS
STVENWVAKHPTSALKTHVKRAEAFEVDYAEFLDDSLVEAFDLRASMEKNDQLIAEGKER
EVEWWILKPSMSDRGQGIRLFSTMEELQSIFDSWEVESDSEDEDDDDASSSNADNTSSNA
DDTSDAGSDHGDGNGNGINTSHLRHFIAQPYIHPPLLLPELSNRKFHIRVYVLAIGALKV
YVYKDMLALFAGVPYTSPTTSSSSDPDSNPDSEPAGELDLSAHLTNTCLQTYLSPNAAEN
SVHRFWDLSSLSPTHFPQSKAENIWDQICEVTGDLFEAAARGMMIHFQPMEQAFEVYGLD
FLVDADDESGRNTAWLLEVNAFPDFKQTGELKGVVGGFWEGVVREAVGGFVGVQKGDSDE
KMRLVRDVDLGRRW*
>FG09516.1
MHILVTNDDGPPSPHSSPYVHCLIQQLQQAGHTVSPLYYRPSSVVHGDESPGTTHHRPSP
SGDVEEWVLVDGTPASCVQIGLHHFFQDKGPIDLVVSGPNYGRNTTAVFALSSGTLGAAL
EAAVCQKKSIALSFAFFTRNHDPVIIEAACRRSVKVIENLYKQWPTDGSADLYSVNVPLI
EGLENNKAIWTNVLQNYWREGGCFQEIEGEAGDENEEEERIREGVGGEVDDAARPSSRKG
HTHKHFKWAPKFTDVYKSVEESEPGNDGWAVKEGLTSITPLKANFMLGAGELFNQKEFEL
DSGSVANQSTQEMALRPKGPSIQAVISYEDAYVQPLILSALNSIFPEGVFNVITEVPESD
EPALAKIVPSEENILQITAYESIDFEYAGSHERTTLINSYMIRKALIRKHFLSTTVDHWV
AKHPESVLKTHIKRSEAFEVDFAEFLDDALVEAFDLRESMDRNEEQSDPSSKEWWILKPG
MSDRGQGIKLFSSMDELQNIFDIWEEDQPDTDDEDEVADNDNDGDGITTSHLRHFVAQPY
IHPPLLVDGEKRKFHIRTYVMCSGSLDVWVYKHMLALFAGKPYTAPADAPEDIESFLTNT
CLQDSPNENTVRRFWDLPLSNDMRDDIFRQICDVTGEIFEAAAKAMPIHFQTMPNAFEVY
GLDFMVDAQGTAWLLEVNAFPDFKQTGGDLKEIVSGFWKGVMRHGVAPFFGIESKIRDQE
GAEDMVPVRKSIKGNQSPPYHNLAATREFKVALKKDNNSQLGLRSNYTATM*
>AN4967.1
MHILVVNDDGPPSNESSPYVHSFVHTLQSAGHTVSVVLPHQQRSWIGKAHLIGAAVKPTY
FRPGTLHKDDGTTHEYPRGDNDPDGDEWILINSTPASCVQIGLYHYFQDRGPVDLVVSGP
NYGRNTTALFAMSSGTIGAAMEGAACGKRSIALSYAFSSRNHDPVIIAEASRHSVRVIEY
LAKNWDEGVDLYSVNVPLEPGVSESKVMYTEMLDNRWSSGSCFDAVDAEVPVENPEQREQ
TLRDQEEKLEEDPTANPNAKSGRKSRIAHKHFIWAPKFTDVYRSVEESAPGNDGWTVKEG
MTSVTPLKANFMHTPGIKGEIKLSDNEEPAFYSIVDCDDPYVQELVEQALRFSMGSRCRS
VSSISELPSRSTPVFQYREYERLDFEHAMTNPTTSLINAYIIRKALIRKHYLANTVSNWV
TKHPESVLAKHVKVSVDFELDYAEFLDDALLEAYELRECFEENESRPDSEKVWWILKPGM
SDRGQGIRLFNSEDQLREIFEEWEPDSEDEEEDECGEDDGNDETDKAATSGVVTSQLRHF
IAQPYIDPPLLLPSLNNRKFHIRTYVLATGSLKVYVFKEMLALFASKPYVSPTSQNNETE
DSIADLTRHLTNTCFQDKSLPESETVRRFWCLPSIPPPNTHLTATWKEDIYEQICAVTGE
LFTAAARGMMIHFQTMPNAFEVFGVDFLVDDTGNVWLLEVNAFPDFGQTGEELRDVVVGG
LFKGVIGVAVKGFFGEEGKTEENGMRLVAELDLGRKN*

>MG01294.1
MSDSNRGRGRGRGDGGSRGDRGGSRGRGGGDRGGDRGRGGSFRGDRGGGDRGGDRGRGGG
FRGDRGGGDRGGDRGRGGGFRGDRGGGDRGGDRGRGGFRGDRGGRGGGGFRGGRGGYQDS
GPSIYKPAGGVPAPDKSITALEDKWAENAANISVTELTERAGKLGITSSTLDTTQVTAIL
PQRPAYGTTGTPVILWANYFSMNVKSQTLFKYALKVKRSGSDEDVVGKLLRTIVRKALDQ
VAVQNPKNKIVSEFKAKVVSQGKLILPPGGGPVLVEHTGRKRAEEYQVTFSPPEDIDVAK
LVEWLRTMNDRLDDIVPTFPKFASTIDAIGIIMGHYARTSPGVVPLGASSARFFPSEANE
LREFVQMSAMKNLVRGYFSSARPATGRLLLNVNVTHAVFQRSGNAMELIKALLNECRVRD
LRSPALRAASRKIAKLKMEVTLFDSKGKKCGVSERSILELSRTTASTTMFKLDKPKDAAA
ETWADGSPIQYGGNVSVAEYYRKKFKQTVDDRLPLIVTGSHKRKLYFPSDFCRILPDQCS
NGKLSAGEASAMINFACRGPARNAESIVKVGAETLGISGAVNEILKAFGITVDANLITVQ
GRVLTAPTLAYLNKTNKSQTVTVRDGSWNLRDLKVVKGGTVPSWGCLTIIDPNDRYGDVS
FDDVKITMSAFVQFLNQNMGIRIPGLTNAADQLKTCQIQEGDEYKTIKQGFEQLKGLKPM
MVFVILPDSKDAAVYNAVKRIADIDLGVHTVCMVRKNLFKNGPGQNPQYYANVGLKVNLK
AGGINHKLSQDIPVSKGGKAMFVGWDVIHPTNLGVDKDSGLPSVVGLVSSIDEHLAQWPA
VAWAQKGGQEMADSRLEERFGSRLVLWQKHNQGRLPERIIIFRDGVSEGQFATVENTELP
LVRKACAKVYGNKPKPKITIVVSVKRHQTRFFPTDEADMSRSMNNKSGTVVDRGVTVARY
WDFFLQSHDALKGTARPARYTVIHNEIFPTVKGANAADELERLTHSLSFLFGRATKAVSI
CPPAYYADIVCTRHRAHLGELFDSVITGGSVVSGQTGSSASSNNEQQILARADALNVHRN
LIDSMYWI*
>FG08752.1
MADRGDRGGRGGRGGRGGGYQDGRGGGRGGGDGRGRGDGGFRGGRGRGEGGRGDFRGG
FSGPRGGRGGSDRGRGGYRGGRGGGQFSNEPSFFKIESGVPQPDAAITKLEDEVVKNQNN
SVAQLTSKMSKLGVEEKEHLSKFPPRPAFGNKGRPVTLWANYYQIDTNIPMLFKYTISVK
EIVAESEEKTDKPTVAQPKGKGKGKGKPGKPKKSGSVEVKGRKLFLVIKETLNELTKKDK
SLLLATEFKSQLISLRKLDLGLDNSFRVNLPSSANPDKTEVFEVTLNGPEVARVDEMLKY
VKSNTSASHDAPGKLADDDDDAAKALAFPKFPDVVDALNVIFGFGPRSNEDISAVGNSRF
FSFKNGGICRDMSMRGRPLQAVRGTFQSVRLGTGRLLLNTNITVGIFKISGNCAKLFQDL
NVFEAQKSEWRKVNNAKLMNKFLPKTRVLATMKFANDKKVQRRKAIYSLAYAPEIERACR
GNDHPPRFTKGYEYPGPGNVSFYMVSDSKGTGEYITVREFYKRKYNDTTLKDYPLLNLGT
AANPNFTPAEYVEILPGQSVKAKLNSQESTAMVDFACRSPYANALSITKDARETLGLDDE
KLGQFGIQVGKQLLTVHGRVLNAPGVSYYDSNTRLVQVHPREGSWNMSAKQVYKPGKSIQ
KWTYVNVKPGGRSGPVPKRTVMEFAQVMRQMNIGISSNPVDPCTEVITQEDYAQGRSDAF
FKWAKQNRIEFILVILGTSESETYGRIKTLGDCTYGIHTCCAQAEKFGFNRNPLPYFANC
ALKWNLKAGGVNHKLHNEFGLIKEGKTMLVGYDVTHPTNMPSGQGDDAPSLVGLVATIDR
DMGQWPAYSWEQSSKQEMLDETLTEAFKSRLALWQMHNRQQLPENIVIFRDGVSEGQFAQ
VLQRELPRIRIACNAKYPKNKPPRISLIVSVKRHQTRFYPTSSESMTSKNNIENGTIVDR
GVTQARYWDFFLTAHSSIKGTARPAHYTVLLDEVFRAKYGAEAANELERYAHELCYLFGR
ATKAVSICPPAYYADVVCTRARCYRPEFFEISDVESVSTAGPGLGASDPKQVHADLANSM
YYI*
>AN1519.1
MSSAGGSPQRGSRGRGNDRGRGRGLFHGDRGRGRGGGRGLFNDLPHRPAPGDPGRGGSRG
RAGRGGRGGGGGLDQGPPIYLPPDGAPQPNVKVTQTENSQAAALVKKEKTAGYPERPGYG
TQGHPIQLFANYLELKSSGKSLFRYHINIDGGGRKPSSRKAKQIICLLLEDHFSPFRHSI
VTDYRSNLISHLEILDHEQPSVKYNVTYRSEKEDEPRDTSETYRITCKFTGRLDPADLLN
YLTSSNAASMLQEKAEILQALNIVLGHHPKSTGSIASVGTNKHYAIHDNAAEKFDLGAGL
EALRGYFVSVRAATARLLVNIQVKYVACYQDGPLYQVIREFQCANGRNVYALKRFLGRLR
VEVTHIKRKNKRGEYIPRIKTITNLATPQDGTQDGNKCKDAPKVKFIGAGPNDVSFFLDD
PEQGKGSKKAGPKPSGTYITVAEFFKEYYRIQVDPDMPVVNVGSIAKPSYLPVEVCDVLS
GQPAKTKLSSNQTRQMLNFAVRSPAQNAHSIVTKGTQILGLRDPTAATLVDFGIQTNPNL
ITVPGRVLAPPTVYYKDEKSKDKEIAPMSGSWNMKSIRFSTSSNLQSWACILITAGPKQH
FQSPDDLEDCLYRFTKKLREVGVNANPPVFKVRVQVTKENAETVIDAEIRKILHQHRPKL
ILTILPFNDTALYNCIKRACDVRHGVRNINVLAEQFCKRNEQYFANVGLKFNLKLGGVNQ
VVRPSQLGIIGEGKTMLIGIDVTHPSPGSAKGAPSVAAMVASVDSSLGQWPAEIRIQKEA
RKEMVDALDSMLKAHLRRWAANHKAAYPENIIVYRDGVSEGQYDHVTDEELPLLKNACKN
IYPAPDTARNLPRFSIIIVGKRHHTRFYPTLQEDADRFNNPVNGTVVDRGITEARNWDFF
LQAHTALKGTARPAHYYTVWDEIFLRQKVIPPAKNAADMLEAMTHHMCYLFGRATKAVSI
CPPAYYADLVCTRARCYLSSAFEPSTPSGSVIGAEDSTVKVANDDVLIHPNVRDTMFYI*
>NCU04730.1
MSKLSLSEKEKANNLPVRPGHGTMGEKVKLWANYFKINIKSPAIYRYTIKVAATEEKLGK
EAEVASKKVEVVVGKLLKQIEANVKSVAIASDFKVHLVTTTKLKVPENRIFEVTWTEPSS
NQNLPSKPQTWVVKVEESVETCDFGKVLNELTTLDPKLDGDFPKYNVELDALNTIVTHHA
RADDNVAVVGRGRFFAIGDDLIEQVRPHDSPLVILRGYFASVRPATGRLLLNTNITHGVF
RPGVKLAQLFQELGLDVMDKCNAWNEVTKNQLNDKMRRVHKVLAKGRVELNAPFLIDGKI
VYKKCYRTLNGIANRGDERGKQKDGKEVRYPPLFGIPGVQVGGPTSCQFYLRARETKDGA
APPPTPGLPSNAYITVANYYKQRYGITANASLPLVNVGTKEKAIYVLAEFCTLVKGRSVK
AKLTANEADNMIKFACRAPSLNAQSIVTKGRQTLGLDKSLTLGKFKVSIDKELITVVGRE
LKPPMLTYSGNKTVEPQDGGWLMKFVKVARPCRKIEKWTYLELKGSKANEGVPQAMTAFA
EFLNRTGIPINPRFSPGMSMSVPGSEKEFFAKVKELMSSHQFVVVLLPRKDVAIYNMVKR
AADITFGVHTVCCVAEKFLSTKGQLGYFANVGLKVNLKFGGTNHNIKTPIPLLAKGKTMV
VGYDVTHPTNLAAGQSPASAPSIVGLVSTIDQHLGQWPAMVWNNPHGQESMTEQFTDKFK
TRLELWRSNPANNRSLPENILIFRDGVSEGQFQMVIKDELPLVRAACKLVYPAGKLPRIT
LIVSVKRHQTRFFPTDPKHIHFKSKSPKEGTVVDRGVTNVRYWDFFLQAHASLQGTARSA
HYTVLVDEIFRADYGNKAADTLEQLTHDMCYLFGRATKAVSICPPAYYADLVCDRARIHQ
KELFDALDENDSVKTDDFARWGNSGAVHPNLRNSMYYI*

>FG06663.1
MLIALAAVSRASPTNNDECNCYLTNGSNSAYYSQHKFYDFRNLSKYAQVPSPIRNSSRKA
GVTSDYFKSDTWDNTWSIQDWNNRGKGGVSLSGDATVFMANSPNNVYIQKNDDDDAASDT
FLTMRTMRMPGFQSAAEFESVSTYHYVSVRMLARVTGGAGACMALFTYLEGEDLADVQEA
DIEILTRDPKNRIQYTNQPSFTDDGDDIPKATRNGTLPKGLGWDDWVVHRLDWTPERSVW
YVQGKEVASIEFQTPKDPAQIILNAWSDGGSWSGNMSLGDAAYMQIQWIDMVYNTTKDGE
DKRSLDGEMPERSVGRESSLFRRGDNDDECKVVCSIDDVDKAGETKVLSQSAASHMVMHW
TKAIVIGCVILMAIYILLATSRNDVVSPSLEHITPSSLCHQTLLLYNHHHKQQQPFSPAI
P*
>NCU00061.1
MARESGTDWAPCCYIQPSRGRTHTSLPSLITLVAIVSPLASAVPSLTDDSKCECYLTNGT
QASFFATHEFLDFRNLAEHAGIPPTITKPNDSGSAPVTSEYFTSKEWTEAFWVLSWNNSH
QIREDATVLMVNSPNNVYIEANSDRNPSSQTWLTLRTQRLKDFQTASEIESVAPGFQYLS
VRMLARTVGSPGAITALFTYRDADRLADVQESDLEVRTMDPRNTIQYTNQPSYTEEGEEI
DHATKNVTLPPGHPDWTHWAVHRLDWTPKETVWYVDGTEVATISFQTPRDPSKVILNAWS
DGGEWSGNMSVNDAAYLQIQWLEIVYNATDENPAKRKRNDSGVAAVMPRERDEDGEGGCD
VVCSIDETPEPGKPVMLWNNGAMRMVGGGLVAWIPSLVTFGMLALISGGLW*
>AN2690.1
MRLRPSLHPLPVLSLVASTVSIVLPPANTTLTTTAANVPNLAPRIPSHYPCDCYIVSGDE
PGYFTDYQFWDFRDVPLPQSLISDGYGPSTVSHWEAETVPLSQTPFSKDWQTQSWSRQET
TDSTVPMVNDDANAFFAKHPNLPAASQLVLRTTRLEDYSSSAEVESQHGNYFHVSIRVRM
RLMSGEAISRRPWDETPDVNEVPKGACAGIFTYRSATCESDVEFLTSDPPNTIHYANQPD
YDAENDIIIPGASEVVTTVPVPWSEWVTHRMDWFANETVWYADDELQAVVSKSVPDRPSI
LALNLWSDGGLWTGDMQVDDSVYMGIEWIEIAYNTSAAGDAPIETGQRHRVRPSERTKRS
SHRKRQTSGDDAGESDGVITQLSRAQKSIYKYFVRLLSLEFDVEQR*
>MG08823.1
MNLRTVTGSASSILLILSSIHVAVADLPLVSDSKCGCYVTNGSNSAFFTSHRFYDFRSLS
QYVGSGIPDAPSSSPQAASNASITSGYFNTSAFTDDWSIMTWNNSASVGGDAAILMVNSP
TNLFIQKNNDTDADSDTYLTMRTARNEDFQSASEIESNSNRYRFVSMRMLARTTGAPGAC
TALFTYRSAKDYADVQEADIEVLTKDPTDKIMYTNQPSYVKVGNDTQNIPESNINGTAPV
DWTQWSVHRLDWTPLRTTWYVDGRQVASISFQTPRDASRINLNSWSDGGDWTGLMDVGKE
ARMQIKWLEMVFNQTSSGTSRMRPRDHTGNSGRGSGNGNSGGASDSGSTEGLSDGKCRTV
CSIDDTDQRGMAAVVADGGAATLKLAAWIPISVALVTMAASMF*
>FG10621.1
MRPSRKVVILALAATAIAQNDTQPEEKPEDWESDQCDCYLTDGRDPGYYTKHRFWDFRNL
AEYAGIPDTIAGENASAEADVTSKYFKTKAWKNFWSIQSWSNRRSHDTLSYGAMFPMVNS
PNNIYIRSNPDKNPSSETYLSMRTNRQEEFQVASEFDSIDRFQFLSIRFMGRVRGAPGAC
MAMFTYVPAKEIKDVQEADMEILTREDHDRVHYTNHPGYSATELFPKATRNITLPDGLKW
NEWVEHRLDWTPTQSIWYANGIEAANISFQVPRDPSLLIFNSWGDGGVWTQNMTTGQEAY
MELQWLQMVYNESDVSETKRKRDMDQDVGPRGRFLRRRGEAVENVCRMVCSIDEDDGVGA
VTWLWGNNTAASISLFLD*
>NCU07134.1
MTMSTWPWGLLVTTGYLINLTFAAYPLTTDANCHCYKTNATSTNYFRQHKFFDFRNLQQY
ANTPPNPISTFEGNAAAPPTSSYFDSNQWKDNWGIQTWNNTELMRLNNTDVNDATVPMVN
SFNNIYIERSSDRNANGQTYLVMRTVRHSLDGPSTPYGSNSSSGFQSSAEFESKLTSYQF
LSLRMLARTRGSPGAVTAMFTYRPPPQPQQLALVQEADLEIRTQDPSNFVQYTNQPAWNS
TSDIKEATRNASMPSGRKWSDWAFYRMDWTPGQSTWFVNGVEACKINFQAPRDPSQVMFN
VWSDGGSWSGIMGQGGKAEMQVQWIEMVYNSTEASGPVTGPWGDKKGCVNICSIDETTQL
GTPVLISNPNGE*
>FG00184.1
MPSKSILASLLAVAPLALAQNPSCDCYMTEGQYYKDHSFFDFRSLSQHAGVPALIDSIQG
NAEAGFTSDFFNWESDFSKTWGAQKWNNGNEEFPMQNTYNNLYIEENNDESPASDTWLTM
RTARHNGFQTASEFESMVKHQYVSLRMYARTKGSPGACTAMFTYRAGEDLASVQEADIEV
LTKDPPNVIHYTNQPSYTSEGGEVDGAHLGAALPDGITWSTWAKHTLDWTPDTTIWRVND
KELWRNSFQVPRDPATLSFNAWSNGDTWTGTIPEGGAAYQQIQWIELLSGRTDGATCSSV
CSVDKGEPGKAVQV*
>AN6819.1
MSSLLKISIILAGMAAAQNTECGCYSTDGVTAATYTNRIYHDFRSLDETGTLYTGEPANV
TNDDESAAAPVQSGYLTSDGFVNDFGIQTWGSPASEDTPLRKQYSNANVYIEHDGSSSST
HLTLRSYRNADFVSTAEIDSKLQNIFHASITVRARVRGAAGACAGIFTYLDDKTESDIEI
LTRDPTNHIRYTNQPGLDSDGNEIPGASTDAVLPNGAVWTDWVDHRLDWTPELSAFYANG
ELVETKTYGIPDAPSSFIVNLWGDGGSWSGTPDIESAAYLDIQWIEVLFNTSDASA*

>NCU03310.1
MSNNPYQDLLNKIQQTARSRGGGGFPGGRPPIPPRGLGPALTGFALLGGGAWVLSNSLFN
VDGGHRAIKYRRVNGVSKEIYGEGTHLMIPWFETPITYDVRAKPRNVSSLTGTKDLQMVN
ITCRVLSRPEVTALPQIYRTLGTDYDERVLPSIVNEVLKSVVAQFNASQLITQREMVAKL
VRENLAKRAARFNILLDDVSLTHLAFSPEFTAAVEAKQVAQQEAQRAAFIVDKARQEKQA
MVVKAQGEARSAELIGEAIKKSKSYVELKKLENARAIANIIQEAGGKNRLLLDSEGLGLN
VFEDRRSKDN*
>AN6073.1
MSRIPKDQWERLQLILQSRNRFNGFRPGGGGGGGIGASAALIVLGLGGWALSNSLFNVDG
GHRAIKYSRFGGVKKEIYSEGTHFAIPLIETPIIYDVRAKPRNIASLTGTKDLQMVNITC
RVLSRPRVDALPQIYRTLGQDFDERVLPSIVNEVLKSVVAQFNASQLITQRENVARLVRD
NLARRAARFNIALDDVSLTHLTFSPEFTAAVEAKQVAQQEAQRAAFLVDKARQEKQAFIV
RAQGEARSAELIGDAIKKSKSYIELRRIENARHIAQIIQENGGRNKLYLDSQGLGLNVNA
GADGESK*
>FG01119.1
MSNNNWQEEAMRRLRQMQQARGGGGGGPQMPRAAGGALAGGLLLAGGALFLSNSLFNVDG
GQRAIKYQRLTGVSKEIYNEGTHINIPWFETPIVYDVRAKPRNVASLTGTKDLQMVNITC
RVLSRPQIDALPQIYRTLGTDYDERVLPSIVNEVLKSVVAQFNASQLITQRENVARLVRE
NLARRAARFNILLDDVSLTHLAFSPEFTAAVEAKQVAQQEAQRAAFVVDKARQEKQAMVV
KAQGEARSAELIGEAIKKNKAYLELKKIENARLIAAQLQEAGSKNRLMLDSEGLGLNVFD
KNDKS*
>MG09886.1
MSQNPFQEYMRQLQRIQQRAGGGGGMPPRGTIGLGAAAALGAAGIWVVSNSLFNVDGGHR
AIKYRRISGVSKEIFGEGTHFAIPWFETPIVYDVRAKPRNVSSLTGTKDLQMVNITCRVL
SRPEVKALPQIYRTLGSDYDERVLPSIVNEVLKSVVAQFNASQLITQRENVARLIRENLS
RRAALFNIVLDDVSLTHLAFSPEFTAAVEAKQVAQQEAQRAAFVVDKARQEKQAMVVKAQ
GEARSAELIGEAIKKSKSYVELKKLENARAIAQTLQEAGGRNRLLLDAEGLGLNVFEKTD
RKD*
>AN0686.1
MAANGLYNLQRLAIPIGLGAMAVNASLYDVKGGTRAVIFDRLSGVQEQVVNEGTHFLIPW
LQKAVIYDVRTKPRNISTTTGSKDLQMVSLTLRVLHRPEVPKLPAIYQSYGTDYDERVLP
SIGNEVLKAIVAQFDAAELITQREAVSNRIRTDLMKRASQFNIALEDVSITHMTFGKEFT
RAVEQKQIAQQDAERARFIVEKAEQERQANVIRAEGEAESADIISKAVAKAGNGLIEIRR
IEASKDIAHTLASNPNVTYLPGGEGKDGGKSTSLLLGLRS*
>FG10306.1
MAGAARALGFMYRMAVPASAAVFLGSQALYDVKGGTRAVIFDRLSGVKEEVINEGTHFLI
PWLQKSIIFDVRTKPRNIATTTGSKDLQMVSLTLRVLHRPNVKALPKIYQNLGADYDERV
LPSIGNEVLKAIVAQFDAAELITQREAVSDRIRNDLTLRAAEFNIALEDVSITHMTFGRE
FTKAVEQKQIAQQDAERARFIVERAEQERQANVIRAEGESESAEAISKAIQKAGDGLIQI
RKIEASREIAATLSSNPNVAYLPGGSGKQGGQYLLSVGRA*
>MG06004.1
MRHTMAASQFRLPFLAGAGALAFATAQASLYDVKGGTRAVIFDRLSGVKDTVVNEGTHFL
IPWLHRAIIFDVRTKPRMIATTTGSKDLQMVSLTLRVLHRPEVKALPKIYQNLGTDYDER
VLPSIGNEVLKSIVAQFDAAELITQREAVSQRIRTDLMKRASEFNIALEDVSITHMTFGK
EFTKAVEQKQIAQQDAERARFIVEKAEQERQANVIRAEGEAESAETISRAIAKSGDGLVQ
IRKIEASREIAQTLASNPNVAYLPGGKQGTNILLNAGRA*
>NCU08946.1
MAARGLDMITKFAIPATVGVALLQNSIYDVRGGSRAVIFDRVAGVKDTVVNEGTHFLIPW
LQKAIIFDVRTKPRIIPTTTGSKDLQMVSLTLRVLHRPEVQALPKIYQNLGPDYDERVLP
SIGNEVLKSIVAQFDAAELITQREAVSQRIRADLVKRAAEFNIALEDVSITHMTFGKEFT
KAVEQKQIAQQDAERARFIVERAEQERQANVIRAEGEAESAETISKSIAKAGDGLIQIRK
IEASREIAQVLAANPNVAYLPGGGKGTNLLMNVGRA*

Familia 279

>NCU01634.1
MTRISSSGGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETR
GVLKTFLEGVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG*
>AN2426.1
MSGRGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLK
SFLESVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG*
>FG05491.1
MTGRGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLK
TFLEGVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG*
>MG01160.1
MTGRGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLK
SFLEGVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG*
>MG06293.1
MTGRGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLK
SFLEGVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG*
>NCU00212.1
MTGRGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLK
TFLEGVIRDAVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG*
>AN0734.1
MSGRGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLKTFLEGVIRDA
VTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG*
>FG04289.1
MRDNIQGITKPAIRRLARRGGVKRISAMIYEETRGVLKTFLEGVIRDAVTYTEHAKRKTV
TSLDVVYALKRQGRTLYGFGG*

ANEXOS

ANEXO 1 - Caracteres IUPAC para secuencias

Tras la línea de cabecera y los comentarios, una o más líneas pueden seguir para describir la secuencia: cada línea de una secuencia debería tener algo menos de 80 caracteres. Las secuencia pueden corresponder a secuencias de proteínas (estructura primaria de las proteínas) o de ácidos nucléicos, y pueden contener huecos (o *gaps*) o caracteres de alineamiento. Normalmente se espera que las secuencias se representen en los códigos estándar IUB/IUPAC para aminoácidos y ácidos nucléicos, con las siguientes excepciones: se aceptan letras minúsculas y se mapean a mayúsculas; un único guión o raya puede usarse para representar un hueco; y en secuencias de aminoácidos, 'U' y '*' son caracteres aceptables (ver más abajo). No se admiten dígitos numéricos, pero se utilizan en algunas bases de datos para indicar la posición en la secuencia.

Los códigos de ácidos nucléicos soportados son:

Código de ácido nucléico	**Significado**
A	**A**denosina
C	**C**itosina
G	**G**uanina
T	**T**imidina
U	**U**racilo
R	G A (pu**R**ina)
Y	T C (pirimidina/p**Y**rimidine)
K	G T (cetona/**K**etone)
M	A C (grupo a**M**ino)
S	G C (interacción fuerte/**S**trong interaction)
W	A T (interacción débil/**W**eak interaction)
B	G T C (no A) (**B** viene tras la A)
D	G A T (no C) (**D** viene tras la C)
H	A C T (no G) (**H** viene tras la G)
V	G C A (no T, no U) (**V** viene tras la U)
N	A G C T (cualquiera/a**N**y)
X	máscara
-	hueco (gap) de longitud indeterminada

Los códigos de aminoácidos soportados son:

Código de aminoácido	Significado
A	Alanina
B	Ácido aspártico o Asparagina
C	Cisteína
D	Ácido aspártico
E	Ácido glutámico
F	Fenilalanina
G	Glicina
H	Histidina
I	Isoleucina
K	Lisina
L	Leucina
M	Metionina
N	Asparagina
O	Pirrolisina
P	Prolina
Q	Glutamina
R	Arginina
S	Serina
T	Treonina
U	Selenocisteína
V	Valina
W	Triptófano
Y	Tirosina
Z	Ácido glutámico o Glutamina
X	cualquiera
*	parada de traducción
-	hueco (*gap*) de longitud indeterminada

ANEXO 2 - Contenido de bases de datos disponibles en BLAST

**** Nucleotide Sequence Databases***

nr: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant".

est: Database of GenBank+EMBL+DDBJ sequences from EST division.

est_human: Human subset of GenBank+EMBL+DDBJ sequences from EST division.

est_mouse: Mouse subset of GenBank+EMBL+DDBJ sequences from EST division.

est_others: Non-Mouse, non-Human sequences of GenBank+EMBL+DDBJ sequences from EST Division.

gss: Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.

htgs: Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished phase 3 HTG sequences are in nr.

pat: Nucleotides from the Patent division of GenBank.

yeast: Saccharomyces cerevisiae genomic nucleotide sequences.

mito: Database of mitochondrial sequences.

vector: Vector subset of GenBank(R), NCBI, in ftp://ftp.ncbi.nlm.nih.gov/blast/db/. **ecoli:** Escherichia coli genomic nucleotide sequences.

pdb: Sequences derived from the 3-dimensional structures from the Brookhaven Protein Data Bank.

drosophila genome: Drosophila genome provided by Celera and Berkeley Drosophila Genome Project (BDGP).

month: All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.

alu: Select Alu repeats from REPBASE, suitable for masking Alu repeats from query equences. It is available by FTP from ftp://ftp.ncbi.nlm.nih.gov/blast/db/alu.n.Z. See "Alu alert" by Claverie and Makalowski, Nature 371: 752 (1994).

dbsts: Database of GenBank+EMBL+DDBJ sequences from STS division.

chromosome: Searches Complete Genomes, Complete Chromosome, or contigs form the NCBI Reference Sequence project.

wgs_anopheles: Anopheles gambiae (mosquito) whole genome shotgun sequences.

**** Peptide sequence database***

nr: All non-redundant GenBank CDS translations + PDB + SwissProt + PIR+PRF.

swissprot: Last major release of the SWISS-PROT protein sequence database (no incremental updates).

pat: Proteins from the Patent division of GenBank**.**

yeast: Saccharomyces cerevisiae genomic CDS translations.

e.coli: Escherichia coli genomic CDS translations

pdb: Sequences derived from the 3-dimensional structures from the Brookhaven Protein Data Bank

drosophila genome: Drosophila genome proteins provided by Celera and Berkeley Drosophila Genome Project (BDGP).

month: All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days

Bibliografía consultada y otros recursos

- A Primer on Molecular Evolution and Phylogenetics: http://bioinfo.cipf.es/fransua/Courses/doku.php?id=start
- Alex Sánchez - IntroducciÓn a la BioinformÁtica (http://www.ub.edu/stat/docencia/Biologia/introbioinformatica/#inicio).
- Bioinformatics (Second Edition).2002. A. D. Baxevanis and B. F. Ouellette, eds.
- Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2001, Second Edition Andreas D. Baxevanis, B.F. Francis Ouellette
- Bookshelf – NCBI: http://www.ncbi.nlm.nih.gov/sites/entrez?db=books
- Claverie, Jean-Michel, y Cedric Notredame. 2007. *Bioinformatics for dummies*. For Dummies.
- Current Protocols in Bioinformatics (2003) 6.1.1-6.1.13. Introduction to Inferring Evolutionary Relationships.
- Current Protocols in Bioinformatics. 2006
- Eddy, 2004 Nature Biotechnology. What is a Hidden Markov Model?
- Eduardo Rodriguez-Tello. Curso de Bioinformática. http://www.tamps.cinvestav.mx/~ertello/bioinfo.php
- EMBL – EBI: http://www.ebi.ac.uk/
- Emboss: http://emboss.sourceforge.net/
- ExPASy Proteomics Server: http://www.expasy.org/links.html
- Glossary of Bioinformatics TermsGlossary of Bioinformatics Terms: http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/genejargon.shtml
- Joaquín Cañizares Sales – Bioinformática I http://personales.upv.es/jcanizar/bioinformatica/
- Korf, Ian, Mark Yandell, y Joseph Bedell. 2003. *BLAST*. O'Reilly Media, Inc.
- Lesk, AM., “Bioinformatics”, Primera edición, Oxford University Press, 2002.
- Lesk, Arthur M. 2002. *Introduction to bioinformatics*. Oxford University Press.
- Lope Andrés Flórez Weidinger - http://bioinformate.uniandes.edu.co/.
- NCBI YouTube Channel: http://www.youtube.com/user/NCBINLM
- NIH Biomedical Information Science and Technology Initiative Consortium: http://www.bisti.nih.gov/
- Polanski, Andrzej, y Marek Kimmel. 2007. *Bioinformatics*. 1º ed. Springer.
- Rodrigo Lopez - 2Can Support Portal – Bioinformatics (http://www.ebi.ac.uk/2can/home.html)
- Wikipedia: http://es.wikipedia.org/wiki/Wikipedia:Portada