Ramón Tamarit Agusti

GENOMICA COMPUTACIONAL

PEC1 — Primera Prueba de evaluación continúa

INDICE

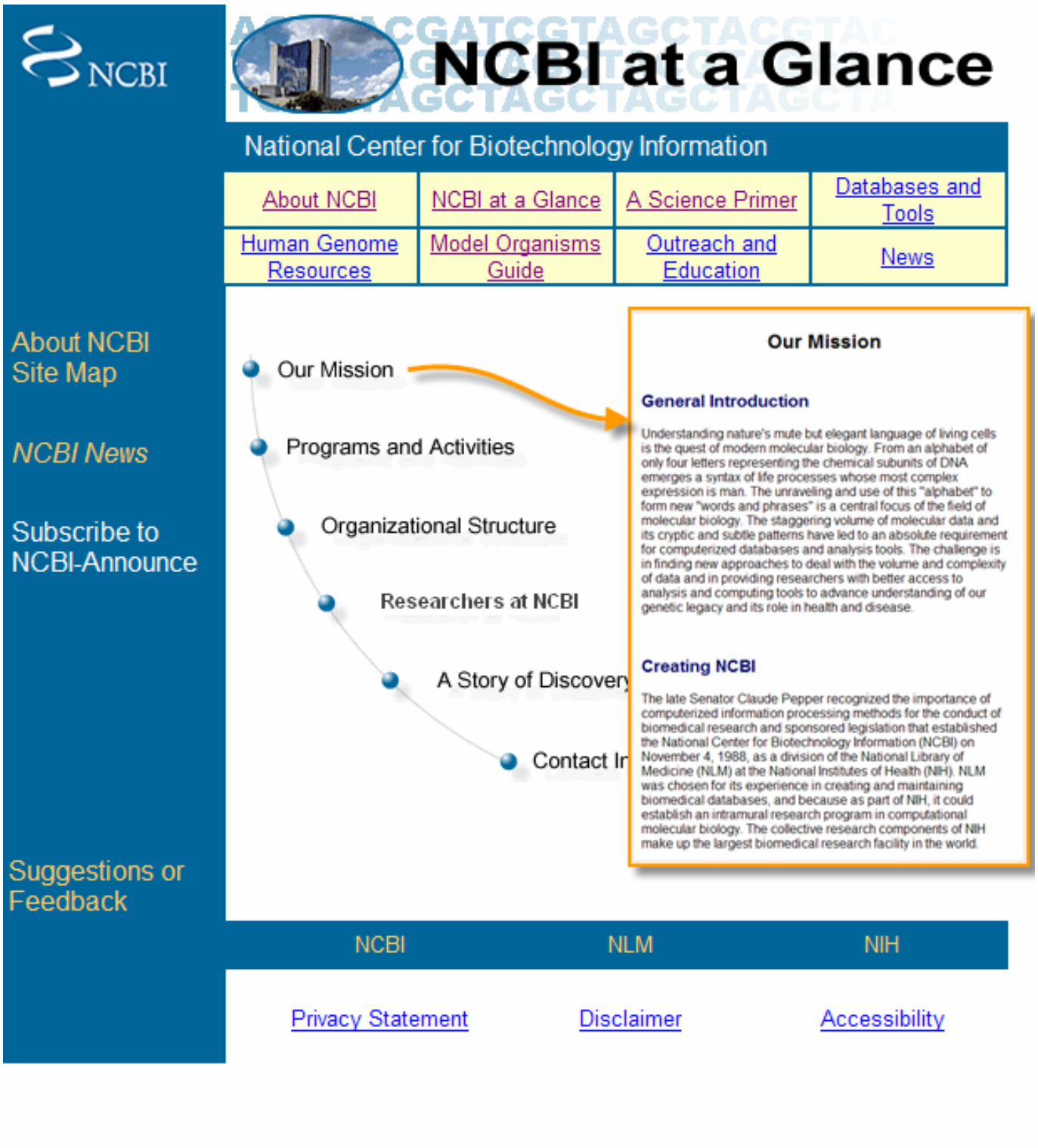
1 PARTE I

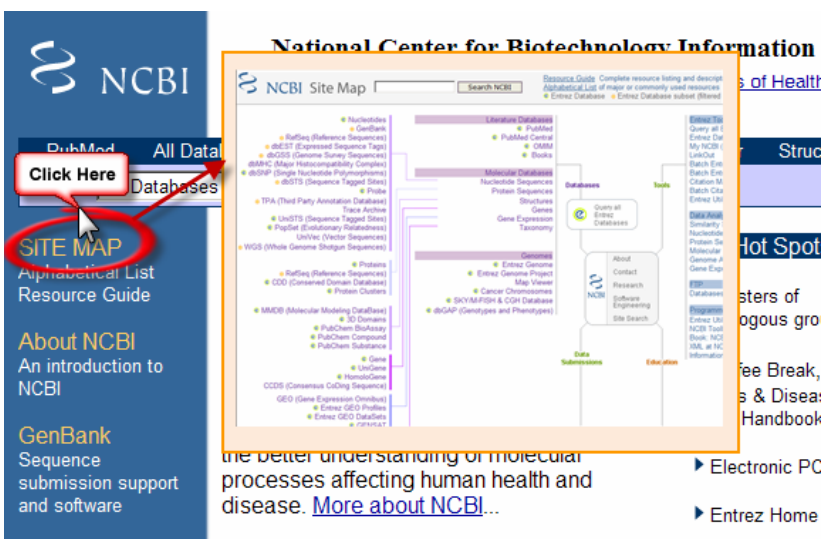
1.1 EJERCICIO 1

1.1.1 Accede a la base de datos de secuencias NCBI:

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

Este fue uno de los primeros portales WEB pioneros en la recolección de secuencias biológicas. Aquí se pueden encontrar informaciones cruzadas con la mayoría de bases de datos existentes en genómica y proteómica.





Una opción interesante es que se puede acceder a un mapa completo e intuitivo de todos los recursos disponibles en el sitio desde “SITE MAP”.

El mapa del sitio es totalmente muy ilustrativo de todo lo que hace el NCBI, y de cuales son las relaciones entre todas las bases de datos y recursos disponibles.

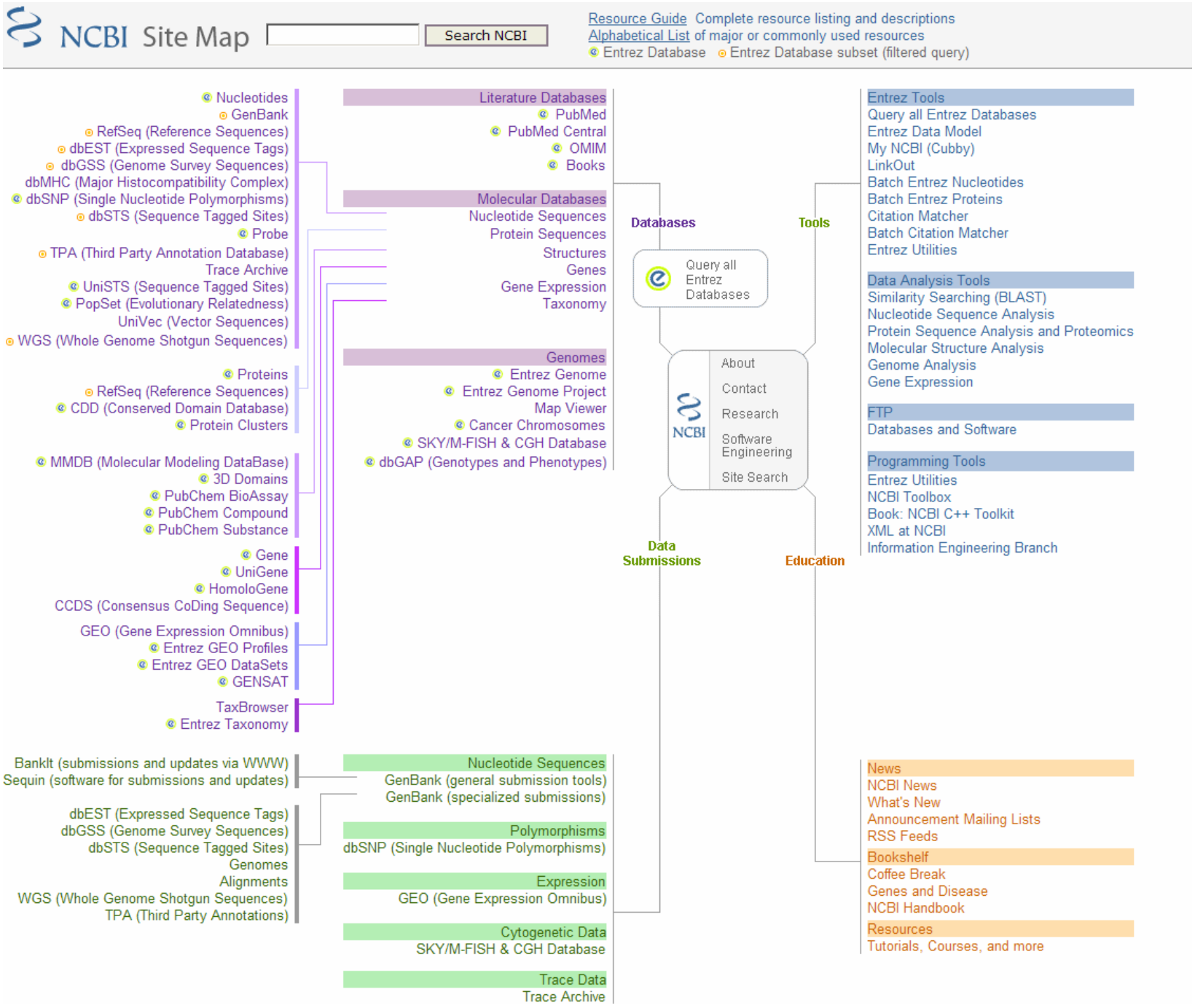


Imagen del mapa completo de los recursos disponibles, agrupados por:

- Herramientas
- Sistemas de consulta / Bases de datos primarias
- Envió de datos
- Literatura
- Educación

1.2 Intentad encontrar información del proyecto REFSEQ (Reference Sequence Project).

Debéis encontrar un enlace (link) a una página dedicada a REFSEQ. Es importante recordar que la mayoría de bancos de datos biológicos, debido a la inherente ambigüedad de la informacion que contienen, son muy redundantes y a veces incoherentes. Explicad la misión de REFSEQ en este contexto.

Descripción detallada de RefSeq:
http://nar.oxfordjournals.org/cgi/content/full/33/suppl_1/D501?ijkey=06NkMzN3kUcez&keytype=ref
Kim D. Pruitt , Tatiana Tatusova , and Donna R. Maglott
NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins
Nucl. Acids Res. 33: D501-D504.

El link directo hacia RefSeq desde el Mapa del Sitio (de ahí la utilidad del mapa)

http://www.unav.es/genetica/bioinfo/RefSeqesp.html
RefSeq pretende ser una colección integrada, completa y no redundante de secuencias de DNA genómico, RNA y proteínas, para los principales organismos. El NCBI genera RefSeqs para más de 1100 virus y 150 bacterias además de organismos superiores (humano, ratón, rata, zebrafish, etc).

Las **principales características** de la colección RefSeq son:
1. No-redundancia
2. Conexión entre secuencias nucleotídicas y sus correspondientes secuencias proteicas
3. Formato constante
4. Datos validados
5. Números de acceso específicos
6. "Curación" (mantenimiento) a cargo del personal del NCBI y colaboradores externos

Los números de acceso de RefSeq tienen el formato: **XX_123456** (dos letras, guión bajo y 6 números). Las dos letras iniciales indican el tipo de secuencia:

Secuencias **revisadas manualmente**:
```
NC 123456  Moléculas genómicas completas (genomas, cromosomas, plásmidos)
NG 123456  Región genómica incompleta
NM 123456  RNA mensajero
NR 123456  RNA no codificante (ribosomal, pseudogenes, etc)
NP 123456  Proteína (en futuras versiones se añadirán dos números más)
```

Secuencias **no revisadas**:
```
NT 123456  Secuencias de clones usados para ensamblar genomas (BACs)
NW 123456  Ensamblajes parciales de genomas (shotgun)
XM 123456  RNA mensajero deducido de las anotaciones del DNA genómico
XR 123456  RNA no codificante deducido de las anotaciones
XP 123456  Proteínas dedicidas de las anotaciones del DNA genómico
```

También hay registros que contienen todas las secuencias de un proyecto de secuenciación de un genoma completo, con el formato **NZ_ABCD12345678**, en el que ABCD son las letras que identifican el proyecto. ZP_12345678 se refiere a las proteínas deducidas de los registros del proyecto.

http://www.unav.es/genetica/bioinfo/buildprocess.html
Explicación de los pasos
1. Recogida de datos desde el propio NCBI y de colaboradores externos, produciendo un registro con nombre y símbolo oficiales, nombres alternativos, enlaces a bases de datos de SNPs, asociación con enfermedades (OMIM), localización citogenética, localización genómica, tipo de ontología (GO) y localización intracelular, números de acceso de las secuencias, marcadores STS y clusters de UniGene.

2. Se seleccionan manualmente las secuencias representativas de esta región, y a partir de ellas se genera la secuencia "semilla" que se usa para producir el registro inicial de RefSeq.

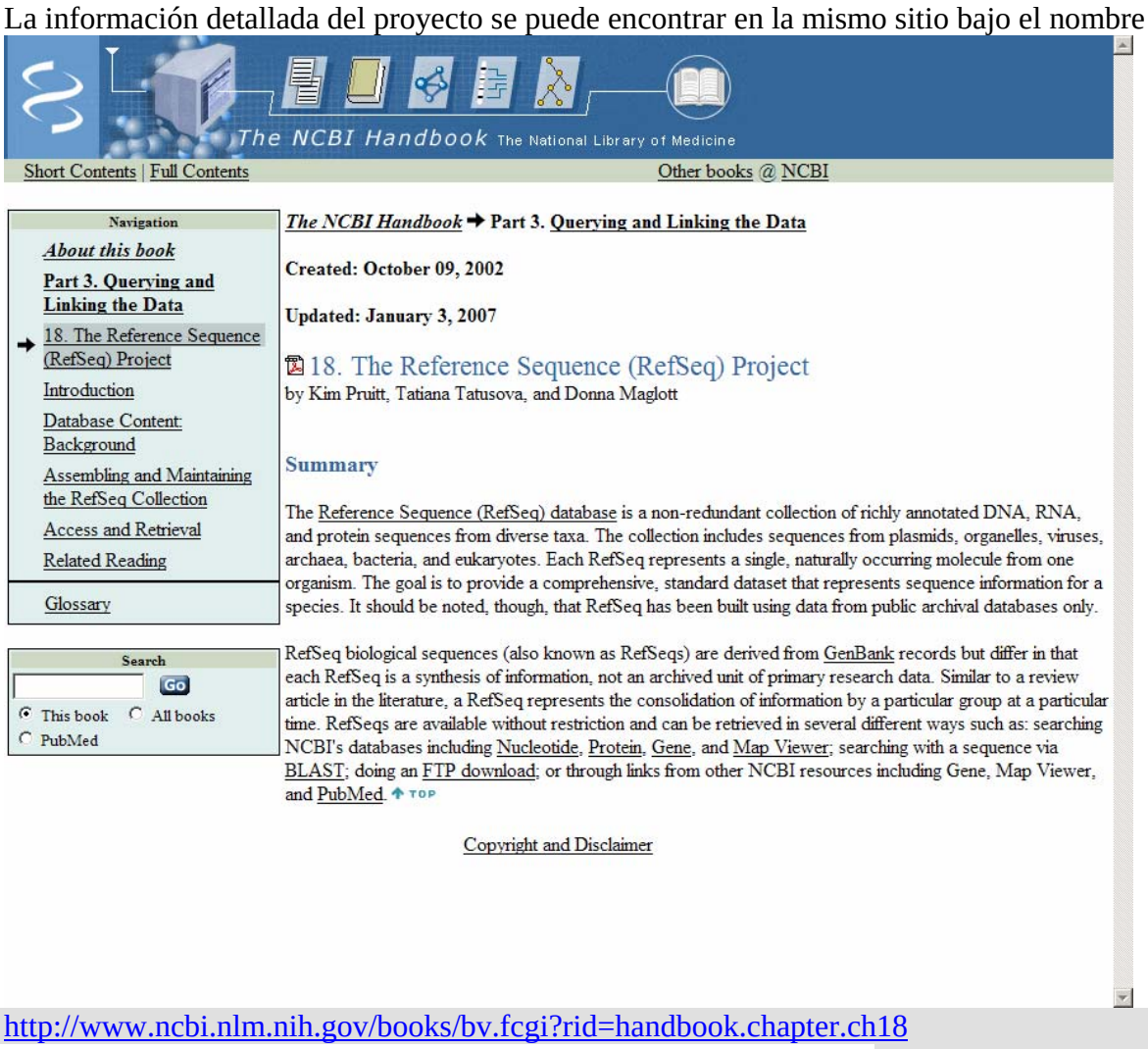
3. La secuencia semilla se usa como query en una búsqueda BLAST. Para los RNA mensajeros, se selecciona el resultado que identifica el alineamiento más largo sin huecos y con el mínimo número de desemparejamientos.

4. Las secuencias seleccionadas en el paso 3 que contienen una región codificante completa dan lugar a los registros de mRNA y de proteínas marcados como PROVISIONAL o PREDICTED (XM_ ó XP_), mediante un proceso automático. Estos registros son públicos y contienen anotaciones como referencias bibliográficas, nombres alternativos, número de identificaciónd e LocusLink, número de OMIM, localización física, nombre y símbolo oficiales.

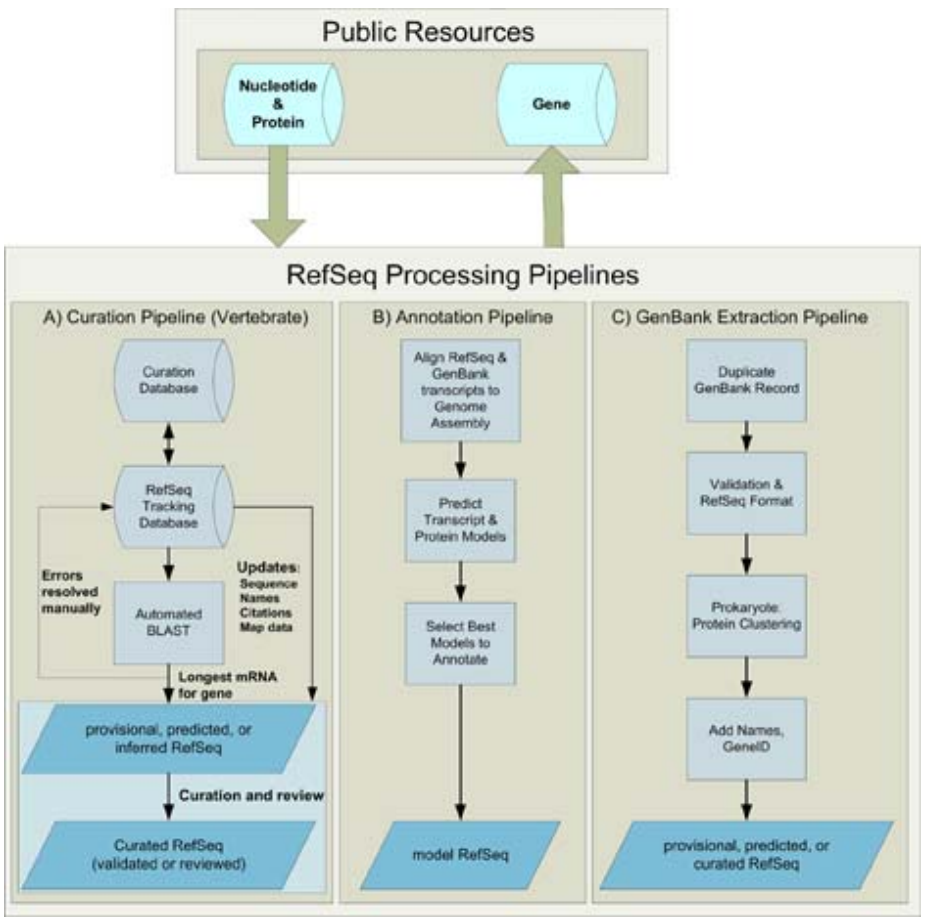
5. Las secuencias que no entran en el paso anterior (por ejemplo, por contener regiones codificantes incompletas) son revisadas manualmente para comprobar si se puede obtener esa información combinando varias secuencias incompletas de diferentes registros.

6. Los registros marcados como REVIEWED son el producto final. En principio, todos los registros provisionales serán sometidos a revisión, generando variantes de splicing y purificando la calidad de la secuencia (eliminar errores de secuenciación, por ejemplo), añadiendo referencias bibliográficas, resumiendo las funciones, añadiendo anotaciones e indicando si el proceso es definitivo. Estos son los mensajeros y proteínas con identificadores NM_ y NP_, respectivamente.

La información detallada del proyecto se puede encontrar en la mismo sitio bajo el nombre de NCBI Handbook.



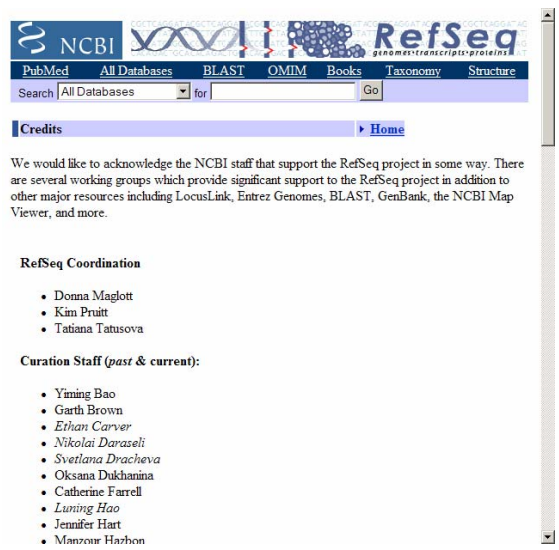
http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch18

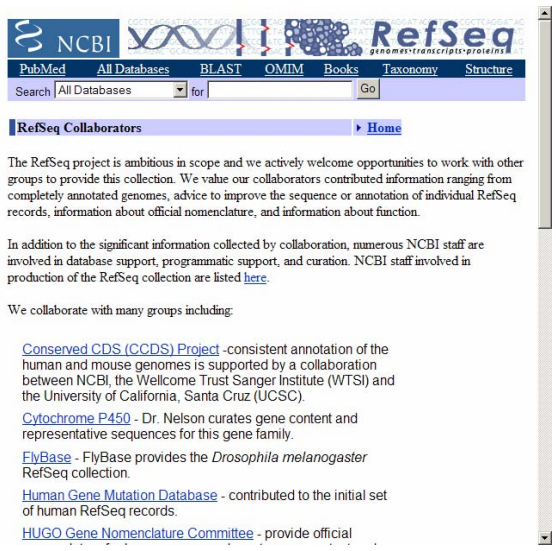


The NCBI Handbook ➜**Part 3. Querying and Linking the Data** ➜18. The Reference Sequence (RefSeq) Project

Figure 2.RefSeq Processing Pipelines. Once sequence data are deposited in the public archival databases, it is available for RefSeq processing. Processing pipelines include the vertebrate curation pipeline, the computational genome annotation pipeline, and extraction from GenBank. These pipelines generate new and updated RefSeq records that become publicly available in Entrez Nucleotide, Protein, and Gene databases. (**A**) Once a gene is defined and associated with sufficient sequence information in an internal curation database, it can be pushed into the RefSeq pipeline. The RefSeq process is initiated by an automated BLAST step, which uses the stored sequence data as a query against GenBank to identify the longest mRNA for each locus. This initial RefSeq record has a status of provisional, predicted, or inferred. Subsequent curation may result in a sequence or annotation update (as described in Box 2) and a status of validated or reviewed. Records are updated if the underlying GenBank accession number is updated or if other associated data are updated, including nomenclature, publications, or map location. (**B**) Available RefSeq and GenBank data are aligned to an assembled genome, *ab initio* gene prediction is carried out that uses alignment data, and an analysis program integrates all available data to define the annotation models. New "model" RefSeq records are generated by this pipeline. (**C**) When a complete, annotated genome becomes available in GenBank, a set of corresponding RefSeq records are generated by duplicating the GenBank submission, followed by validation and addition of cross-references to Entrez Gene (via a dbXref citing the GeneID) and, in some cases, more informative and standardized protein names.

El staff de refseq. Y los colaboradores …. Gracias a todos ellos disponemos de toda esta preciosa información

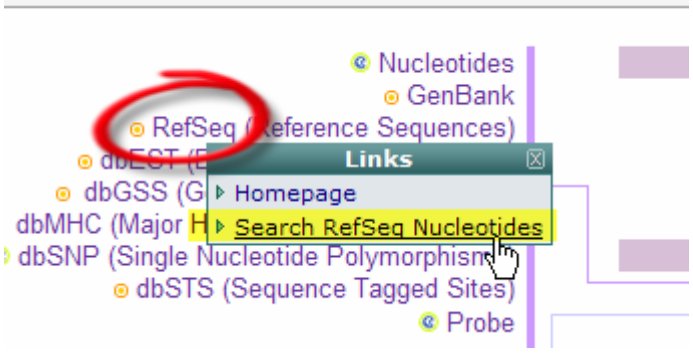




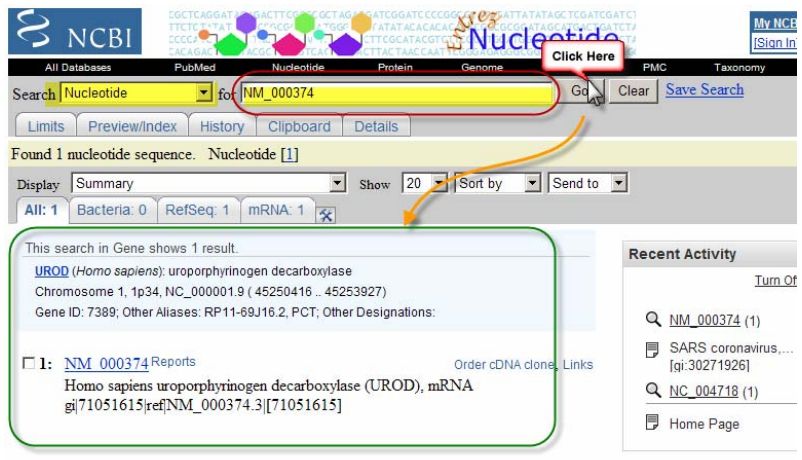
1.2.1 Seleccionad Búsquedas de secuencias de Nucleótidos

Buscaremos una secuencia concreta por su identificador (ID) en el siguiente paso. Seleccionad en el cuadro Search qué tipo de datos se desea buscar.

Otra vez desde el mapa pulsamos sobre RefSeq y elegimos Nuecleotides.



1.2.2 Introducir este identificador NM_000374

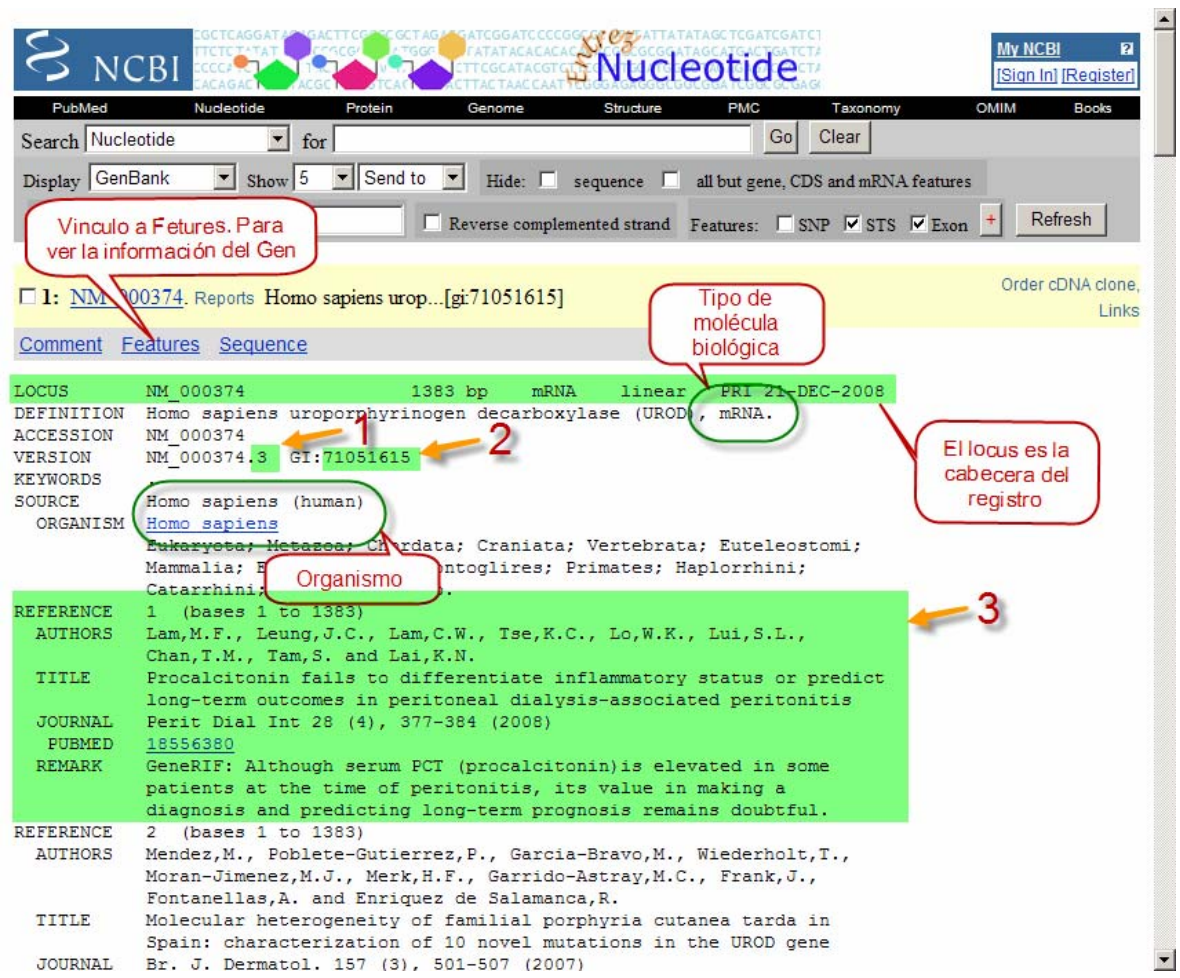


Aquí encontramos una descripción de lo que significa la cabecera de acceso.
http://www.geneinfinity.org/sp_seqformat.html
The description line (or header line) is often used to add information,
>gi|identifier|namespace|accession.version|name description *(NCBI)*
Example:
>gi|412163|emb|CAA00606.1| albumin [Homo sapiens]

1.2.3 Acceder a esta entrada. Debéis deducir el tipo de molécula biológica y especie a la que corresponde.

Echadle un vistazo a la ficha de esta secuencia. Éste es el típico formato donde se especifica a la izquierda, el nombre de cada atributo, y a su derecha, el

valor de éste. Estrictamente hablando esto es una colección de secuencias, más que una base de datos informática. Describid qué valores estáis observando (campo a campo).



Básicamente la información que contiene el registro es muy descriptiva, no obstante, en la página **http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html** se puede consultar una descripción completa de que es cada campo, aquí únicamente enumero las partes más importantes.

Campo	**Descripción**
Locus (cabecera del registro)	
Name	Un nombre único para la secuencia
Sequence length	Longitud de la secuencia
Molecule Type	ADN, genómico, mRNA, etc.
Genbank Division	División del GenBank a la que pertenece la secuencia
Modification Date	Fecha de la última modificación
Definition	Breve descripción de la secuencia
Accession	Identificador único de la entrada, no varía aunque se modifique la secuencia
Version	Número de versión de la secuencia
GI	Identificador único de la secuencia, cambia con las modificaciones
Keywords	Palabras clave que describen la secuencia
Source	Nombre del organismo
Organism	Nombre científico del organismo
Reference	Publicaciones relacionadas.

Otro de los campos más importantes es el de features:

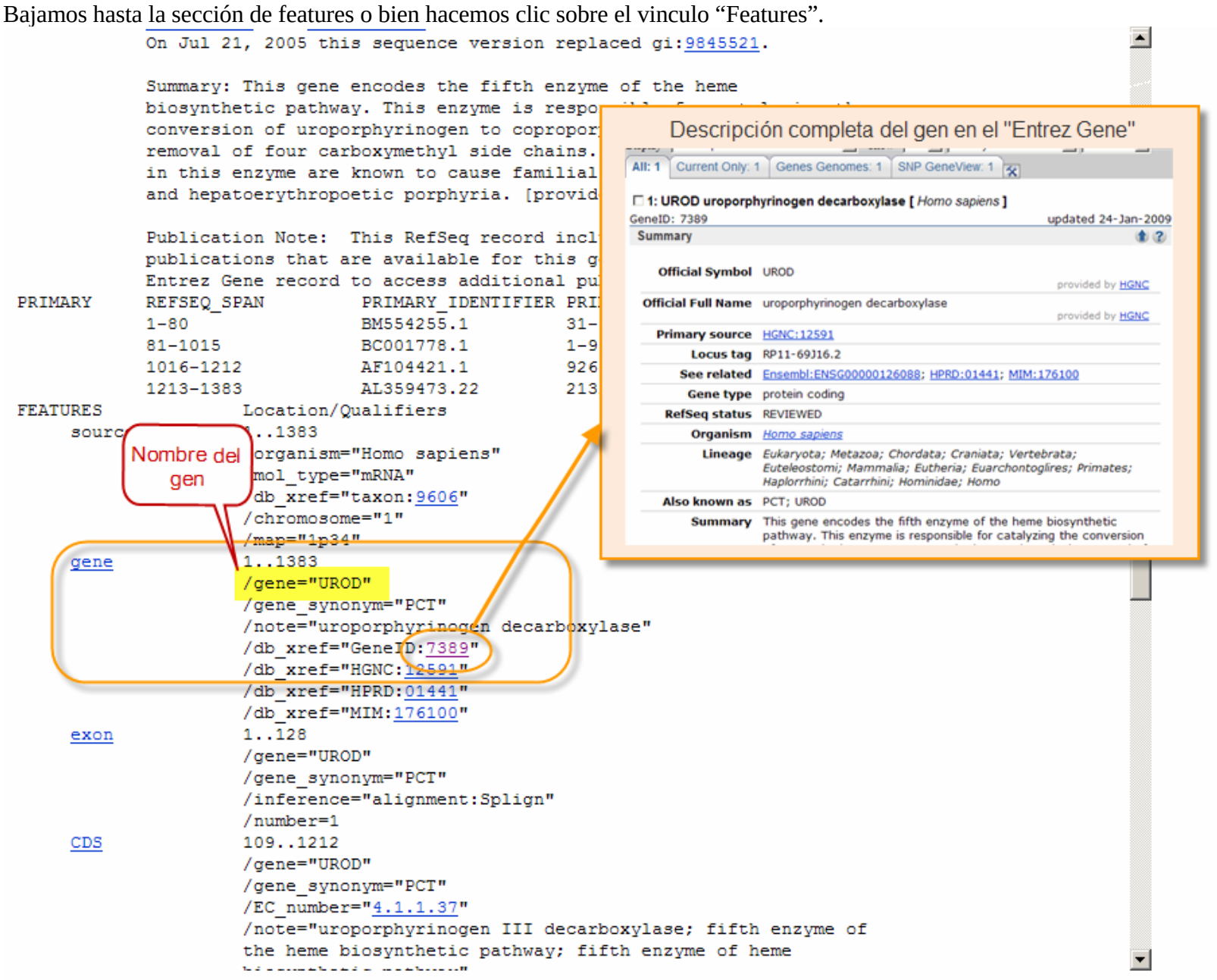
Features	Información sobre las regiones de interés	
	source	Longitud de la secuencia, nombre del organismo, taxón ID
	CDS	Secuencia codificante
	protein_id	Identificador de la secuencia protéica
	gene	Región cubierta por un gen

```
            publications that are available for this gene. Please see the
            Entrez Gene record to access additional publications.
PRIMARY     REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN       COMP
            1-80                BM554255.1         31-110
            81-1015             BC001778.1         1-935
            1016-1212           AF104421.1         926-1122
            1213-1383           AL359473.22        21321-21491
FEATURES             Location/Qualifiers
     source          1..1383
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="1"
                     /map="1p34"
     gene            1..1383
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /note="uroporphyrinogen decarboxylase"
                     /db_xref="GeneID:7389"
                     /db_xref="HGNC:12591"
                     /db_xref="HPRD:01441"
                     /db_xref="MIM:176100"
     exon            1..128
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /inference="alignment:Splign"
                     /number=1
     CDS             109..1212
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /EC_number="4.1.1.37"
                     /note="uroporphyrinogen III decarboxylase; fifth enzyme of
                     the heme biosynthetic pathway; fifth enzyme of heme
                     biosynthetic pathway"
                     /codon_start=1
                     /product="uroporphyrinogen decarboxylase"
                     /protein_id="NP_000365.3"
                     /db_xref="GI:71051616"
                     /db_xref="CCDS:CCDS518.1"
                     /db_xref="GeneID:7389"
                     /db_xref="HGNC:12591"
                     /db_xref="HPRD:01441"
                     /db_xref="MIM:176100"
                     /translation="MEANGLGPQGFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLP
                     EFRETRAAQDFFSTCRSPEACCELTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVP
                     GKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRVPLIGFAGAPW
                     LMTYMVEGGGSSTMAQAKRWLYQRPQASHQLLRILTDALVPYLVGQVVAGAQALQLFE
                     SHAGHLGPQLFNKFALPYIRDVAKQVKARLREAGLAPVPMIIFAKDGHFALEELAQAG
                     YEVVGLDWTVAPKKARECVGKTVTLQGNLDPCALYASEEEIGQLVKQMLDDFGPHRYI
                     ANLGHGLYPDMDPEHVGAFVDAVHKHSRLLRQN"
     exon            129..241
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /inference="alignment:Splign"
                     /number=6
     exon            745..882
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /inference="alignment:Splign"
                     /number=7
     exon            883..983
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /inference="alignment:Splign"
                     /number=8
     exon            984..1050
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /inference="alignment:Splign"
                     /number=9
     exon            1051..1383
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /inference="alignment:Splign"
                     /number=10
ORIGIN
        1 atggcgacgc tcttggttcc ctacagaaag gggcggagcc tggactgggg ggcaggctca
       61 gattcaggtt aaattgtgga ttgagctcgc agttacagac agctgaccat ggaagcgaat
      121 gggttgggac ctcagggttt tccggagctg aagaatgaca cattcctgcg agcagcctgg
      181 ggagaggaaa cagactacac tcccgtttgg tgcatgcgcc aggcaggccg ttacttacca
      241 gagtttaggg aaacccgggc tgcccaggac tttttcagca cgtgtcgctc tcctgaggcc
      301 tgctgtgaac tgactctgca gccactgcgt cgcttccctc tggatgctgc catcattttc
      361 tccgacatcc ttgttgtacc ccaggcactg ggcatggagg tgaccatggt acctggcaaa
      421 ggacccagct tcccagagcc attaagagaa gagcaggacc tagaacgcct acgggatcca
      481 gaagtggtag cctctgagct aggctatgtg ttccaagcca tcacccttac ccgacaacga
      541 ctggctggac gtgtgccgct gattggcttt gctggtgccc catggaccct gatgacatac
      601 atggttgagg gtggtggctc aagcaccatg gctcaggcca agcgctggct ctatcagaga
      661 cctcaggcta gtcaccagct gcttcgcatc ctcactgatg ctctggtccc atatctggta
      721 ggacaagtgg tggctggtgc ccaggcattg cagctgtttg agtcccatgc agggcatctt
      781 ggcccacagc tcttcaacaa gtttgcactg ccttacatcc gtgatgtggc caagcaagtg
      841 aaggccaggt tgcgggaggc aggcctggca ccagtgccca tgatcatctt tgctaaggat
      901 gggcattttg ccctggagga gctggcccaa gctggctatg aggtggttgg gcttgactgg
      961 acagtggccc caaagaaagc ccgggagtgt gtggggaaga cggtgacatt gcagggcaac
     1021 ctggacccct gtgccttgta tgcatctgag gaggagatcg ggcagttggt gaagcagatg
     1081 ctggatgact ttggaccaca tcgctacatt gccaacctgg gccatgggct ttatcctgac
     1141 atggacccag aacatgtggg cgcctttgtg gatgctgtgc ataaacactc acgtctgctt
     1201 cgacagaact gagtgtatac ctttaccctc aagtaccact aacacagatg attgatcgtt
     1261 tccaggacaa taaaagtttc ggagttgaac tattgtgtag ttttgtttgt gaaagattgt
     1321 gcccatatcc tcagttcttc ttagcctctg ctccttccct gggaaccctc tctatatcct
     1381 ctt
//
```

Vista de grafico (mostrando los 10 exones) del gen UROD

1.2.4 Localizad el nombre del gen.

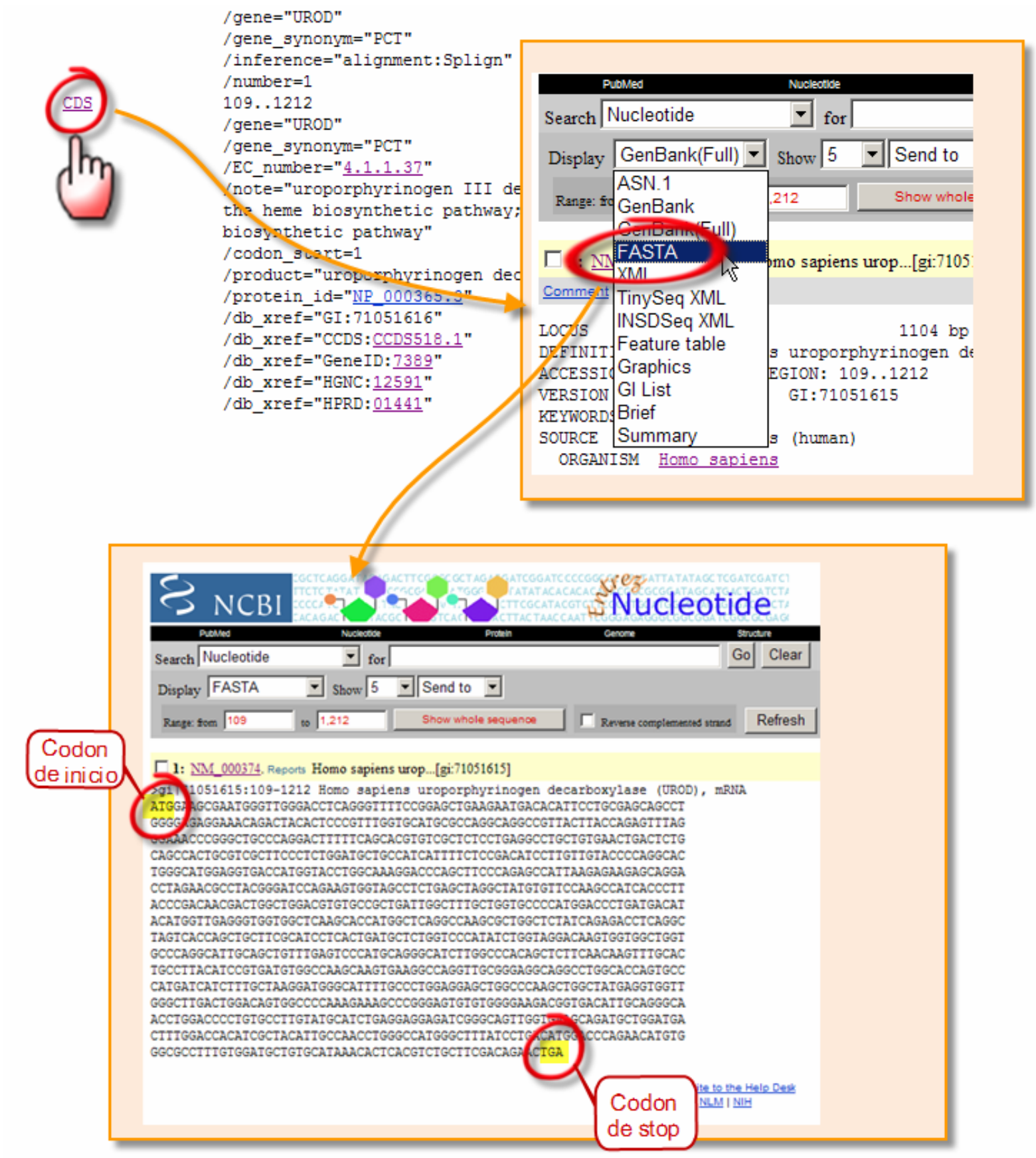
Bajamos hasta la sección de features o bien hacemos clic sobre el vinculo “Features”.



1.2.5 Ahora debéis distinguir la parte codificante para proteína del gen (buscad el codón ATG de inicio de traducción). ¿Cuál es el codón STOP en este caso?

La parte codificante se denomina CDS (Coding Sequence). Jugad un poco con la extracción de subsecuencias de la secuencia principal con el cuadro Range. Finalmente, debeis extraer una secuencia que comience por el START y STOP codons anotados en esta ficha (CDS).

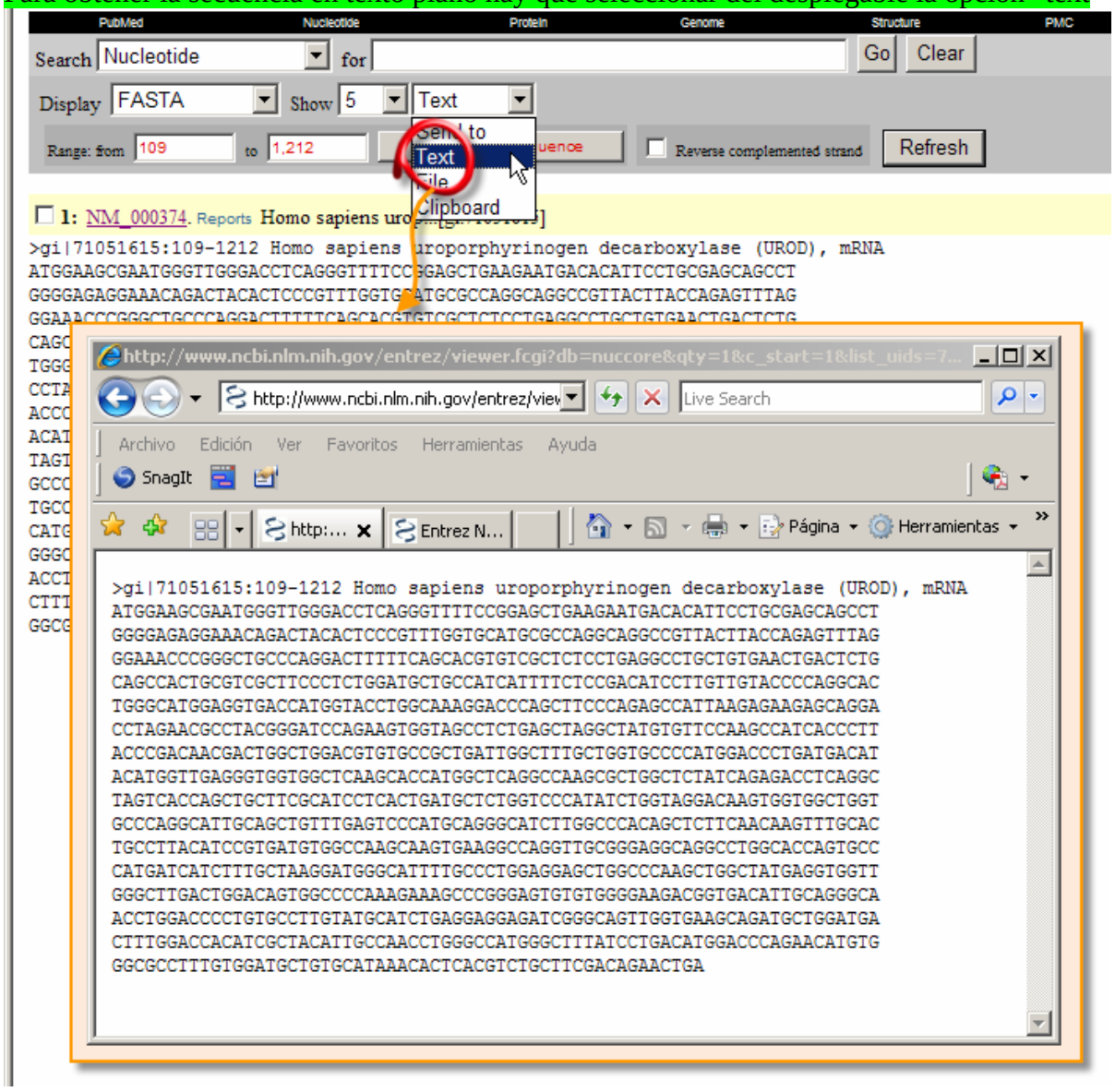
Para obtener la secuencia codificante con los codones de Start y stop pinchamos sobre CDS, y luego elegimos FASTA.



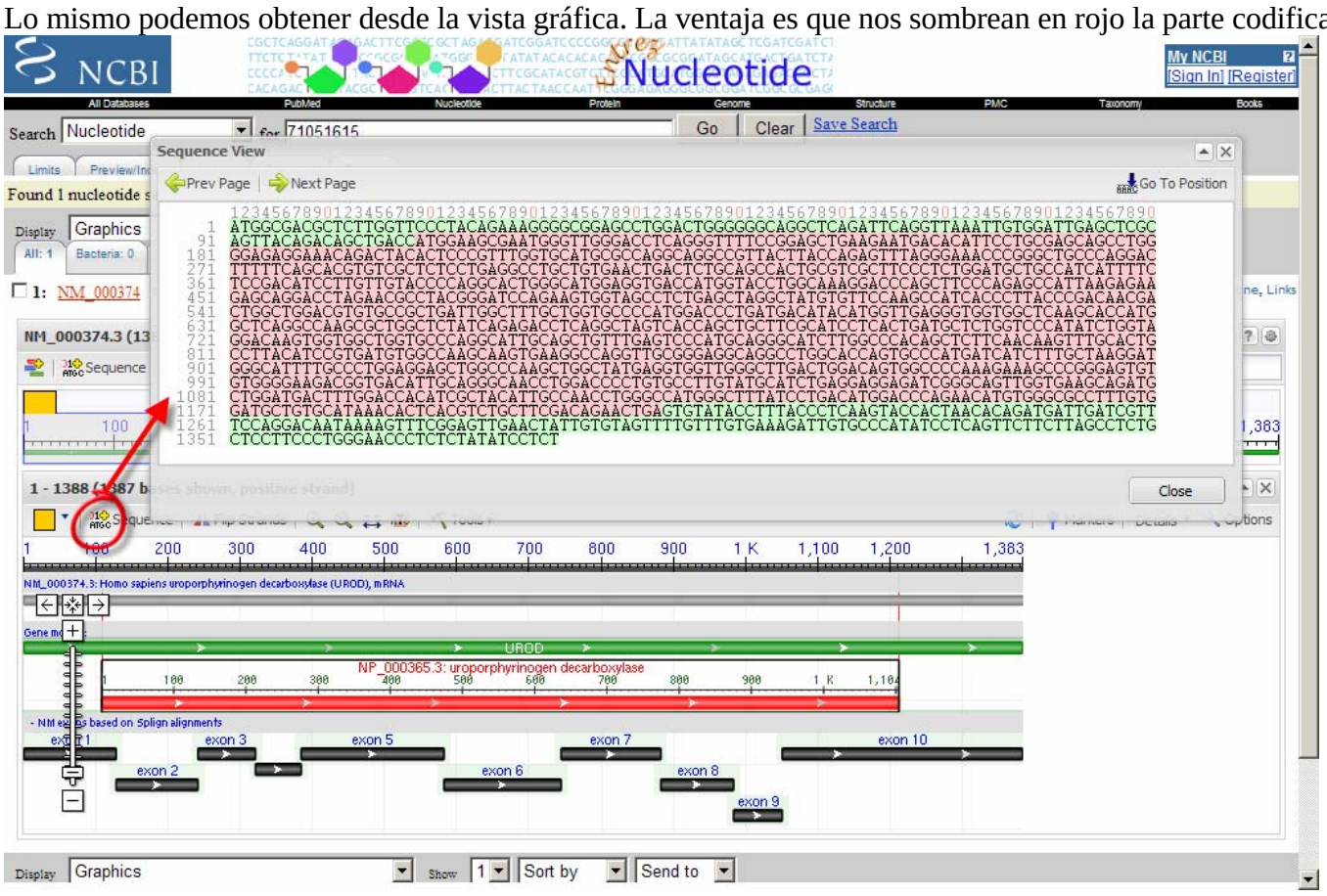
1.2.6 Obtened la secuencia en pantalla en formato FASTA y en texto plano (no HTML).

El formato FASTA es el formato clásico para representar una secuencia de nucleótidos o de aminoácidos. El formato de texto plano es el más usado por las aplicaciones bioinformáticas comunes para intercambiar datos (no usan normalmente ni HTML ni WORD).

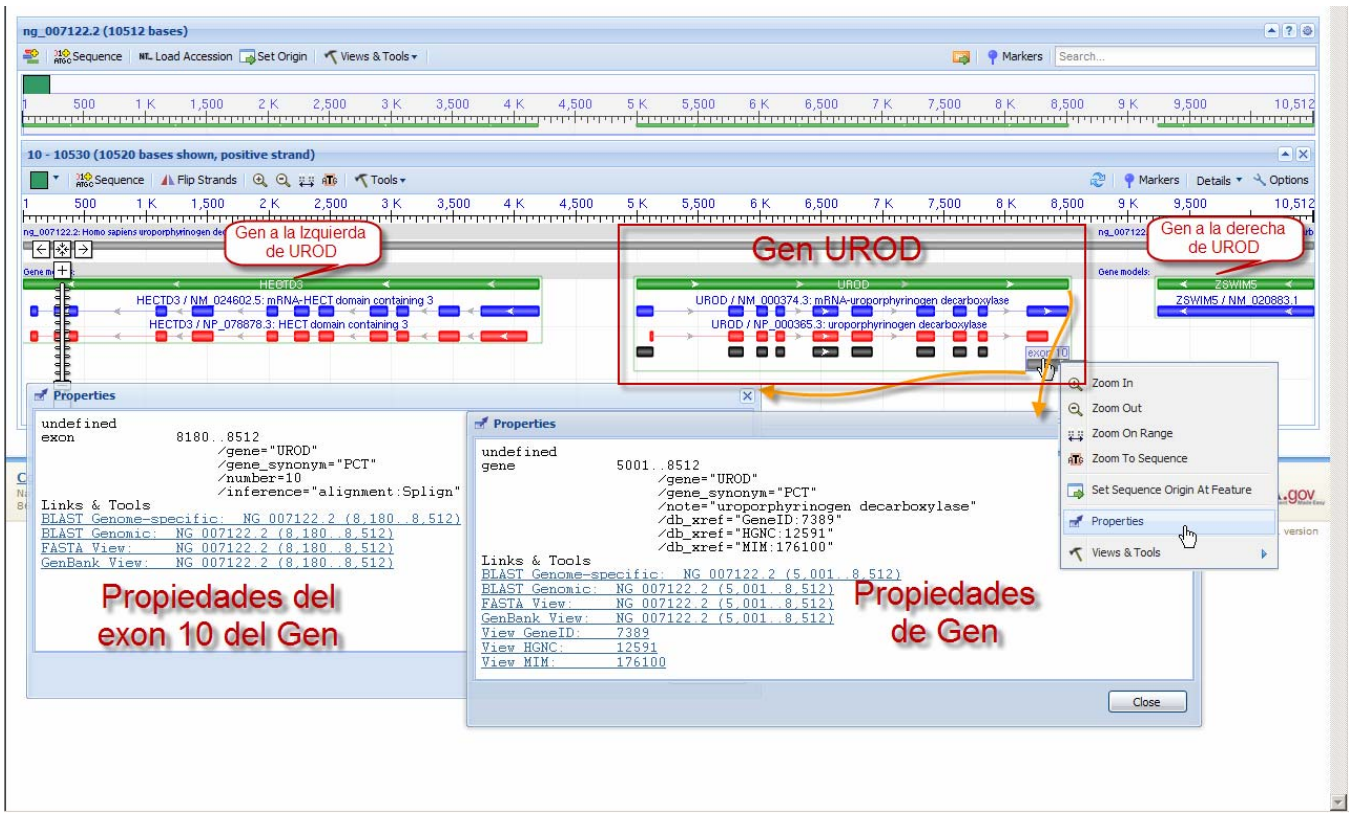
Para obtener la secuencia en texto plano hay que seleccionar del desplegable la opción “text”



Lo mismo podemos obtener desde la vista gráfica. La ventaja es que nos sombrean en rojo la parte codificante:



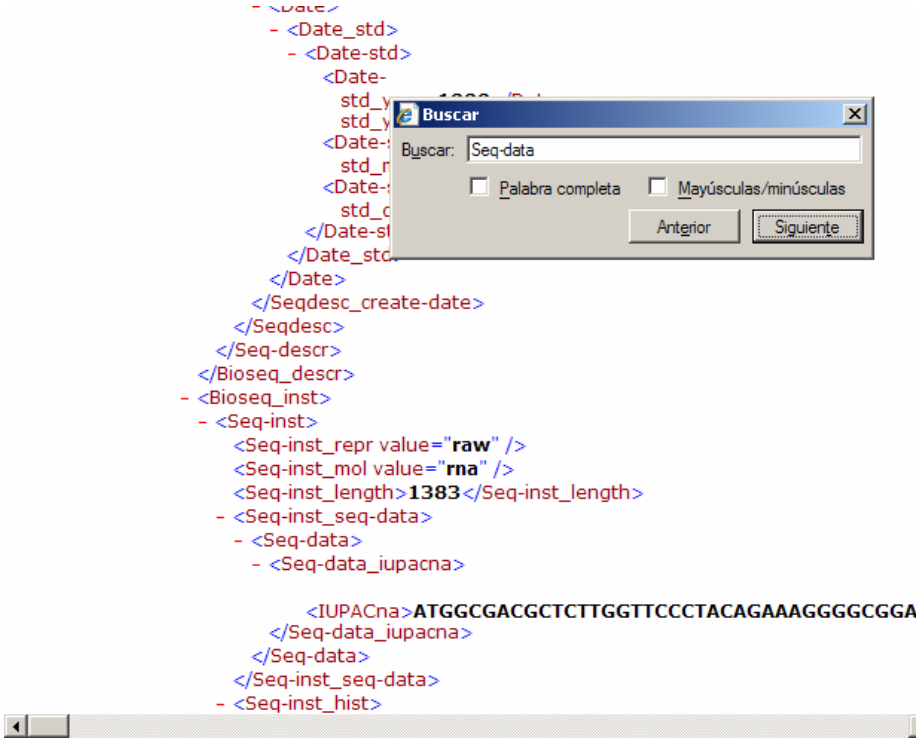
Otra opción interesante de la vista grafica es el menú “properties”, desde donde se puede acceder a todas de las propiedades del gen, así como hacer zoom sobre el mismo.



Igualmente podemos poner marcas que luego se visualizan en la vista de texto y nos ayudan a “situarnos” dentro de la misma.

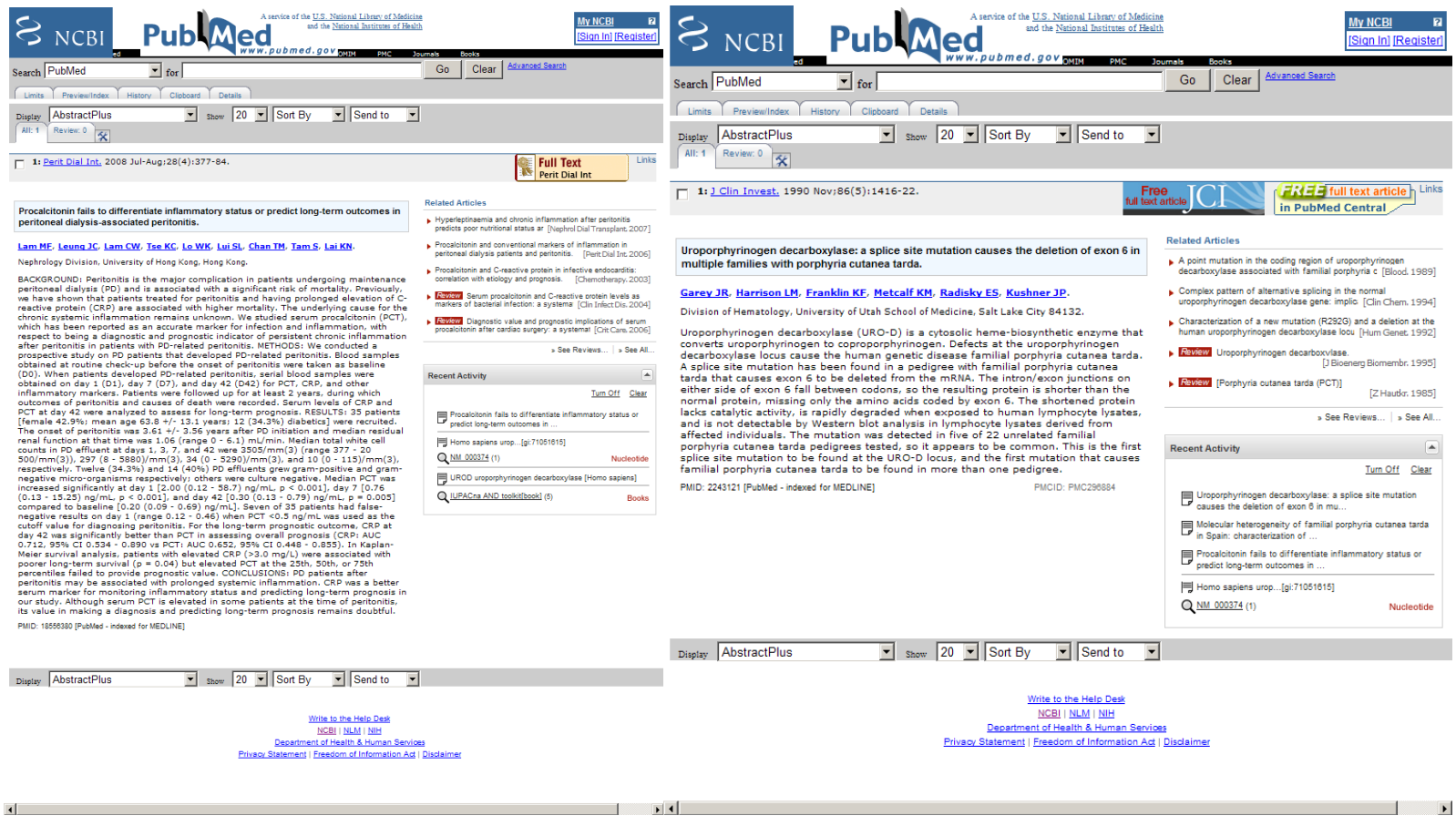
1.2.7 Seleccionad Ver secuencia en formato XML. Localizad la secuencia del gen en este formato de una forma inteligente, sin buscar en todo el fichero manualmente.

Bah¡¡. A priori, buscas “ATG” y te debe llevar al principio de la cadena, pero encuentro más elegante buscar Seq-data, o “IUPACna”, también se puede usar un editor xml para explorar el DOM.



1.2.8 Seleccionad alguno de los links hacia PUBMED. Explicad qué es PUBMED.

Alguno links a pubmed desde el registro de UROD:



Que es pubmed

PubMed, esta disponible por medio del sistema de consulta Entrez del NCBI. Se desarrolló por el National Center for Biotechnology Information (NCBI) en la National Library of Medicine (NLM), del U.S. National Institutes of Health (NIH) Entrez es la herramienta de búsqueda y consulta basada en texto del NCBI para los servicios de PubMed, Nucleotide y Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, y muchos otros. PubMed proporciona el acceso a las citas de la literatura biomédica. LinkOut proporciona el acceso a los artículos en los sitios web originales de las revistas cientificas. PubMed también proporciona el acceso y links a otros recursos de Entrez.

http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html

Introduction

PubMed, available via the NCBI Entrez retrieval system, was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), located at the U.S. National Institutes of Health (NIH). Entrez is the text-based search and retrieval system used at NCBI for services including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others. PubMed provides access to citations from biomedical literature. LinkOut provides access to full-text articles at journal Web sites and other related Web resources. PubMed also provides access and links to the other Entrez molecular biology resources.

Publishers participating in PubMed electronically submit their citations to NCBI prior to or at the time of publication. If the publisher has a web site that offers full-text of its journals, PubMed provides links to that site as well as biological resources, consumer health information, research tools, and more. There may be a charge to access the text or information.

Use the Batch Citation Matcher to match citations to PubMed using bibliographic information such as journal, volume, issue, page number, and year, or the Entrez Programming Utilities that provide access to Entrez data outside of the regular Web query interface.

PubMed Coverage

PubMed provides access to bibliographic information that includes MEDLINE, as well as:

- The out-of-scope citations (e.g., articles on plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and chemistry journals, for which the life sciences articles are indexed for MEDLINE.
- Citations that precede the date that a journal was selected for MEDLINE indexing.
- Some additional life science journals that submit full text to PubMedCentral and receive a qualitative review by NLM.

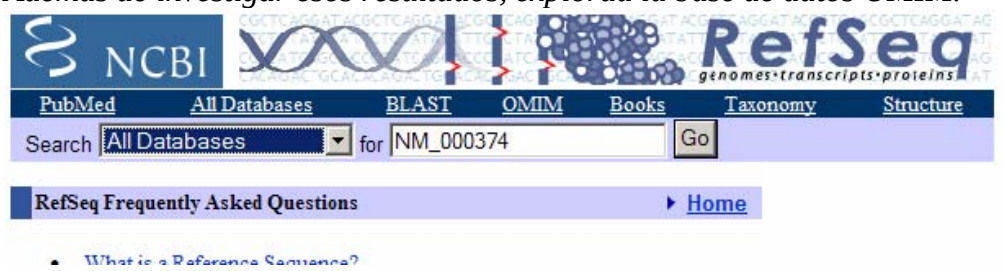
For additional information, please see the NLM Fact Sheet: What's the Difference Between MEDLINE and PubMed?

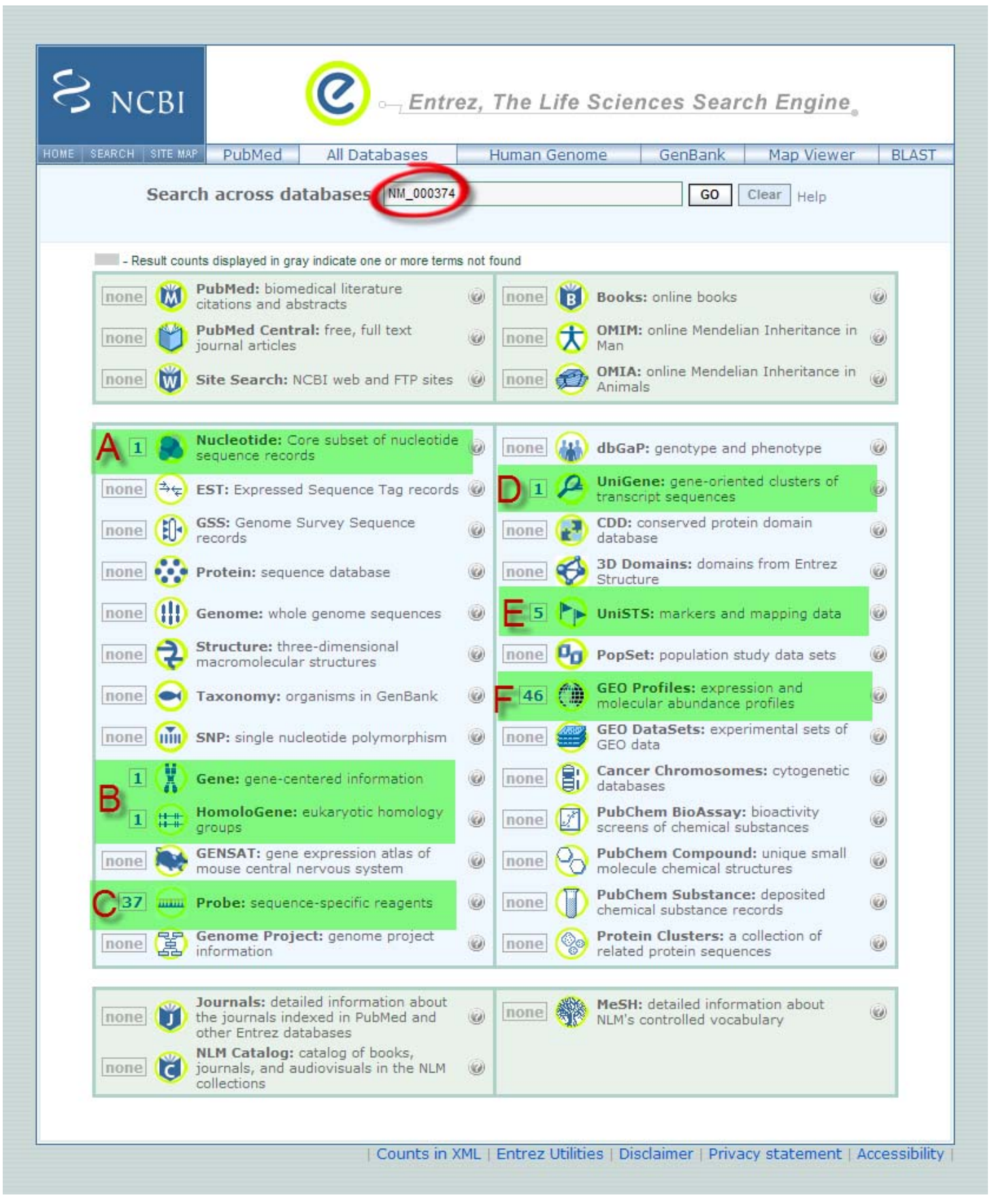
MEDLINE

MEDLINE is the NLM's premier bibliographic database that contains references to journal articles in the life sciences with a concentration on biomedicine. A distinctive feature of MEDLINE is that the records are indexed with NLM's Medical Subject Headings (MeSH). The database contains citations from 1950 to the present, with some older material. New citations that have been indexed with MeSH terms, publication types, GenBank accession numbers, and other indexing data are available daily (Tuesday through Saturday) and display with the tag [PubMed - indexed for MEDLINE]. See also the MEDLINE/PubMed Resources Guide.

1.2.9 Volved a la página principal del NCBI. Para el mismo identificador, repetid la búsqueda pero ahora sin ninguna restricción (en todas las bases de datos a la vez). Comentad brevemente los nuevos resultados aparecidos.

Además de investigar esos resultados, explorad la base de datos OMIM.



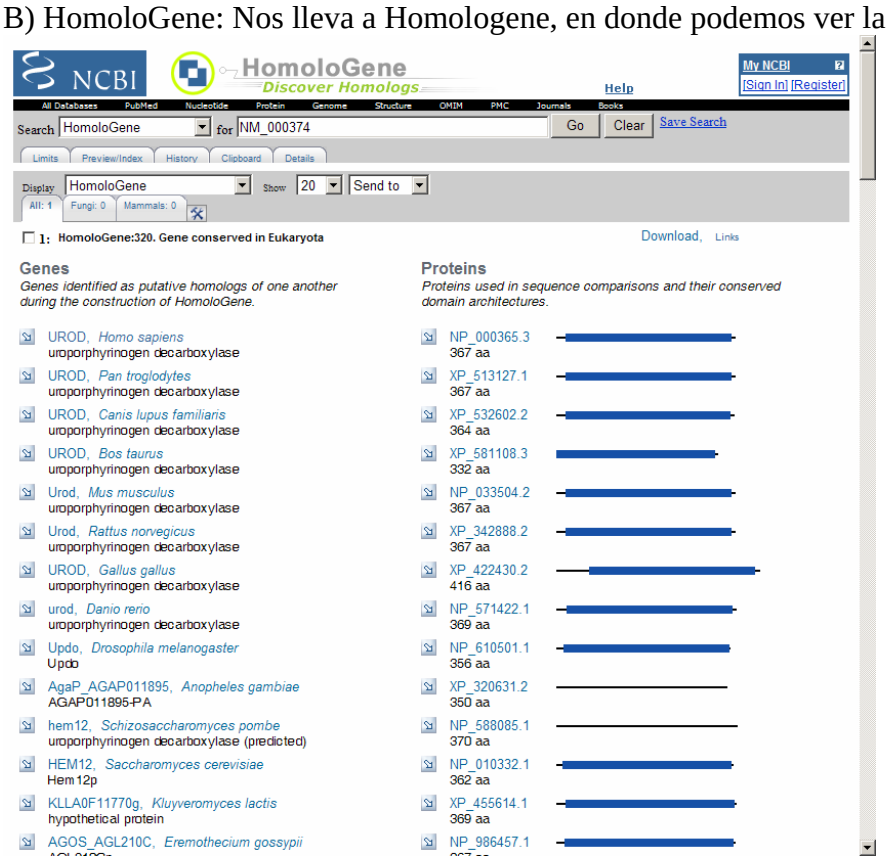


A.- Nos lleva a la búsqueda sobre “Nucleotide” que ya habíamos visto.

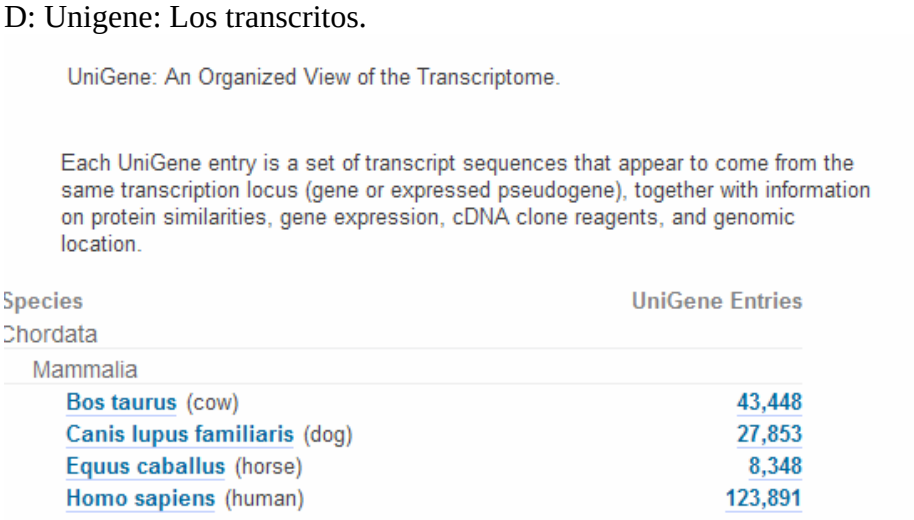
B.- Gene : Nos lleva a la pagina de Entrez Gene de UROD

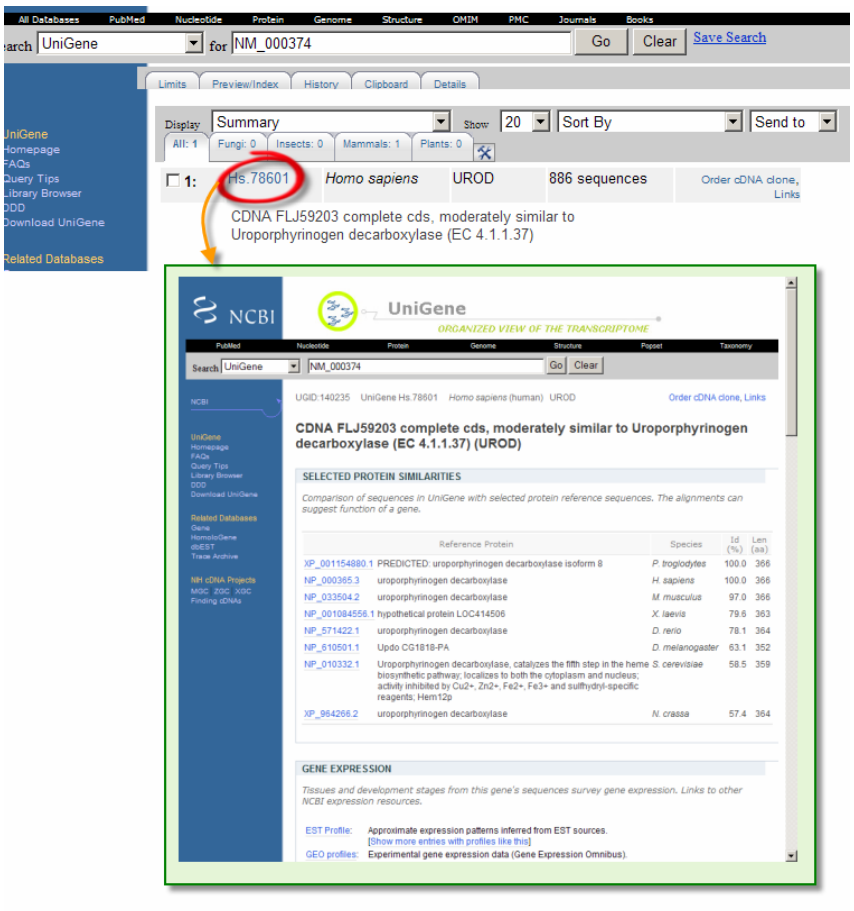
Official Symbol	UROD provided by HGNC
Official Full Name	uroporphyrinogen decarboxylase provided by HGNC
Primary source	HGNC:12591
Locus tag	RP11-69J16.2
See related	Ensembl:ENSG00000126088; HPRD:01441; MIM:176100
Gene type	protein coding
RefSeq status	REVIEWED
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	PCT; UROD
Summary	This gene encodes the fifth enzyme of the heme biosynthetic pathway. This enzyme is responsible for catalyzing the conversion of uroporphyrinogen to coproporphyrinogen through the removal of four carboxymethyl side chains. Mutations and deficiency in this enzyme are known to cause familial porphyria cutanea tarda and hepatoerythropoetic porphyria. [provided by RefSeq]

B) HomoloGene: Nos lleva a Homologene, en donde podemos ver la información de homólogos, fenotipos, etc.



D: Unigene: Los transcritos.

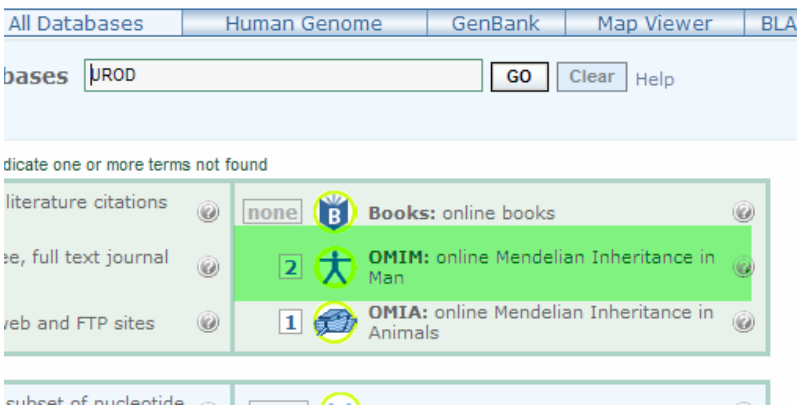




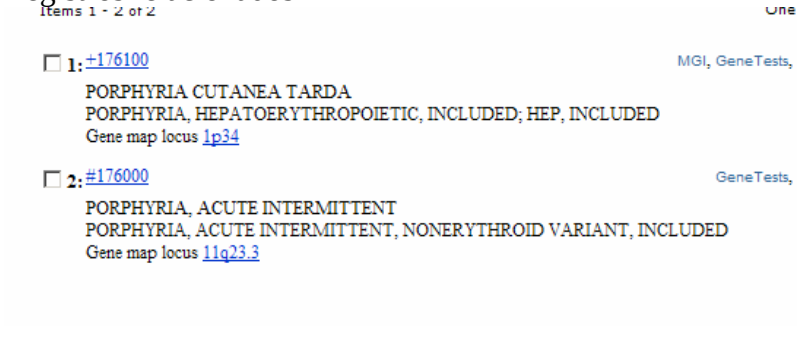
1.2.9.1 Exploración de OMIM

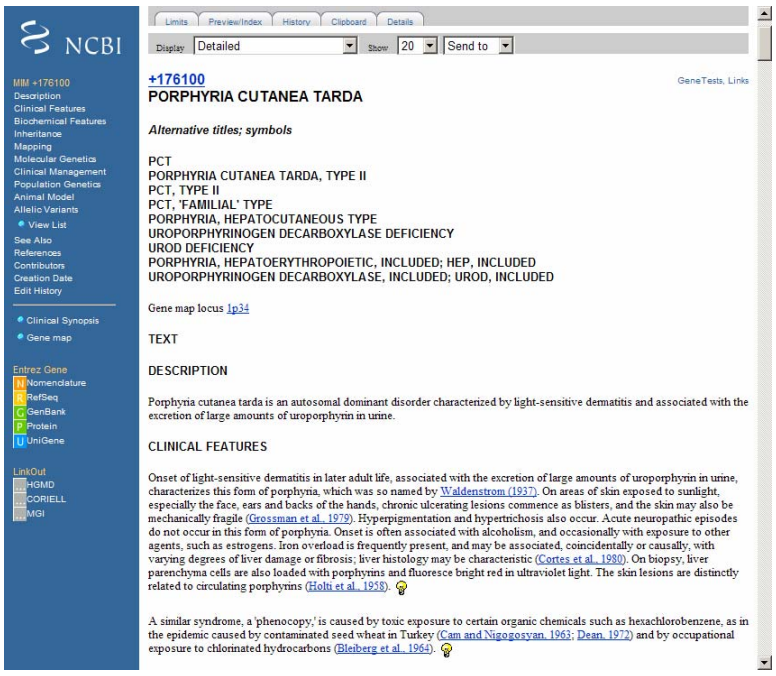
Básicamente OMIM es una base de datos de relaciones fenotipo – genotipo. Es decir, contiene referencias cruzadas entre enfermedades genéticas y genes.

Como OMIM no sale explícitamente repetimos la búsqueda con “UROD”, y entonces si que obtenemos entradas para explorar.



Registros relacionados




NCBI
Limits
Preview/Index
History
Clipboard
Details
Display
Detailed
Show
20
Send to
MIM +176100
Description
Clinical Features
Biochemical Features
Inheritance
Mapping
Molecular Genetics
Clinical Management
Population Genetics
Animal Model
Allelic Variants
View List
See Also
References
Contributors
Creation Date
Edit History
Clinical Synopsis
Gene map
Entrez Gene
Nomenclature
RefSeq
GenBank
Protein
UniGene
LinkOut
HGMD
CORIELL
MGI
+176100
GeneTests, Links
PORPHYRIA CUTANEA TARDA
Alternative titles; symbols
PCT
PORPHYRIA CUTANEA TARDA, TYPE II
PCT, TYPE II
PCT, 'FAMILIAL' TYPE
PORPHYRIA, HEPATOCUTANEOUS TYPE
UROPORPHYRINOGEN DECARBOXYLASE DEFICIENCY
UROD DEFICIENCY
PORPHYRIA, HEPATOERYTHROPOIETIC, INCLUDED; HEP, INCLUDED
UROPORPHYRINOGEN DECARBOXYLASE, INCLUDED; UROD, INCLUDED
Gene map locus 1p34
TEXT
DESCRIPTION
Porphyria cutanea tarda is an autosomal dominant disorder characterized by light-sensitive dermatitis and associated with the excretion of large amounts of uroporphyrin in urine.
CLINICAL FEATURES
Onset of light-sensitive dermatitis in later adult life, associated with the excretion of large amounts of uroporphyrin in urine, characterizes this form of porphyria, which was so named by Waldenstrom (1937). On areas of skin exposed to sunlight, especially the face, ears and backs of the hands, chronic ulcerating lesions commence as blisters, and the skin may also be mechanically fragile (Grossman et al., 1979). Hyperpigmentation and hypertrichosis also occur. Acute neuropathic episodes do not occur in this form of porphyria. Onset is often associated with alcoholism, and occasionally with exposure to other agents, such as estrogens. Iron overload is frequently present, and may be associated, coincidentally or causally, with varying degrees of liver damage or fibrosis; liver histology may be characteristic (Cortes et al., 1980). On biopsy, liver parenchyma cells are also loaded with porphyrins and fluoresce bright red in ultraviolet light. The skin lesions are distinctly related to circulating porphyrins (Holti et al., 1958).
A similar syndrome, a 'phenocopy,' is caused by toxic exposure to certain organic chemicals such as hexachlorobenzene, as in the epidemic caused by contaminated seed wheat in Turkey (Cam and Nigogosyan, 1963; Dean, 1972) and by occupational exposure to chlorinated hydrocarbons (Bleiberg et al., 1964).


NCBI
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
OMIM
Search OMIM for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Entrez
OMIM
Search OMIM
Search Gene Map
Search Morbid Map
Help
OMIM Help
How to Link
FAQ
Numbering System
Symbols
How to Print
Citing OMIM
Download
OMIM Facts
Statistics
Update Log
Restrictions on Use
Allied Resources
Genetic Alliance
Databases
HGMD
Locus-Specific
Model Organisms
MitoMap
Phenotype
Human/Mouse/Rat
Homology Maps
Coriell
The Jackson Laboratory
Human Gene
Nomenclature
Human Genome Resources
Entrez Gene
Genes and Disease
Enter one or more search terms.
Use Limits to restrict your search by search field, chromosome, and other criteria.
Use Index to browse terms found in OMIM records.
Use History to retrieve records from previous searches, or to combine searches.
OMIM® - Online Mendelian Inheritance in Man®
Welcome to OMIM®, Online Mendelian Inheritance in Man®. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.
This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.
OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh.
NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions.
OMIM® and Online Mendelian Inheritance in Man® are registered trademarks of the Johns Hopkins University.

1.3 EJERCICIO 2

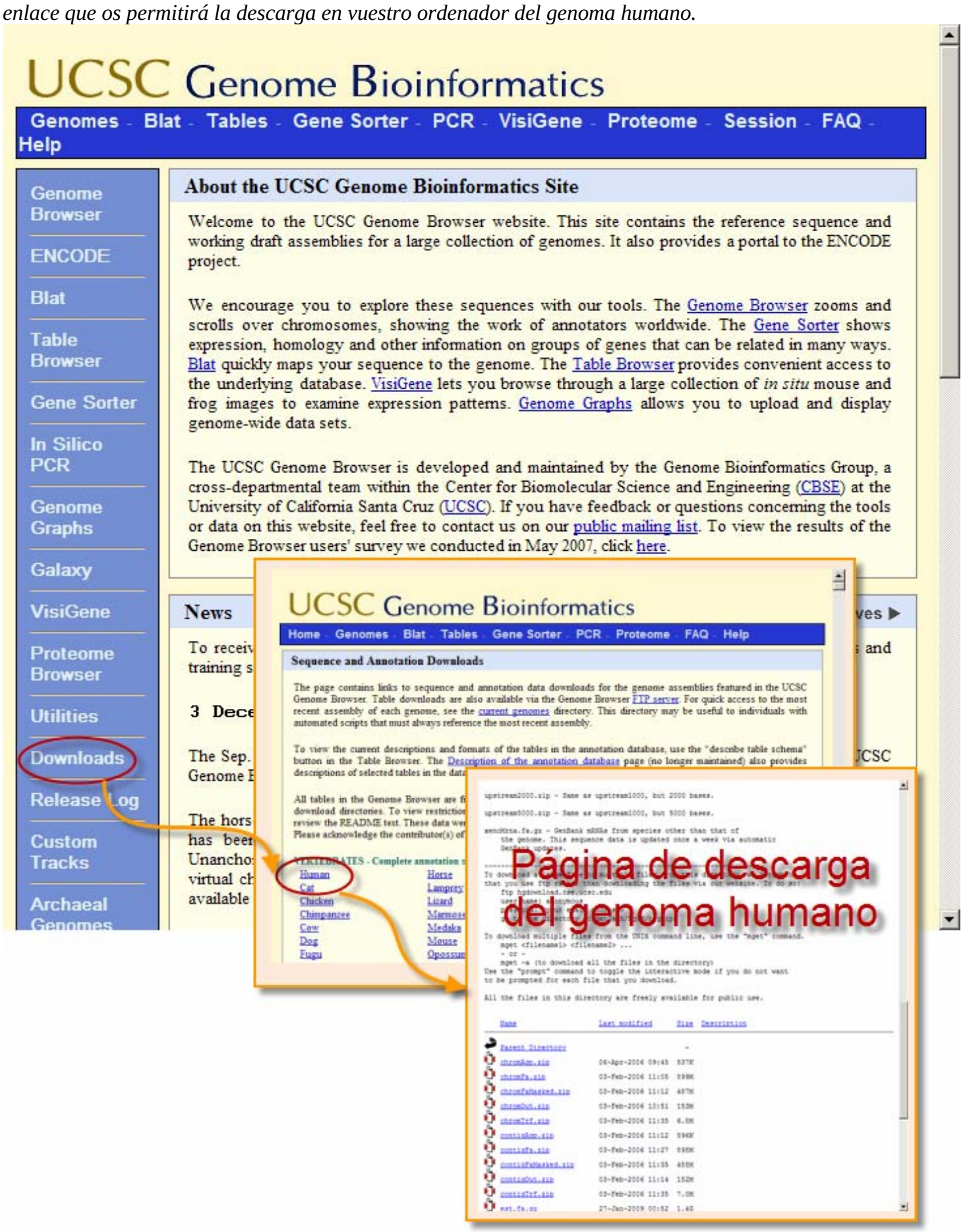
1.3.1 Acceder al portal genómico UCSC:

http://genome.ucsc.edu/

El portal UCSC es junto con ENSEMBL y NCBI uno de los portales más potentes y populares. Las tres WEBS son referencia mundial en temas de anotación genómica.

1.3.2 Seleccionad Genome Browser y Genoma humano.

Antes, sin embargo, echad un vistazo a todas las opciones disponibles desde la página principal. Por ejemplo, buscad el enlace que os permitirá la descarga en vuestro ordenador del genoma humano.



Algunos links visitados:

What does the Genome Browser do?

As vertebrate genome sequences near completion and research re-focuses on their analysis, the issue of effective sequence display becomes critical: it is not helpful to have 3 billion letters of genomic DNA shown as plain text! As an alternative, the UCSC Genome Browser provides a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks (known genes, predicted genes, ESTs, mRNAs, CpG islands, assembly gaps and coverage, chromosomal bands, mouse homologies, and more). Half of the annotation tracks are computed at UCSC from publicly available sequence data. The remaining tracks are provided by collaborators worldwide. Users can also add their own custom tracks to the browser for educational or research purposes.

The Genome Browser stacks annotation tracks beneath genome coordinate positions, allowing rapid visual correlation of different types of information. The user can look at a whole chromosome to get a feel for gene density, open a specific cytogenetic band to see a positionally mapped disease gene candidate, or zoom in to a particular gene to view its spliced ESTs and possible alternative splicing. The Genome Browser itself does not draw conclusions; rather, it collates all relevant information in one location, leaving the exploration and interpretation to the user.

The Genome Browser supports text and sequence based searches that provide quick, precise access to any region of specific interest. Secondary links from individual entries within annotation tracks lead to sequence details and supplementary off-site databases. To control information overload, tracks need not be displayed in full. Tracks can be hidden, collapsed into a condensed or single-line display, or filtered according to the user's criteria. Zooming and scrolling controls help to narrow or broaden the displayed chromosomal range to focus on the exact region of interest. Clicking on an individual item within a track opens a details page containing a summary of properties and links to off-site repositories such as PubMed, GenBank, Entrez, and OMIM. The page provides item-specific information on position, cytoband, strand, data source, and encoded protein, mRNA, genomic sequence and alignment, as appropriate to the nature of the track.

A blue navigation bar at the top of the browser provides links to several other tools and data sources. For instance, the DNA link enables the user to view the raw genomic DNA sequence for the coordinate range displayed in the browser window. This DNA can encode track features via elaborate text formatting options. Other links tie the Genome Browser to the BLAT alignment tool, provide access to the underlying relational database via the Table Browser, convert coordinates across different assembly dates, and open the window at the complementary Ensembl or NCBI Map Viewer annotation.

The browser data represents an immense collaborative effort involving thousands of people from the international biomedical research community. The UCSC Bioinformatics Group itself does no sequencing. Although it creates the majority of the annotation tracks in-house, the annotations are based on publicly available data contributed by many labs and research groups throughout the world. Several of the Genome Browser annotations are generated in collaboration with outside individuals or are contributed wholly by external research groups. UCSC's other major roles include building genome assemblies, creating the Genome Browser work environment, and serving it online. The majority of the sequence data, annotation tracks, and even software are in the public domain and are available for anyone to download.

In addition to the Genome Browser, the UCSC Genome Bioinformatics group provides several other tools for viewing and interpreting genome data:

- BLAT - a fast sequence-alignment tool similar to BLAST. Read more.
- Table Browser - convenient text-based access to the database underlying the Genome Browser. Read more.
- Genome Graphs - a tool that allows you to upload and display genome-wide data sets such as the results of genome-wide SNP association studies, linkage studies and homozygosity mapping. Read more.
- Gene Sorter - expression, homology, and other information on groups of genes that can be related in many ways. Read more.
- Proteome Browser (accessible from Known Genes details pages) - protein property data and links to a wealth of related information. Read more.

About the ENCODE Data Coordination Center (DCC)

The UCSC Genome Browser displays data produced by the Encyclopedia of DNA Elements (ENCODE) Consortium, an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of the functional elements in the human genome.

In September 2007, the ENCODE project was scaled up from a pilot phase to cover the entire human genome. To access genome-wide ENCODE data in the Genome Browser, go to your region of interest and select ENCODE tracks (marked with the NHGRI logo).

During the pilot phase of the ENCODE project (2003-2007), experiments focused on a limited set of genomic regions comprising roughly 1% of the human genome. Data limited to the pilot regions is located in ENCODE-specific track groups in the browser. The ENCODE Pilot Project web pages provide convenient browser access to these regions.

Other key differences from the pilot project are:

- Identification of common cell types to facilitate integrative analysis
- New experimental technologies based on high-throughput sequencing
- A data release policy restricting use of data for nine months following release

Click here to go to the main UCSC Genome Browser site, which provides access to sequence and annotation data for a large collection of genome assemblies. See the Genome Browser User's Guide for information about displaying tracks and navigating in the Genome Browser.

About BLAT

BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 25 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 25 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates.

BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for those heavily involved in repeats. The index takes up a bit less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment. Protein BLAT works in a similar manner, except with 4-mers rather than 11-mers. The protein index takes a little more than 2 gigabytes.

BLAT was written by Jim Kent. Like most of Jim's software, interactive use on this web server is free to all. Sources and executables to run batch jobs on your own server are available free for academic, personal, and non-profit purposes. Non-exclusive commercial licenses are also available. See the Kent Informatics website for details.

For more information on the graphical version of BLAT, click the Help button on the top menu bar or see the Genome Browser FAQ.

BLAT lo usare después en la segunda y tercera parte del PEC

1.3.3 Escribid UROD en el recuadro apropiado e iniciad la búsqueda.

Nuevamente, antes de empezar, echad un vistazo a las diferentes opciones y formatos que se pueden usar para acceder a esta base de datos.

Los formatos posibles para las consultas se indican en el mismo formulario de consulta.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or a cytological band, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
RH18061;RH80175	Displays region between STS markers RH18061;RH80175. This syntax may also be used for other range queries, such as between cytobands and uniquely-determined ESTs, mRNAs, refSeqs, etc.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs
homeobox caudal	Lists mRNAs for caudal homeobox genes
zinc finger	Lists many zinc finger mRNAs
kruppel zinc finger	Lists only kruppel-like zinc fingers
huntington	Lists candidate genes associated with Huntington's disease
zahler	Lists mRNAs deposited by scientist named Zahler
Evans,J.E.	Lists mRNAs deposited by co-author J.E. Evans

Use this last format for author queries. Although GenBank requires the search format *Evans JE*, internally it uses the format *Evans,J.E.*.

1.3.4 Seleccionad el resultado bajo la categoría de REFSEQ Genes.

Los exones de este gen han sido verificados por algún tipo de evidencia experimental o como Mínimo son objeto de evaluación por el proyecto REFSEQ. Algunos genes pueden tener más de una forma alternativa posible.

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	image width	
Mammal	Human	Mar. 2006	UROD	620	submit

Click here to reset the browser user interface settings to their defaults.

add custom tracks | configure tracks and display | clear position

UCSC Genes

UROD (uc009vxm.1) at chr1:45250417-45252835 - Homo sapiens isolate normal patient 1 uroporphyrinogen
UROD (uc009vxl.1) at chr1:45250417-45252722 - cDNA, FLJ79527, moderately similar to Uroporphyrinogen
UROD (uc001cnc.1) at chr1:45251116-45253928 - Homo sapiens isolate normal patient 1 uroporphyrinogen
UROD (uc001cnb.1) at chr1:45250417-45253928 - uroporphyrinogen decarboxylase
UROD (uc001cna.1) at chr1:45250417-45253928 - uroporphyrinogen decarboxylase

RefSeq Genes

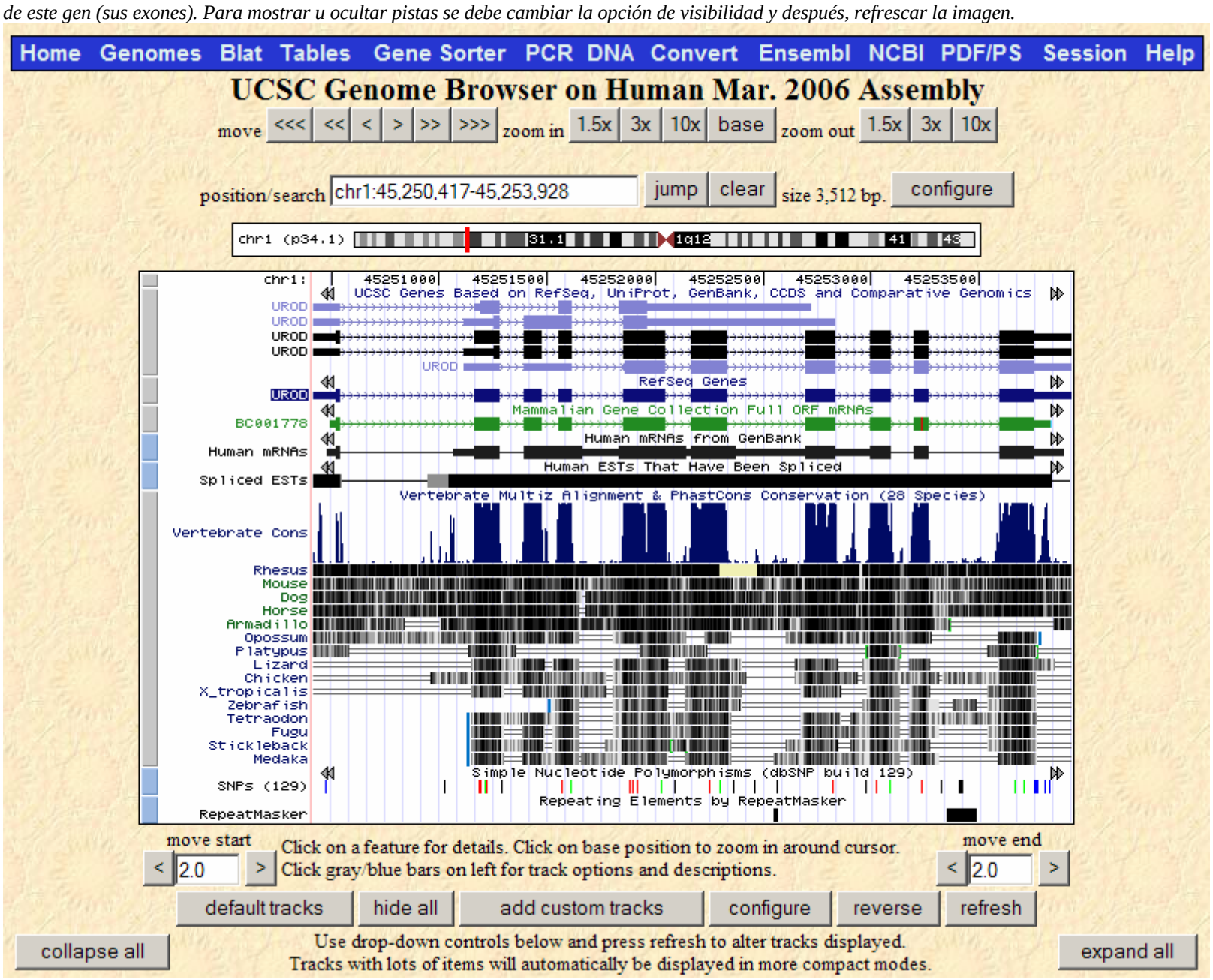
UROD at chr1:45250417-45253928 - (NM_000374) uroporphyrinogen decarboxylase

Non-Human RefSeq Genes

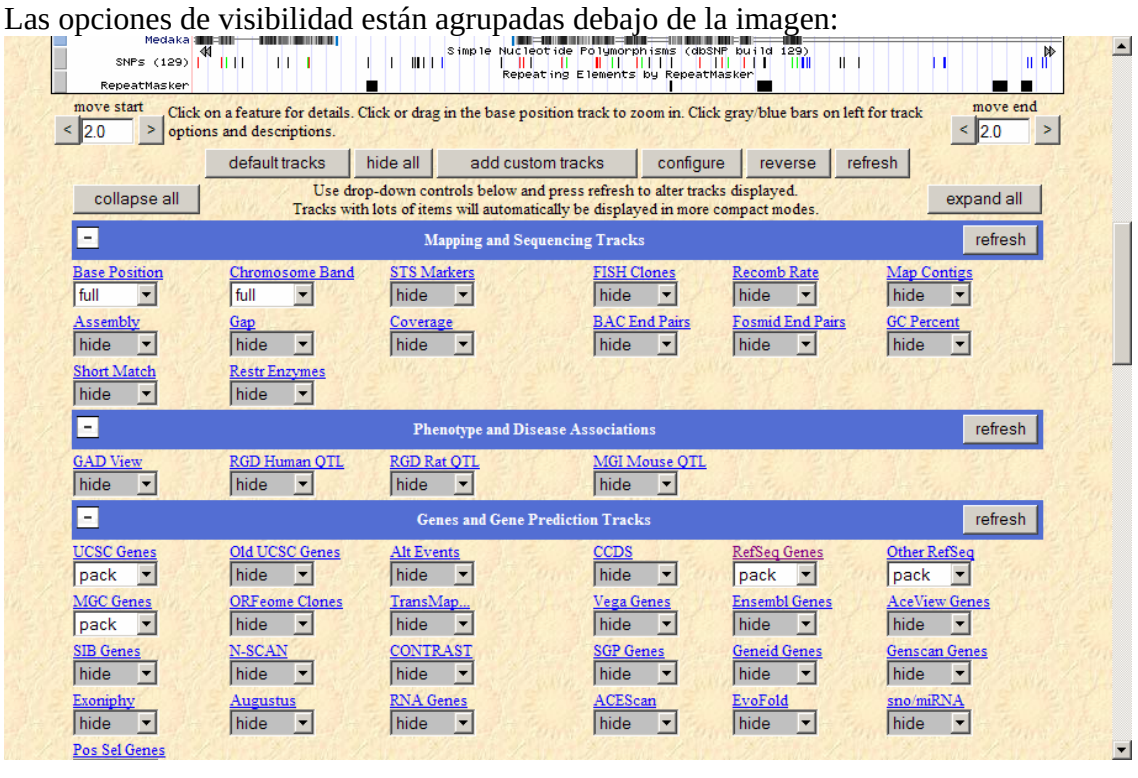
Urod at chr1:45250522-45253794 - (NM_019209) uroporphyrinogen decarboxylase
Urod at chr1:45251165-45253794 - (NM_009478) uroporphyrinogen decarboxylase

1.3.5 Deberíais estar delante de una imagen similar a ésta. Justo después de la imagen, Encontrareis una serie de opciones para mostrar y esconder las diferentes pistas de datos, divididas de forma temática:

Os encontráis delante de la anotación de este fragmento cromosómico (donde se encuentra el gen UROD). La información se muestra en forma de pista donde se presenta una cierta característica genómica. Por ejemplo, las pistas con el nombre de UROD y de REFSEQ contienen la anotación experimental de este gen (sus exones). Para mostrar u ocultar pistas se debe cambiar la opción de visibilidad y después, refrescar la imagen.

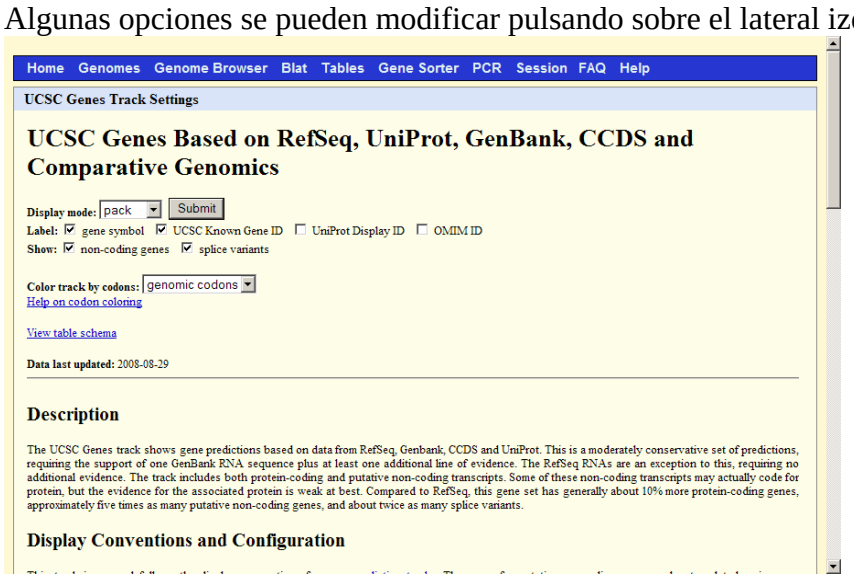


Las opciones de visibilidad están agrupadas debajo de la imagen:



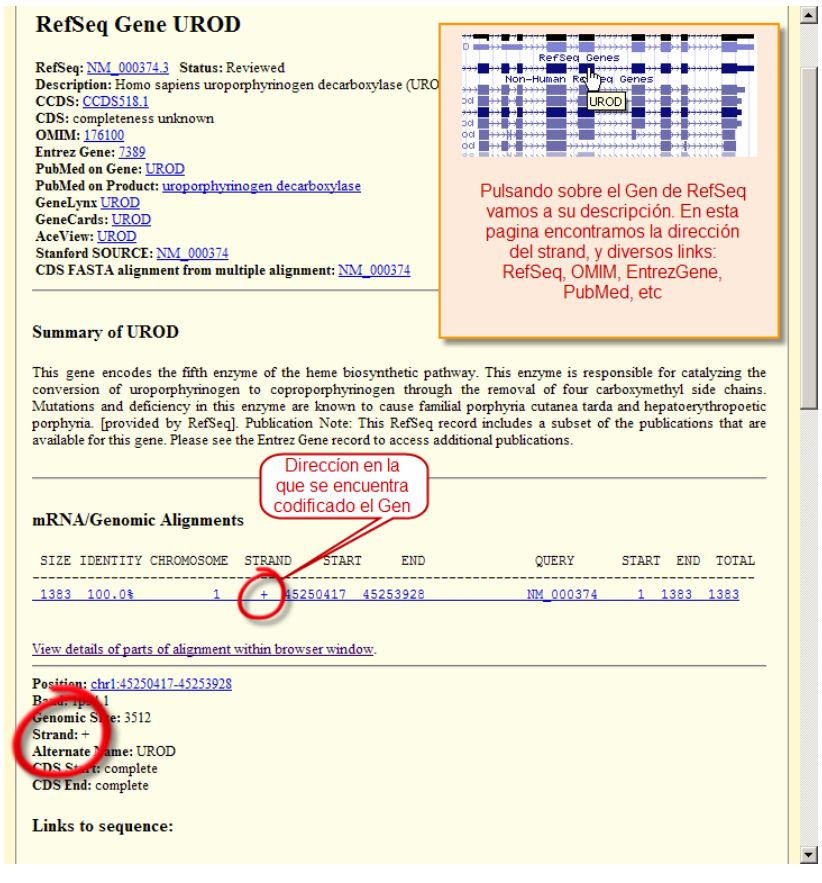
Pulsando sobre el nombre de cada una de ellas se accede a la ayuda

Algunas opciones se pueden modificar pulsando sobre el lateral izquierdo de la imagen:



1.3.6 Buscad los exones del gen URO-D. Explorad los diferentes bloques de opciones.

Buscad ayuda sobre cada opción, seleccionad el link con el nombre correspondiente.
Cada caja negra se corresponde con un exón del gen UROD. Averiguad en cuál de las dos direcciones de la molécula (+ o -) se encuentra el gen.



Un buen punto de partida para obtener ayuda es el material que ofrece OpenHelix

Home
Online Tutorials
Subscriptions
Regional Seminars
Other Services
About Us
Contact Us

UCSC Genome Browser

Tutorial and training materials sponsored by
University of California, Santa Cruz, providers of UCSC Genome Browser

UCSC Genome Bioinformatics

UCSC Genome Browser The UCSC Genome Browser provides a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks (known genes, predicted genes, ESTs, mRNAs, CpG islands, assembly gaps and coverage, chromosomal bands, species homologies, SNPs, and more).

Below are the links for the tutorial and training material the UCSC Genome Browser, including basic software usage and functionality, Gene Sorter, Table Browser and Custom Tracks topics, updated periodically. There are two sets of training materials: Introductory and Advanced topics.

Also consider bringing live onsite hands-on computer seminars on the UCSC Genome Browser to your institution.

UCSC Introductory Genome Browser Training:

Launch	Download	Download	Download
Online Tutorial	PowerPoint Slides (Optimized for Windows®)	Slide Handouts (PDF file)	Hands-on Exercises (PDF file)

UCSC Advanced Topics Training (Table Browser, Custom Tracks & Gene Sorter):

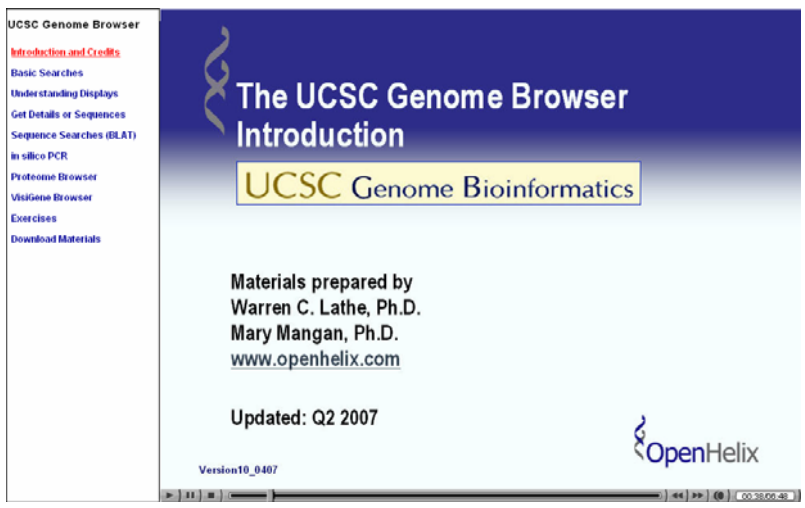
Launch	Download	Download	Download
Online Tutorial	PowerPoint Slides (Optimized for Windows®)	Slide Handouts (PDF file)	Hands-on Exercises (PDF file)

Quick Reference Cards & UCSC Link:

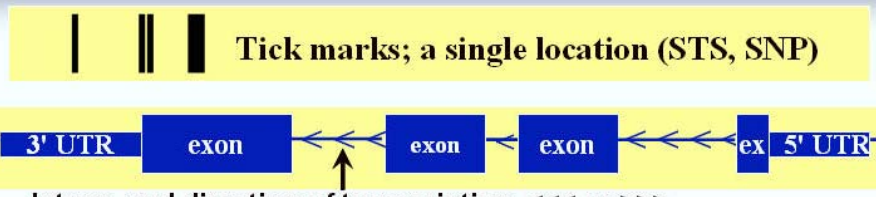
Order	Link
Free Reference Card	Visit the Resource

Click here for technical information on using OpenHelix tutorial and training materials.

Other Online Tutorials
View a catalog of all OpenHelix tutorials.



Visual Cues on the Genome Browser

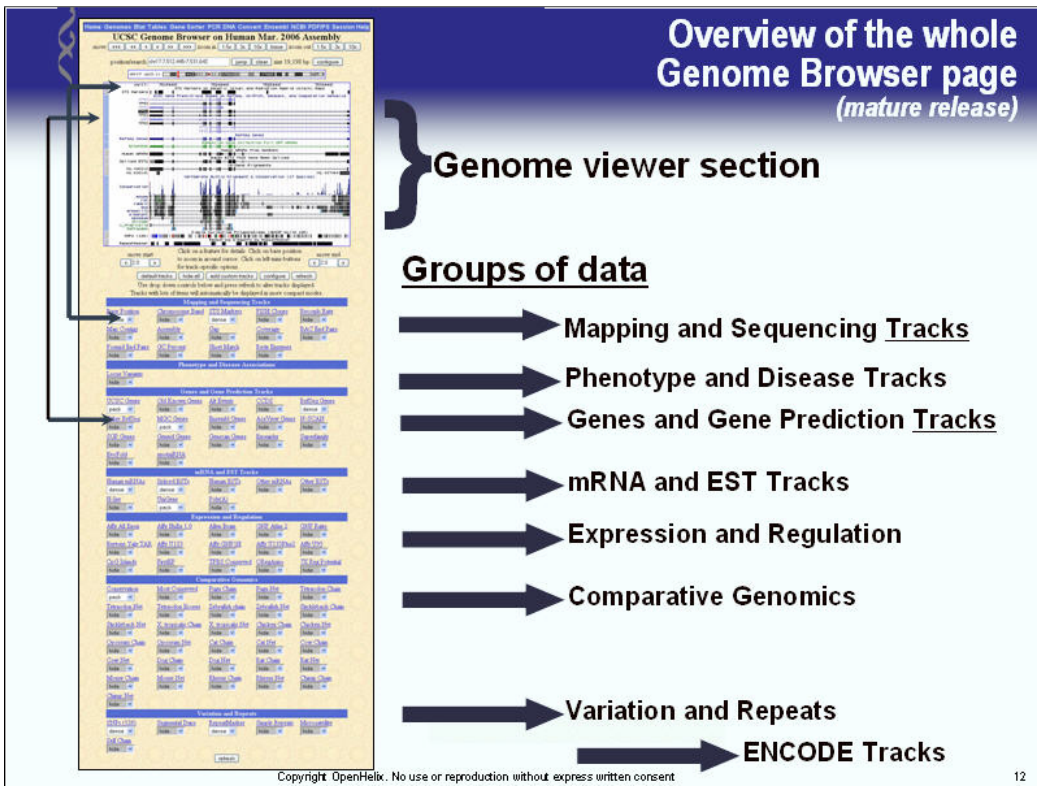


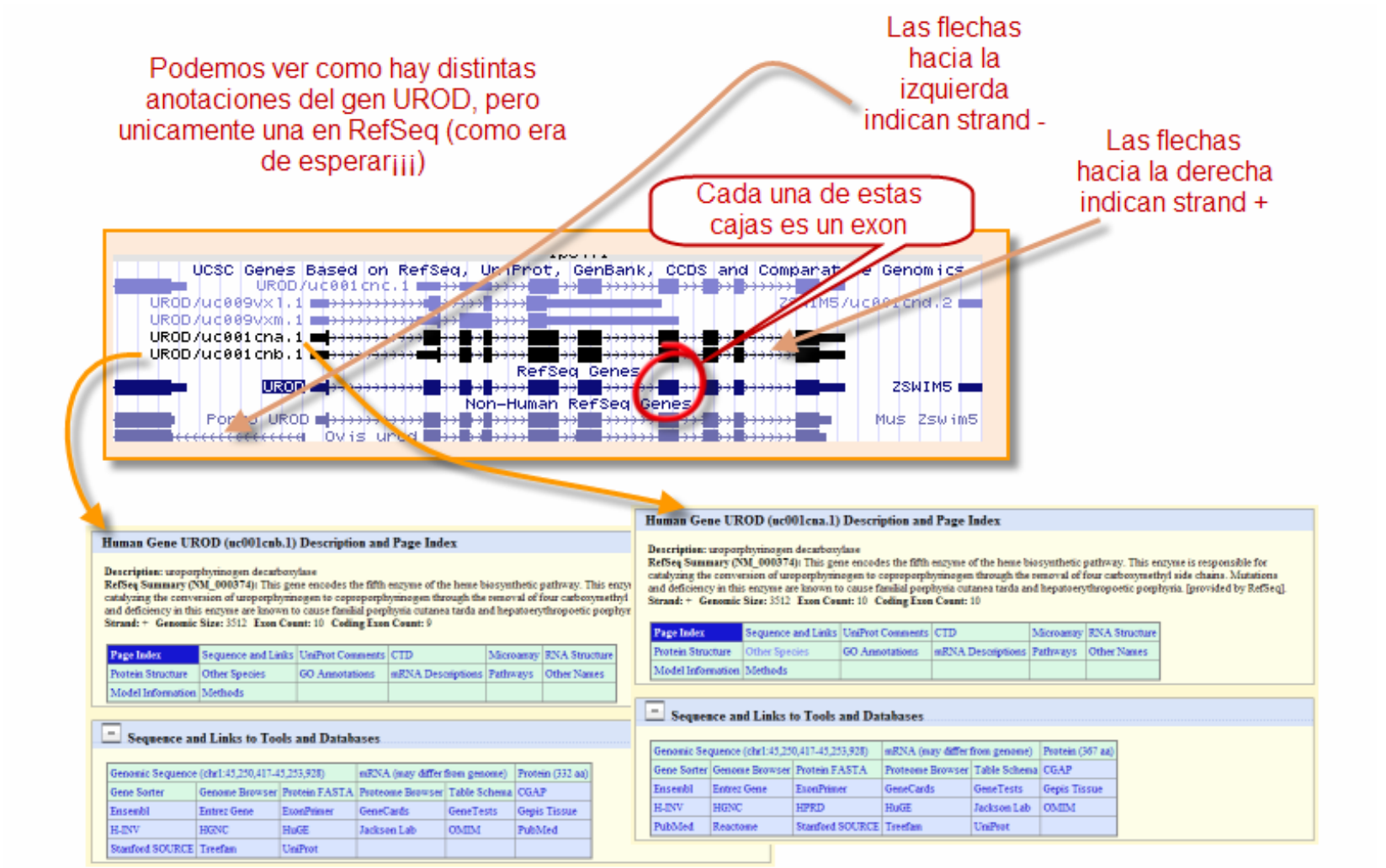
Track colors *may* have meaning—for example, UCSC Gene track:

- If there is a corresponding PDB entry, = black
- If there is a corresponding reviewed/validated seq, = dark blue
- If there is a non-RefSeq seq, = lightest blue

For some tracks, the height of a bar is increased likelihood of an evolutionary relationship (conservation track)

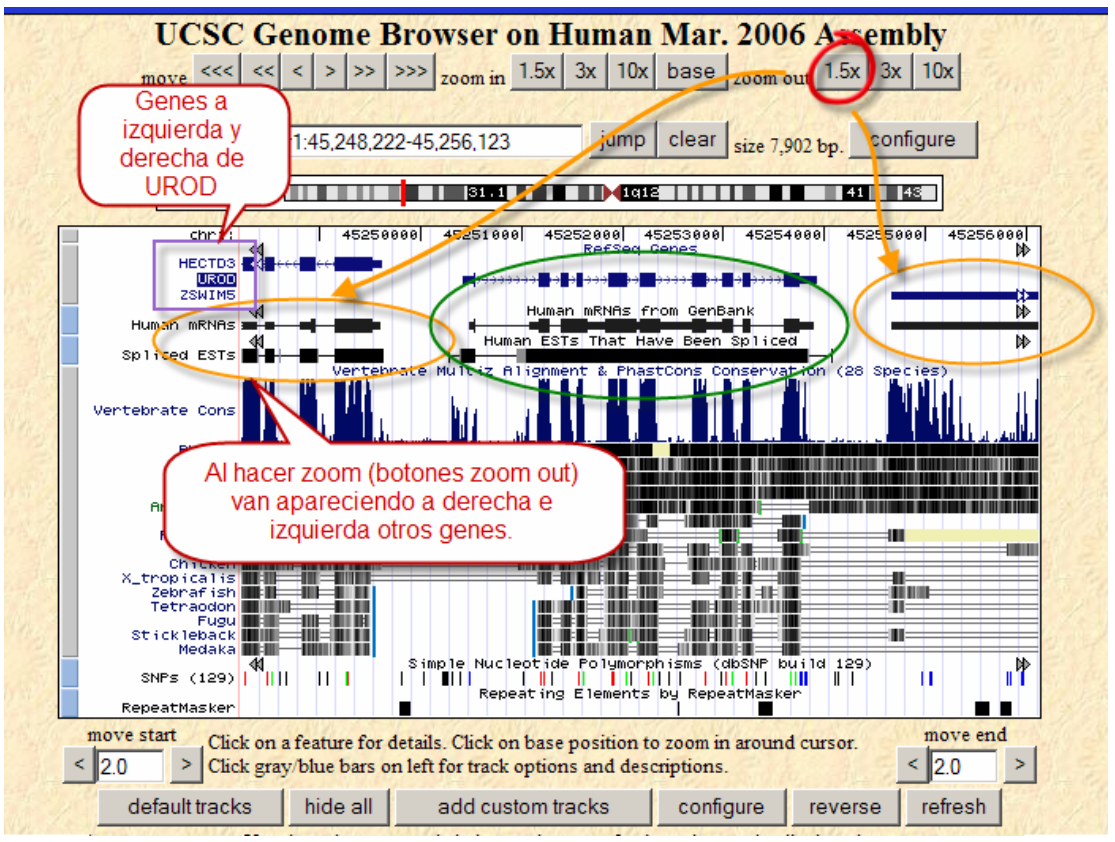
Overview of the whole Genome Browser page *(mature release)*



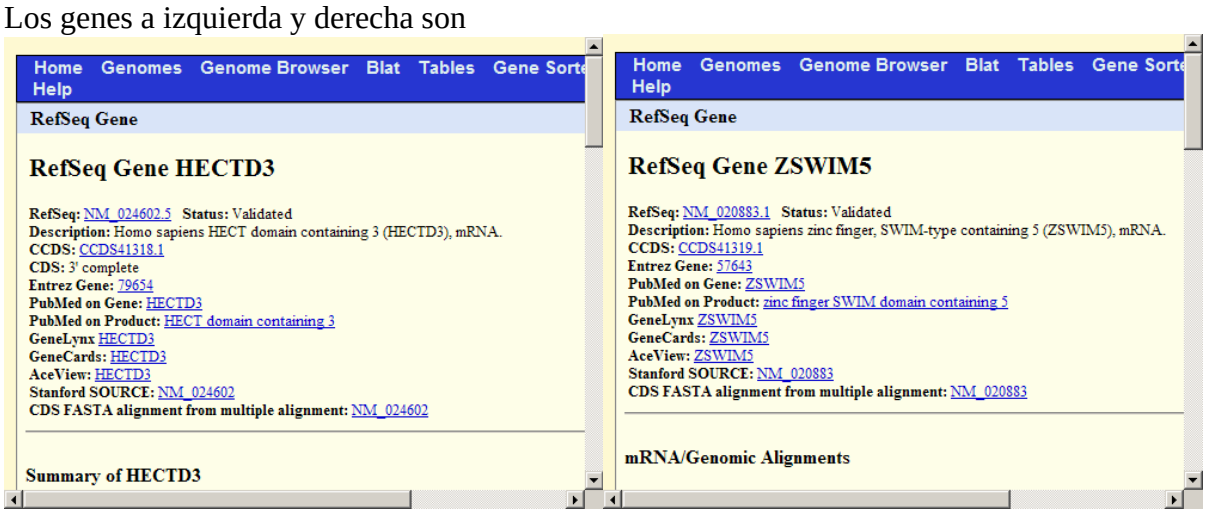

Podemos ver como hay distintas anotaciones del gen UROD, pero unicamente una en RefSeq (como era de esperar¡¡¡)
Las flechas hacia la izquierda indican strand -
Las flechas hacia la derecha indican strand +
Cada una de estas cajas es un exon
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
UROD/uc001cnc.1
UROD/uc009vxl.1
UROD/uc009vxm.1
UROD/uc001cna.1
UROD/uc001cnb.1
RefSeq Genes
UROD
ZSWIM5
Non-Human RefSeq Genes
Mus Zswim5
Human Gene UROD (uc001cnb.1) Description and Page Index
Description: uroporphyrinogen decarboxylase
RefSeq Summary (NM_000374): This gene encodes the fifth enzyme of the heme biosynthetic pathway. This enzy
catalyzing the conversion of uroporphyrinogen to coproporphyrinogen through the removal of four carboxymethyl
and deficiency in this enzyme are known to cause familial porphyria cutanea tarda and hepatoerythropoietic porphyr
Strand: + Genomic Size: 3512 Exon Count: 10 Coding Exon Count: 9
Page Index | Sequence and Links | UniProt Comments | CTD | Microarray | RNA Structure
Protein Structure | Other Species | GO Annotations | mRNA Descriptions | Pathways | Other Names
Model Information | Methods
Sequence and Links to Tools and Databases
Genomic Sequence (chr1:45,250,417-45,253,928) | mRNA (may differ from genome) | Protein (332 aa)
Gene Sorter | Genome Browser | Protein FASTA | Proteome Browser | Table Schema | CGAP
Ensembl | Entrez Gene | ExonPrimer | GeneCards | GeneTests | Gepis Tissue
H-INV | HGNC | HuGE | Jackson Lab | OMIM | PubMed
Stanford SOURCE | Treefam | UniProt
Human Gene UROD (uc001cna.1) Description and Page Index
Description: uroporphyrinogen decarboxylase
RefSeq Summary (NM_000374): This gene encodes the fifth enzyme of the heme biosynthetic pathway. This enzyme is responsible for catalyzing the conversion of uroporphyrinogen to coproporphyrinogen through the removal of four carboxymethyl side chains. Mutations and deficiency in this enzyme are known to cause familial porphyria cutanea tarda and hepatoerythropoietic porphyria. [provided by RefSeq]
Strand: - Genomic Size: 3512 Exon Count: 10 Coding Exon Count: 10
Page Index | Sequence and Links | UniProt Comments | CTD | Microarray | RNA Structure
Protein Structure | Other Species | GO Annotations | mRNA Descriptions | Pathways | Other Names
Model Information | Methods
Sequence and Links to Tools and Databases
Genomic Sequence (chr1:45,250,417-45,253,928) | mRNA (may differ from genome) | Protein (367 aa)
Gene Sorter | Genome Browser | Protein FASTA | Proteome Browser | Table Schema | CGAP
Ensembl | Entrez Gene | ExonPrimer | GeneCards | GeneTests | Gepis Tissue
H-INV | HGNC | HPRD | HuGE | Jackson Lab | OMIM
PubMed | Reactome | Stanford SOURCE | Treefam | UniProt

1.3.7 Jugad con el nivel de detalle (ZOOM).

Anotad qué genes están alrededor del gen UROD.

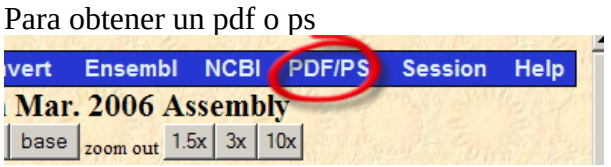


Los genes a izquierda y derecha son



1.3.8 Buscad la opción que permite crear un documento PDF con la anotación gráfica del fragmento actual.

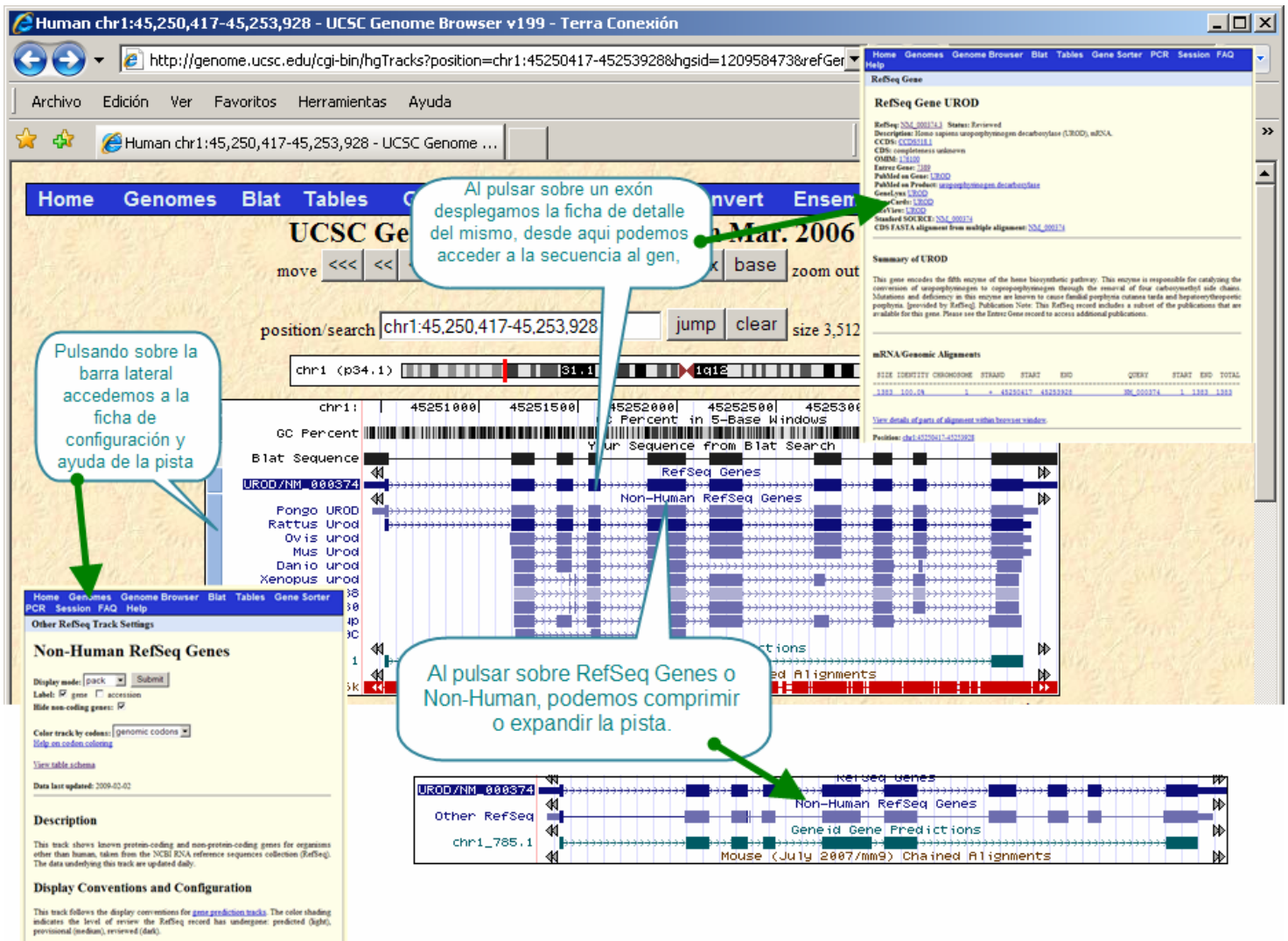
Para obtener un pdf o ps

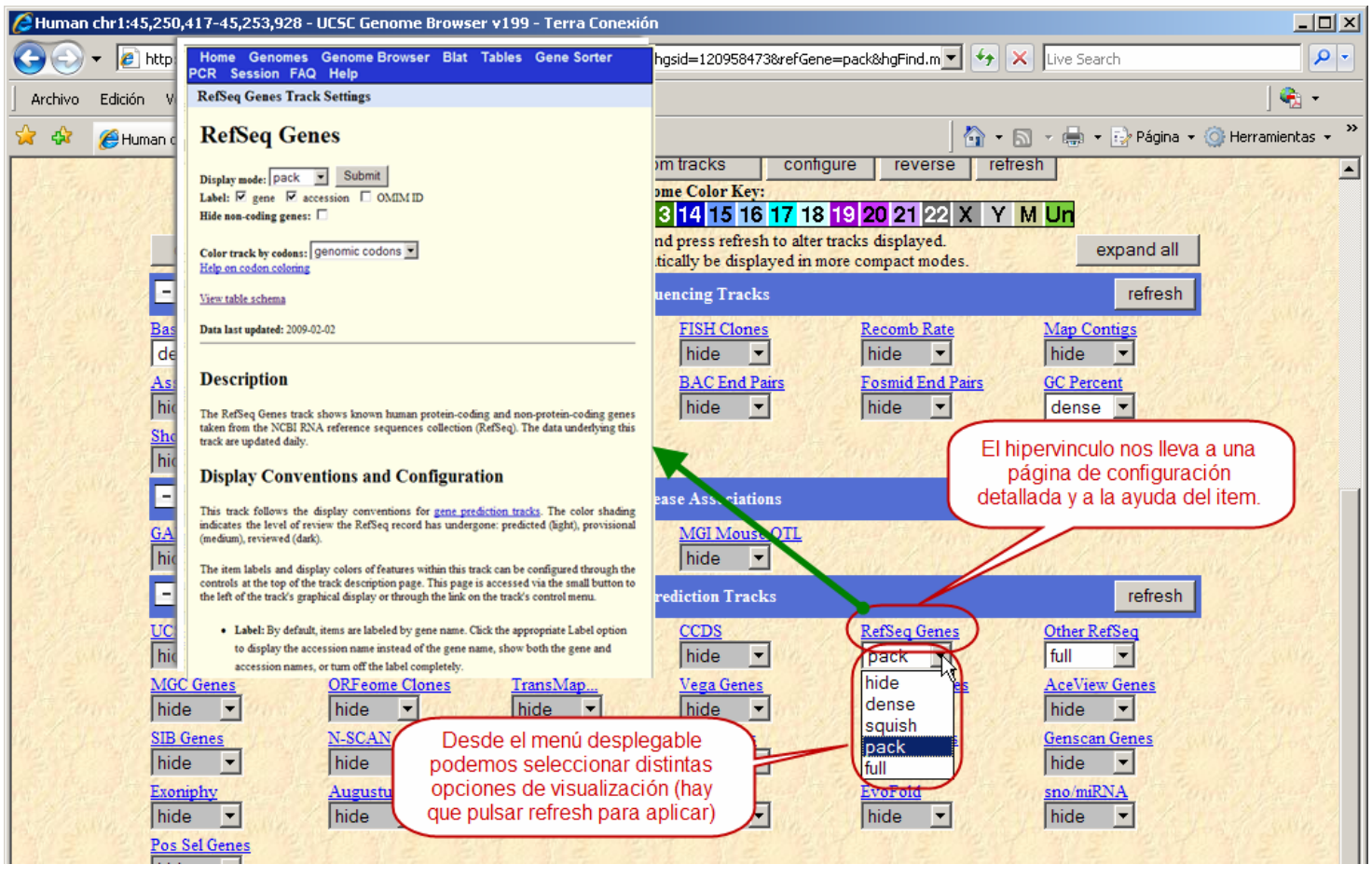


1.3.9 Interaccionad con la imagen (hacer click sobre alguno de los exones).

En la mayoría de casos, seleccionar una caja (feature) produce el pliegue o despliegue de la información asociada a esta pista. A veces, en cambio, los enlaces pueden llevarte a una página específica donde hay más opciones al respecto.

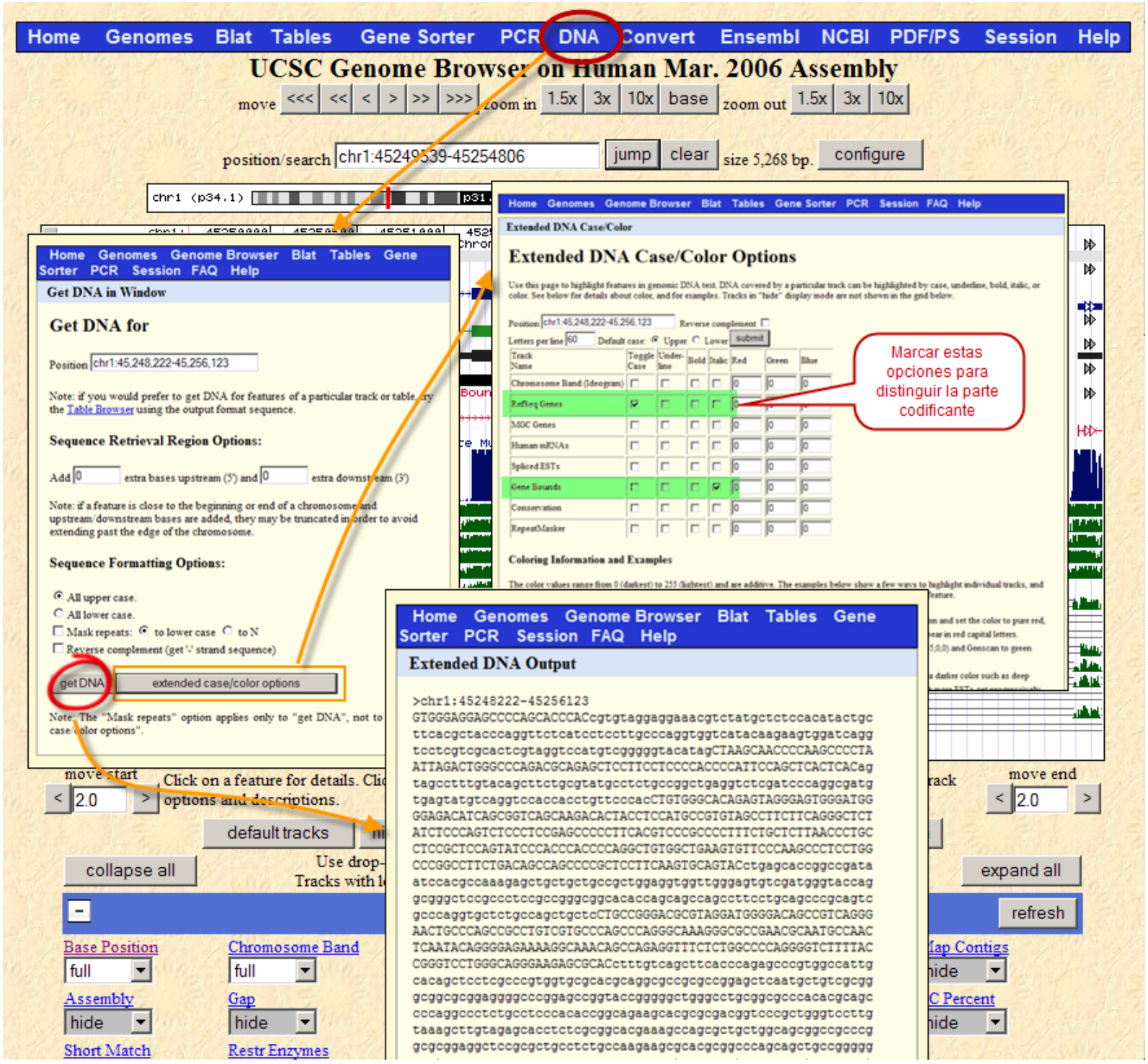
A modo de resumen:





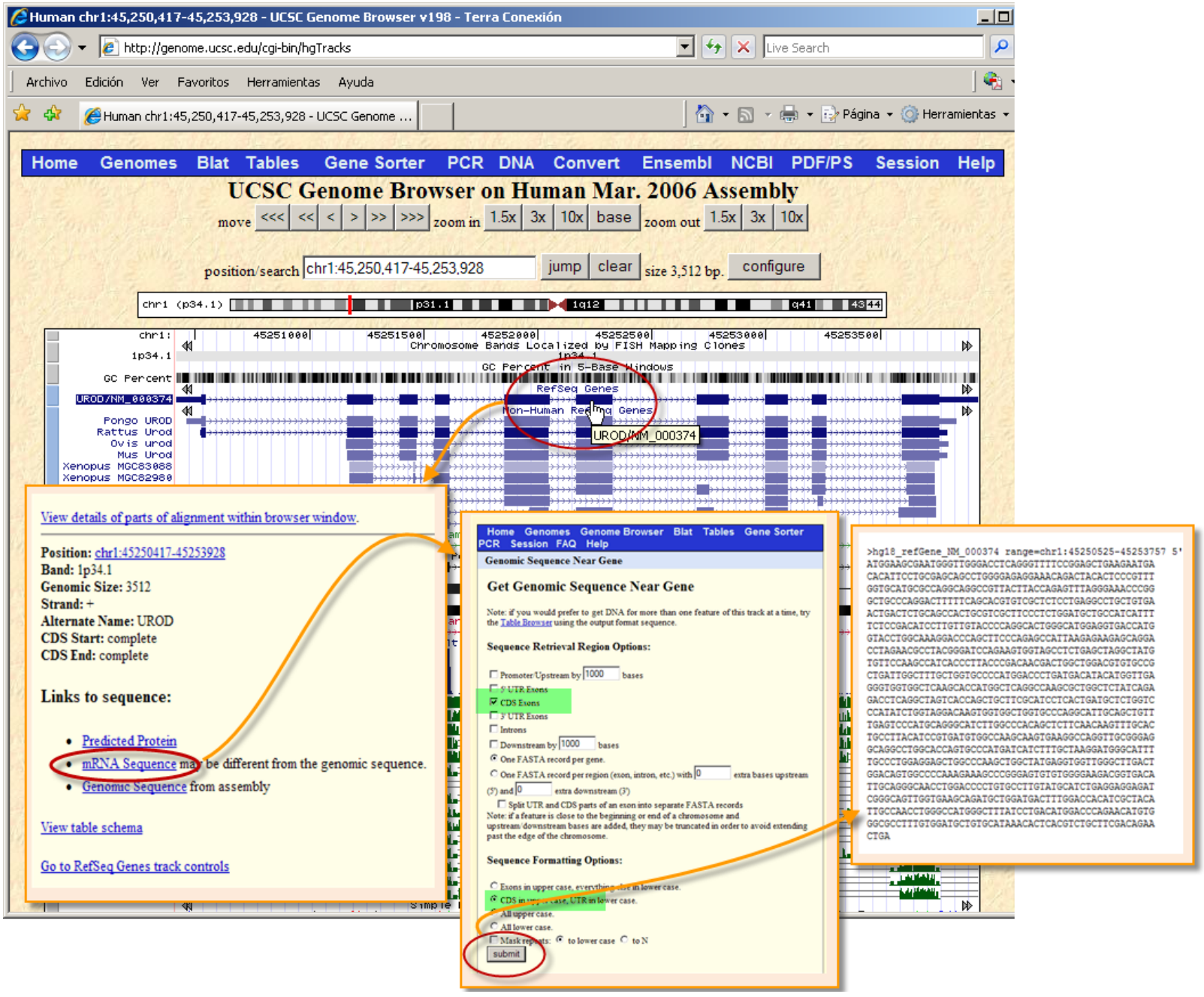
1.3.10 Buscar cómo extraer la secuencia de los exones que pertenecen a la pista REFSEQ. Mostradme solamente la secuencia codificante (CDS).

1.3.10.1 Extracción de toda la secuencia sin quitar lo intrones, nos serviría para comparar



En la figura “Extended DNA Output” se muestran Exones en mayúscula, y en minúscula las partes no codificantes

1.3.10.2 Para obtener la parte codificante

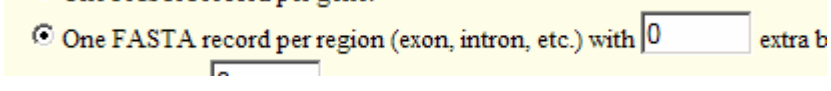



```
>hg18_refGene_NM_000374 range=chr1:45250525-45253757 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGAAGCGAATGGGTTGGGACCTCAGGGTTTTCCGGAGCTGAAGAATGA
CACATTCCTGCGAGCAGCCTGGGGAGAGGAAACAGACTACACTCCCGTTT
GGTGCATGCGCCAGGCAGGCCGTTACTTACCAGAGTTTAGGGAAACCCGG
GCTGCCCAGGACTTTTTCAGCACGTGTCGCTCTCCTGAGGCCTGCTGTGA
ACTGACTCTGCAGCCACTGCGTCGCTTCCCTCTGGATGCTGCCATCATTT
TCTCCGACATCCTTGTTGTACCCCAGGCACTGGGCATGGAGGTGACCATG
GTACCTGGCAAAGGACCCAGCTTCCCAGAGCCATTAAGAGAAGAGCAGGA
CCTAGAACGCCTACGGGATCCAGAAGTGGTAGCCTCTGAGCTAGGCTATG
TGTTCCAAGCCATCACCCTTACCCGACAACGACTGGCTGGACGTGTGCCG
CTGATTGGCTTTGCTGGTGCCCCATGGACCCTGATGACATACATGGTTGA
GGGTGGTGGCTCAAGCACCATGGCTCAGGCCAAGCGCTGGCTCTATCAGA
GACCTCAGGCTAGTCACCAGCTGCTTCGCATCCTCACTGATGCTCTGGTC
CCATATCTGGTAGGACAAGTGGTGGCTGGTGCCCAGGCATTGCAGCTGTT
TGAGTCCCATGCAGGGCATCTTGGCCCACAGCTCTTCAACAAGTTTGCAC
TGCCTTACATCCGTGATGTGGCCAAGCAAGTGAAGGCCAGGTTGCGGGAG
GCAGGCCTGGCACCAGTGCCCATGATCATCTTTGCTAAGGATGGGCATTT
TGCCCTGGAGGAGCTGGCCCAAGCTGGCTATGAGGTGGTTGGGCTTGACT
GGACAGTGGCCCCAAAGAAAGCCCGGGAGTGTGTGGGGAAGACGGTGACA
TTGCAGGGCAACCTGGACCCCTGTGCCTTGTATGCATCTGAGGAGGAGAT
CGGGCAGTTGGTGAAGCAGATGCTGGATGACTTTGGACCACATCGCTACA
TTGCCAACCTGGGCCATGGGCTTTATCCTGACATGGACCCAGAACATGTG
GGCGCCTTTGTGGATGCTGTGCATAAACACTCACGTCTGCTTCGACAGAA
CTGA
```

Comprobamos que ya no están los intrones.

1.3.10.3 Obtención de un multifasta

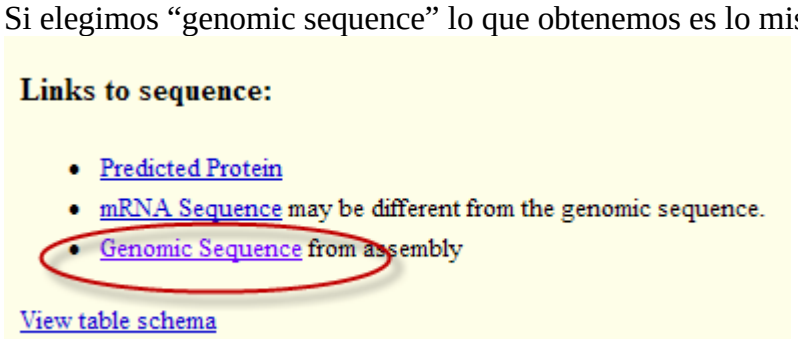
Si marcamos esta opción obtenemos un fichero multifasta (un fasta para cada exón)




```
>hg18_refGene_NM_000374_0 range=chr1:45250525-45250544 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGAAGCGAATGGGTTGGG
>hg18_refGene_NM_000374_1 range=chr1:45251166-45251278 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ACCTCAGGGTTTTCCGGAGCTGAAGAATGACACATTCCTGCGAGCAGCCT
GGGGAGAGGAAACAGACTACACTCCCGTTTGGTGCATGCGCCAGGCAGGC
```

```
CGTTACTTACCAG
>hg18_refGene_NM_000374_2 range=chr1:45251395-45251474 5'pad=0 3'pad=0 strand=+ repeatMasking=none
AGTTTAGGGAAACCCGGGCTGCCCAGGACTTTTTCAGCACGTGTCGCTCT
CCTGAGGCCTGCTGTGAACTGACTCTGCAG
>hg18_refGene_NM_000374_3 range=chr1:45251551-45251613 5'pad=0 3'pad=0 strand=+ repeatMasking=none
CCACTGCGTCGCTTCCCTCTGGATGCTGCCATCATTTTCTCCGACATCCT
TGTTGTACCCCAG
>hg18_refGene_NM_000374_4 range=chr1:45251853-45252050 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GCACTGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTCCC
AGAGCCATTAAGAGAAGAGCAGGACCTAGAACGCCTACGGGATCCAGAAG
TGGTAGCCTCTGAGCTAGGCTATGTGTTCCAAGCCATCACCCTTACCCGA
CAACGACTGGCTGGACGTGTGCCGCTGATTGGCTTTGCTGGTGCCCCA
>hg18_refGene_NM_000374_5 range=chr1:45252168-45252329 5'pad=0 3'pad=0 strand=+ repeatMasking=none
TGGACCCTGATGACATACATGGTTGAGGGTGGTGGCTCAAGCACCATGGC
TCAGGCCAAGCGCTGGCTCTATCAGAGACCTCAGGCTAGTCACCAGCTGC
TTCGCATCCTCACTGATGCTCTGGTCCCATATCTGGTAGGACAAGTGGTG
GCTGGTGCCCAG
>hg18_refGene_NM_000374_6 range=chr1:45252698-45252835 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GCATTGCAGCTGTTTGAGTCCCATGCAGGGCATCTTGGCCCACAGCTCTT
CAACAAGTTTGCACTGCCTTACATCCGTGATGTGGCCAAGCAAGTGAAGG
CCAGGTTGCGGGAGGCAGGCCTGGCACCAGTGCCCATG
>hg18_refGene_NM_000374_7 range=chr1:45252995-45253095 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATCATCTTTGCTAAGGATGGGCATTTTGCCCTGGAGGAGCTGGCCCAAGC
TGGCTATGAGGTGGTTGGGCTTGACTGGACAGTGGCCCCAAAGAAAGCCC
G
>hg18_refGene_NM_000374_8 range=chr1:45253199-45253265 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GGAGTGTGTGGGGAAGACGGTGACATTGCAGGGCAACCTGGACCCCTGTG
CCTTGTATGCATCTGAG
>hg18_refGene_NM_000374_9 range=chr1:45253596-45253757 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GAGGAGATCGGGCAGTTGGTGAAGCAGATGCTGGATGACTTTGGACCACA
TCGCTACATTGCCAACCTGGGCCATGGGCTTTATCCTGACATGGACCCAG
AACATGTGGGCGCCTTTGTGGATGCTGTGCATAAACACTCACGTCTGCTT
CGACAGAACTGA
```

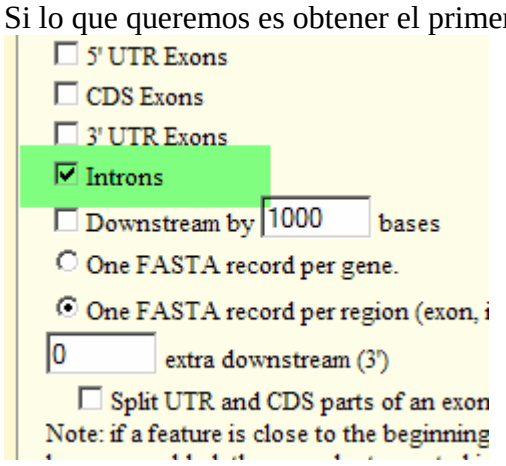
Si elegimos “genomic sequence” lo que obtenemos es lo mismo (aunque se nos advierte que puede no coincidir):




```
>hg18_refGene_NM_000374_0 range=chr1:45250525-45250544 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGGAAGCGAATGGGTTGGG
>hg18_refGene_NM_000374_1 range=chr1:45251166-45251278 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ACCTCAGGGTTTTCCGGAGCTGAAGAATGACACATTCCTGCGAGCAGCCT
GGGGAGAGGAAACAGACTACACTCCCGTTTGGTGCATGCGCCAGGCAGGC
CGTTACTTACCAG
>hg18_refGene_NM_000374_2 range=chr1:45251395-45251474 5'pad=0 3'pad=0 strand=+ repeatMasking=none
AGTTTAGGGAAACCCGGGCTGCCCAGGACTTTTTCAGCACGTGTCGCTCT
CCTGAGGCCTGCTGTGAACTGACTCTGCAG
>hg18_refGene_NM_000374_3 range=chr1:45251551-45251613 5'pad=0 3'pad=0 strand=+ repeatMasking=none
CCACTGCGTCGCTTCCCTCTGGATGCTGCCATCATTTTCTCCGACATCCT
TGTTGTACCCCAG
>hg18_refGene_NM_000374_4 range=chr1:45251853-45252050 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GCACTGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTCCC
AGAGCCATTAAGAGAAGAGCAGGACCTAGAACGCCTACGGGATCCAGAAG
TGGTAGCCTCTGAGCTAGGCTATGTGTTCCAAGCCATCACCCTTACCCGA
CAACGACTGGCTGGACGTGTGCCGCTGATTGGCTTTGCTGGTGCCCCA
>hg18_refGene_NM_000374_5 range=chr1:45252168-45252329 5'pad=0 3'pad=0 strand=+ repeatMasking=none
TGGACCCTGATGACATACATGGTTGAGGGTGGTGGCTCAAGCACCATGGC
TCAGGCCAAGCGCTGGCTCTATCAGAGACCTCAGGCTAGTCACCAGCTGC
TTCGCATCCTCACTGATGCTCTGGTCCCATATCTGGTAGGACAAGTGGTG
GCTGGTGCCCAG
>hg18_refGene_NM_000374_6 range=chr1:45252698-45252835 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GCATTGCAGCTGTTTGAGTCCCATGCAGGGCATCTTGGCCCACAGCTCTT
CAACAAGTTTGCACTGCCTTACATCCGTGATGTGGCCAAGCAAGTGAAGG
CCAGGTTGCGGGAGGCAGGCCTGGCACCAGTGCCCATG
>hg18_refGene_NM_000374_7 range=chr1:45252995-45253095 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATCATCTTTGCTAAGGATGGGCATTTTGCCCTGGAGGAGCTGGCCCAAGC
TGGCTATGAGGTGGTTGGGCTTGACTGGACAGTGGCCCCAAAGAAAGCCC
G
>hg18_refGene_NM_000374_8 range=chr1:45253199-45253265 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GGAGTGTGTGGGGAAGACGGTGACATTGCAGGGCAACCTGGACCCCTGTG
CCTTGTATGCATCTGAG
>hg18_refGene_NM_000374_9 range=chr1:45253596-45253757 5'pad=0 3'pad=0 strand=+ repeatMasking=none
GAGGAGATCGGGCAGTTGGTGAAGCAGATGCTGGATGACTTTGGACCACA
TCGCTACATTGCCAACCTGGGCCATGGGCTTTATCCTGACATGGACCCAG
AACATGTGGGCGCCTTTGTGGATGCTGTGCATAAACACTCACGTCTGCTT
CGACAGAACTGA
```

1.3.11 Repetidlo con el primer intrón.

Si lo que queremos es obtener el primer intron. Hay que seleccionar



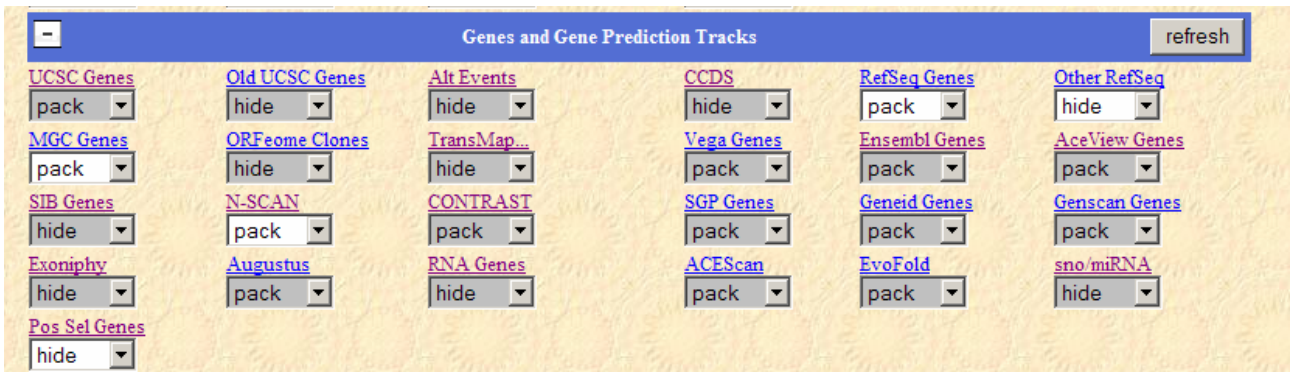
Y asi obtenemos un multifasta con los introne (lo cortamos y pegamos). Señalado en amarillo

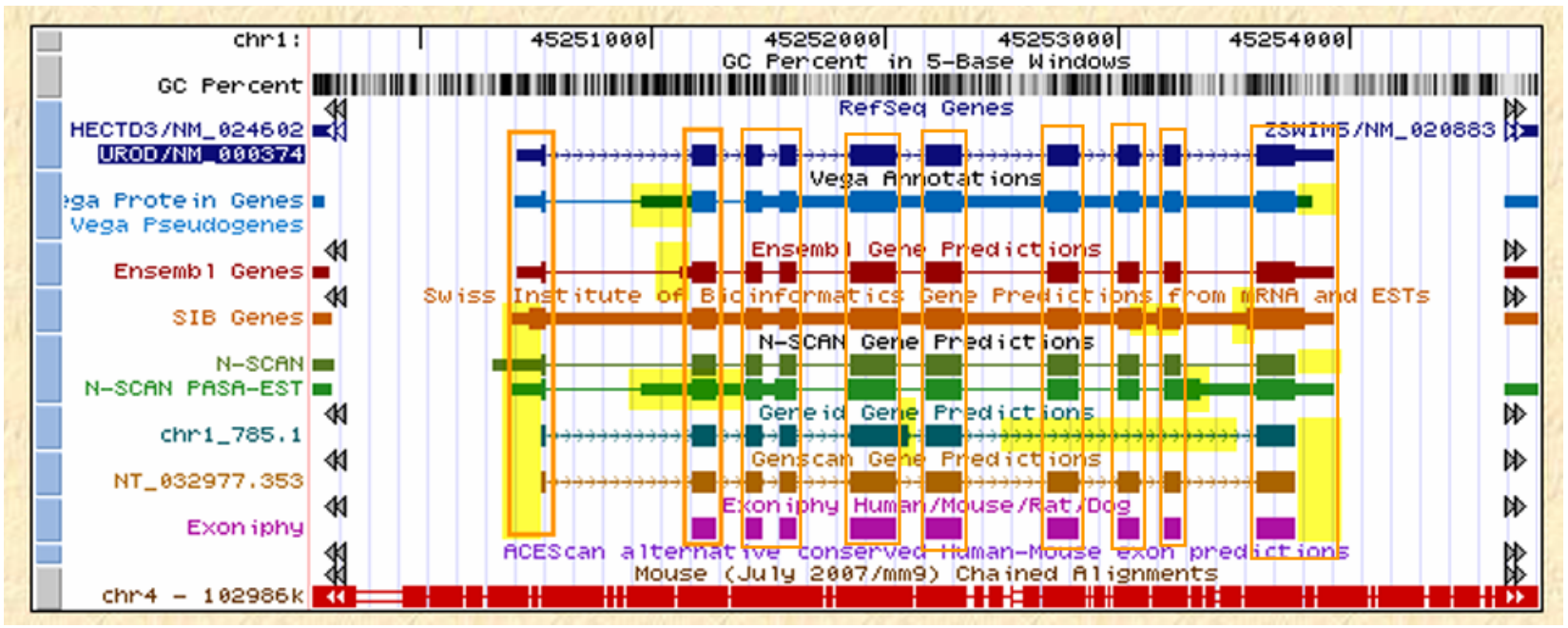
```
>hg18_refGene_NM_000374_0 range=chr1:45250545-45251165 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtgagttctccagagcacgcggtgtggctagccgggcttctaatttgagt
cttccaactcaggactctatccctctactccccttttccccaccctggaga
cctcccaacctgaactccgttagctgggatcctgaatcctaaaaccatgg
atttttgagatgttcatcccagggccttaattcaagggatgcctcaggat
ttccaaccaggatctctattctgggaccatcaactctgatccctctttat
cccccagcctgggtatttctcagcccctgaaccagcccagtgacattccc
gtttctgaggctcactagttcgaagacccccaaactatccttagtgggcc
ttcattccctccccccagtccctctggttgcttcgagcttggaagagtag
agactaagtggagggaagaggccccagggcgggcccttctggagtttgtg
caccacctgataggcagagaggaggcggaacgggcggaaagccagggttt
gggagctggcctggaggaggtagatagcggtcctggactgaatcggcctt
atgaacccgcgctttccccagcctccagcgtagcatactgacacctaccc
ccacccccacctgatcgccag
>hg18_refGene_NM_000374_1 range=chr1:45251279-45251394 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtaagagtcagggtctggaaatctagataaaactccggagggagaaaagt
tttcgaggggcaggggagggctctggagggcctcaaggctgagccctgtc
ttccctctgtatgcag
>hg18_refGene_NM_000374_2 range=chr1:45251475-45251550 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtgaggggtccacaaaagagggaaagatttatgccttcagtctgccacct
agcaacctgtctcctgtttcctacag
>hg18_refGene_NM_000374_3 range=chr1:45251614-45251852 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtacccactcaaacctgatcctagaatataatccaaggacgccttgaaaa
tccttctatcagtccagtcaaggtttacaataagcacttatcctaactgg
atcgagggaaaaactaaggttgaaagaaatggagtttggcagagttttat
tctccttttccttcctcctggaatgagctgaacagaacctttcctcctgg
attccattttgggaacccagatgttttctccccctccag
>hg18_refGene_NM_000374_4 range=chr1:45252051-45252167 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtaatgtgggacagggcagggactcggggcgcggggagatcactctggaa
ggtctggggtagacaaaaggaagggtcagtctggcttctgtgacaccatc
tttctatccttctctag
>hg18_refGene_NM_000374_5 range=chr1:45252330-45252697 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtgagtcctgagagagagagaaataggctgggatttggtctgtaaggacc
agaagcaagagtgtcctaaacctgagagggcaggggtcttaatgccaggg
atgaaagaaccttggcctccagtgatctagcggagcagccaagcccatcc
tgacactgacagtggggcttaatgctctaagtattcagacaccaaagtta
gtgctgggatctgaggaaagtaaatttttttttttttaattactgggttt
ttagggtcaggcagtatcagggattgaagtcatttggggaaaattgaggt
ggattttgtatgtgggggaaacttcctctttgtgtgttacatatttttct
tcaccataccctaactag
>hg18_refGene_NM_000374_6 range=chr1:45252836-45252994 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtgaggattgggatgggttgagtgaaggtggtcctgtggagctttcaggc
taagtcctgcatggactggagtgaccactggagggcagcagaagtacagt
caagaaagattagtggttgtagcaaggccctctgtagcctgagatctgct
tttttctag
>hg18_refGene_NM_000374_7 range=chr1:45253096-45253198 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtaagccatggaagggtgaggccttgaggttgaggtgggggtgttggctg
ggggagctgccatgtatgcagttaccagaacgtggcgctggctttgcttc
cag
>hg18_refGene_NM_000374_8 range=chr1:45253266-45253595 5'pad=0 3'pad=0 strand=+ repeatMasking=none
gtaacagccagggcccctctgtgtgtctgttactgtgcactcctgtggct
gtggctgtattattctgtgtgcacttgtttttaatgtctgtctgtccttt
tcttctactctgtacaacataagccctagaaagaccggacttttgttgc
tgttgttcatttgtgtttatgcttcatgcctgggtccatactagggatct
gataaattttattgaatgactgaataacactgagtagaagcatgcctaca
tatgcgtttgtcactagtatatatagggaggacaaaggcttgctggtcct
cctgtagccagtgccctgttggtccccccag
```

Si el fichero es muy grande con muchos introne y exones entonces hay que localizar sus coordenadas y usar el TABLE BROWNSER

Nuestro filtro seria range=chr1:45250545-45251165

1.3.12 Explorad el bloque de opciones Genes and Gene prediction Tracks. Activad todos los programas de predicción de exones (genes). Analizad si las predicciones coinciden con las anotaciones reales (pista REFSEQ Genes).





En la figura se puede apreciar que ninguno de los probados ofrece un 100% de fiabilidad. (solo he mostrado la imagen de unos cuantos, sino uno se pierde con tantas cajas… ¡yo al menos!.)

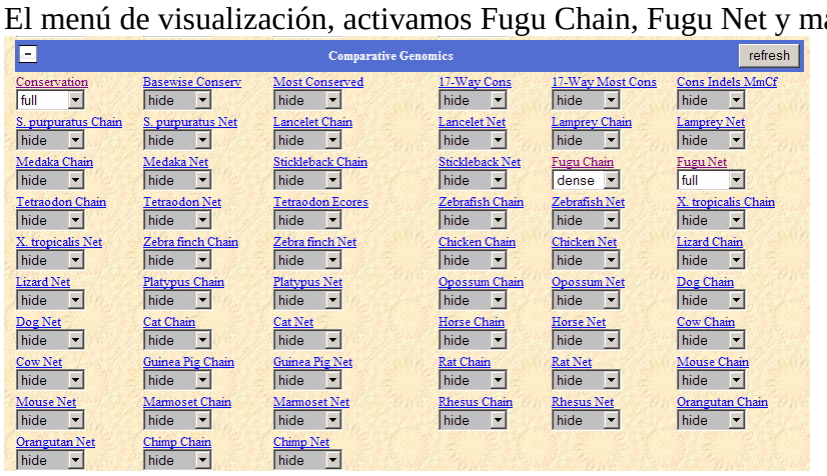
En el caso de Geneid vemos que faltan 3 exones??.
Si ejecutamos geneid con una secuencia de UROD obtenemos:

```
Internal        1710    1860    -0.11   +       0       gi|1322018|gb|U30787.1|HSU30787_1
Internal        1976    2055     0.24   +       2       gi|1322018|gb|U30787.1|HSU30787_1
Internal        2132    2194     0.44   +       0       gi|1322018|gb|U30787.1|HSU30787_1
Internal        2434    2682     4.66   +       0       gi|1322018|gb|U30787.1|HSU30787_1
Internal        2749    2910     3.19   +       0       gi|1322018|gb|U30787.1|HSU30787_1
Internal        3279    3416     0.97   +       0       gi|1322018|gb|U30787.1|HSU30787_1
Internal        3576    3676     3.23   +       0       gi|1322018|gb|U30787.1|HSU30787_1
Internal        3780    3846    -0.96   +       1       gi|1322018|gb|U30787.1|HSU30787_1
Terminal        4179    4340     4.55   +       0       gi|1322018|gb|U30787.1|HSU30787_1
```

Exactamente nueve exones …. ¿Diferente version?
En el resto están señalados en amarillo sobre la imagen las predicciones que no coinciden con los genes reales. Hay que hacer notar que algunos de los programas de predicción ofrecen varias predicciones y otros se basan en una combinación de predicciones ab-initio y búsqueda de homologos. Parece que la predición más acertada es de Genscan.

1.3.13 Explorad las opciones en Comparative genomics. Activad la opción Fugu chain y explicad informalmente qué contiene esta pista y cómo se obtiene.

El menú de visualización, activamos Fugu Chain, Fugu Net y marcamos conservation a Full



Básicamente hay tres bloques de opciones:

- Resúmenes de agrupaciones de distintas organismos modelo

Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)

This track shows multiple alignments of 28 vertebrate species and two measures of evolutionary conservation -- conservation across all 28 species and an alternative measurement restricted to the placental mammal subset (17 species plus human) of the alignment. These two measurements produce the same results in regions where only mammals appear in the alignment. For other regions, the non-mammalian species can either boost the scores (if conserved) or decrease them (if non-conserved). The mammalian conservation helps to identify sequences that are under different evolutionary pressures in mammalian and non-mammalian vertebrates.

Basewise Conservation by PhyloP for 28-Species Multiz Align.

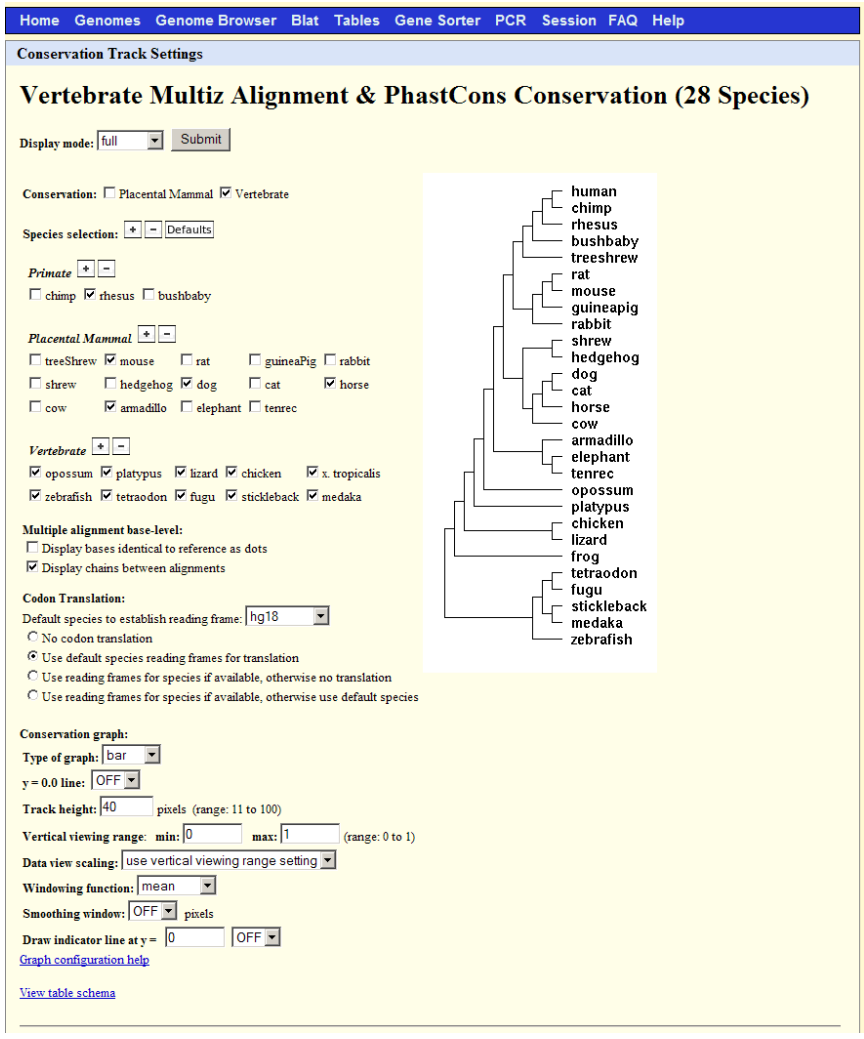
This track shows measures of evolutionary conservation generated using the *phyloP* (Phylogenetic P-Values) program from the PHAST package. Two measurements are provided: conservation across 28 species, and an alternative measurement restricted to the placental mammal subset (17 species plus human) of the multiple alignment.

PhastCons Conserved Elements, 28-way Vertebrate Multiz Alignment

This track shows predictions of conserved elements produced by the phastCons program based on a whole-genome alignment of vertebrates, and for the placental mammal subset of species in the alignment. They are based on a phylogenetic hidden Markov model (phylo-HMM), a type of probabilistic model that describes both the process of DNA substitution at each site in a genome and the way this process changes from one site to the next.

17-Way Cons Track Settings - Vertebrate Multiz Alignment & Conservation (17 Species)

This track shows a measure of evolutionary conservation in 17 vertebrates, including mammalian, amphibian, bird, and fish species, based on a phylogenetic hidden Markov model, phastCons (Siepel *et al.*, 2005). Multiz alignments of the following assemblies were used to generate this track:



17-Way Most Cons Track Settings- -PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment

This track shows predictions of conserved elements produced by the phastCons program. PhastCons is part of the PHAST (PHylogenetic Analysis with Space/Time models) package. The predictions are based on a phylogenetic hidden Markov model (phylo-HMM), a type of probabilistic model that describes both the process of DNA substitution at each site in a genome and the way this process changes from one site to the next.

Cons Indels MmCf Track Settings- Indel-based Conservation for human hg18, mouse mm8 and dog canFam2

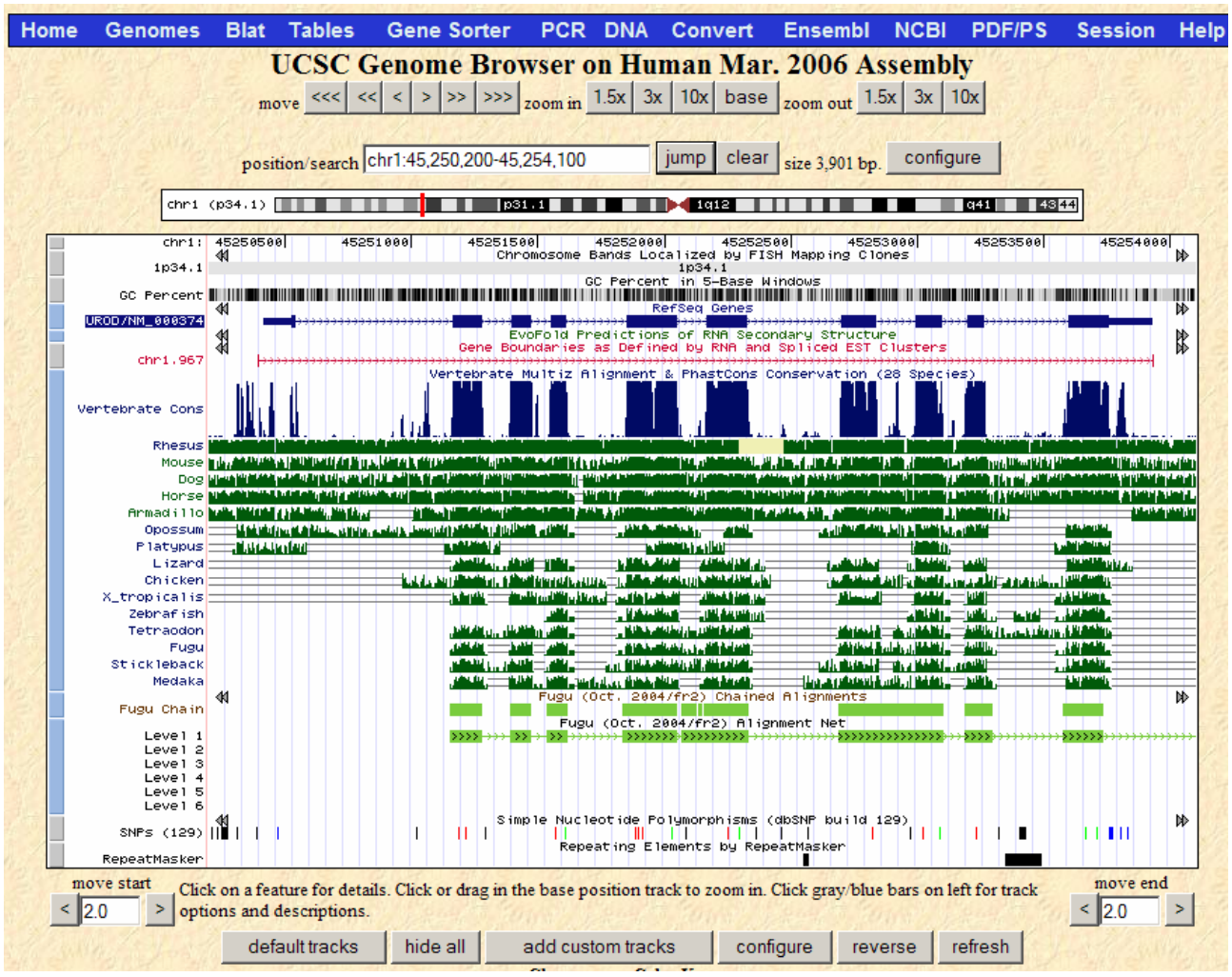
This track displays regions showing evidence for conservation with respect to mutations involving sequence insertions and deletions (indels). These “indel-purified sequences” (IPSs) were obtained by comparing the predictions of a neutral model of indel evolution with data obtained from human (hg18), mouse (mm8) and dog (camFam2) alignments (Lunter *et al*., 2006) The evidence for conservation is statistical, and each region is annotated with a posterior probability. It may be interpreted as the probability that the segment shows the paucity of indels by selection, rather than by random chance. Apart from the underlying alignment, these data are independent of the conservation of the nucleotide sequence itself. Any inferred conservation of the sequence, e.g. as shown by phastCons, is therefore independent evidence for selection. It may happen that sequence is conserved with respect to indel mutations without concomitant evidence of conservation of the nucleotide sequence. The opposite may also happen.

- Visualizaciones de los alineamientos “Chain” para distintos organismos
- Visualizaciones de los alineamientos “net” para los mismos organismos que “Chain”

1.3.13.1 Que son los alineamientos Chain y Net

http://genomewiki.ucsc.edu/index.php/Chains_Nets
Chains and nets are Jim Kent's brainchild, published here: [http://www.pnas.org/cgi/content/full/100/20/11484] They are generated from genomic local alignments computed by Blastz. They used to be generated by a long manual process documented in some of our older makeDb/doc/*.txt files, but are now generated by the script kent/src/hg/utils/automation/doBlastzChainNet.pl .
Here are some musings on the fine points of chains and nets -- these are from Angie's mental model of chains and nets and represent opinions which may be outdated or plain old incorrect. The source code, and the results that we get by running these programs on real data, are the ultimate source of truth about chains and nets.
Chains in a nutshell:
a chain is a sequence of gapless aligned blocks, where there must be no overlaps of blocks' target or query coords within the chain. Within a chain, target and query coords are monotonically non-decreasing. (i.e. always increasing or flat)
double-sided gaps are a new capability (blastz can't do that) that allow extremely long chains to be constructed.
not just orthologs, but paralogs too, can result in good chains. but that's useful!
chains should be symmetrical -- e.g. swap human-mouse -> mouse-human chains, and you should get approx. the same chains as if you chain swapped mouse-human blastz alignments. However, Blastz's dynamic masking is asymmetrical, so in practice those results are not exactly symmetrical. Also, dynamic masking in conjunction with changed chunk sizes can cause differences in results from one run to the next.
chained blastz alignments are not single-coverage in either target or query unless some subsequent filtering (like netting) is done.
chain tracks can contain massive pileups when a piece of the target aligns well to many places in the query. Common causes of this include insufficient masking of repeats and high-copy-number genes (or paralogs).
And nets:
a net is a hierarchical collection of chains, with the highest-scoring non-overlapping chains on top, and their gaps filled in where possible by lower-scoring chains, for several levels. I think a chain's qName also helps to determine which level it lands in, i.e. it makes a difference whether a chain's qName is the same as the top-level chain's qName or not, because the levels have meanings associated with them -- see details page.
a net is single-coverage for target but not for query.
because it's single-coverage in the target, it's no longer symmetrical.
the netter has two outputs, one of which we usually ignore: the target-centric net in query coordinates. The reciprocal best process uses that output: the query-referenced (but target-centric / target single-cov) net is turned back into component chains, and then those are netted to get single coverage in the query too; the two outputs of that netting are reciprocal-best in query and target coords. Reciprocal-best nets are symmetrical again.
nets do a good job of filtering out massive pileups by collapsing them down to (usually) a single level.
"LiftOver chains" are actually chains extracted from nets, or chains filtered by the netting process. Same-species liftOver chains use blat -fastMap as the alignment method, and are generated by kent/src/hg/utils/automation/doSameSpeciesLiftOver.pl, based on a series of scripts that Kate wrote in kent/src/hg/makeDb/makeLoChain/. Cross-species liftOver chains are generated by doBlastzChainNet.pl.
Navigation: back to Implementation_Notes
Retrieved from "http://genomewiki.ucsc.edu/index.php/Chains_Nets"

La pista chain muestra cajas separadas o unidas por simples o dobles líneas. Las cajas encuadran las regiones alineadas. Las líneas simples indican huecos que son principalmente debido a una deleción o una inserción en la cadena. Las líneas dobles representan huecos más complejos posiblemente resultado de inversiones, solapamiento, deleciones, o mutaciones acumuladas, o una parte del genoma no secuenciada en una de los genomas comparados.



Desde el detalla del alineamiento, podemos abrir otro brownser al organismo comparado.

Home Genomes Genome Browser Blat Tables Gene Sorter PC

Fugu (Oct. 2004/fr2) Chained Alignments (1168)

Human position: chr1:45251152-45326996 size: 75845
Strand: +
Fugu position: chrUn:181808351-181829431 size: 21081
Chain ID: 1168
Score: 197786 **Approximate Score within browser window:** 49918

Fields above refer to entire chain or gap, not just the part inside the window.

View details of parts of chain within browser window.
Open Fugu browser at position corresponding to the part of chain that is in this window.

View table schem

Go to Fugu Chain track controls

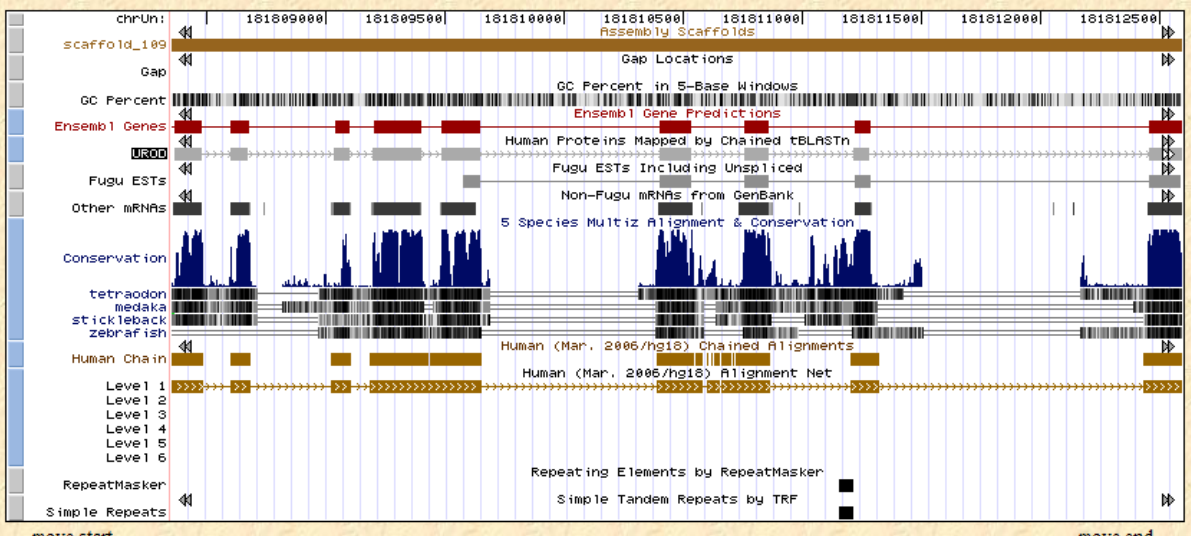
Data last updated: 2007-01-24

Home Genomes Blat Tables PCR DNA Convert PDF/PS Session Help

UCSC Genome Browser on Fugu Oct. 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chrUn:181,808,351-181,812,591 jump clear size 4,241 bp. configure



move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click gray/blue bars on left for track options and descriptions. move end < 2.0 >

default tracks hide all add custom tracks configure reverse refresh

Chromosome Color Key:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y M Un

collapse all — Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. — expand all

Mapping and Sequencing Tracks refresh

Base Position	Scaffolds	Gap	GC Percent	Short Match	Restr Enzymes
full	pack	dense	dense	hide	hide

Genes and Gene Prediction Tracks refresh

TransMap...	Ensembl Genes	Human Proteins
hide	dense	pack

mRNA and EST Tracks refresh

Fugu mRNAs	Spliced ESTs	Fugu ESTs	Other mRNAs
hide	hide	dense	dense

Comparative Genomics refresh

Conservation	Most Conserved	Human Chain	Human Net	Chicken Chain	Chicken Net
pack	hide	dense	full	hide	hide

Mouse Chain	Mouse Net	Zebrafish Chain	Zebrafish Net	Stickleback Chain	Stickleback Net
hide	hide	hide	hide	hide	hide

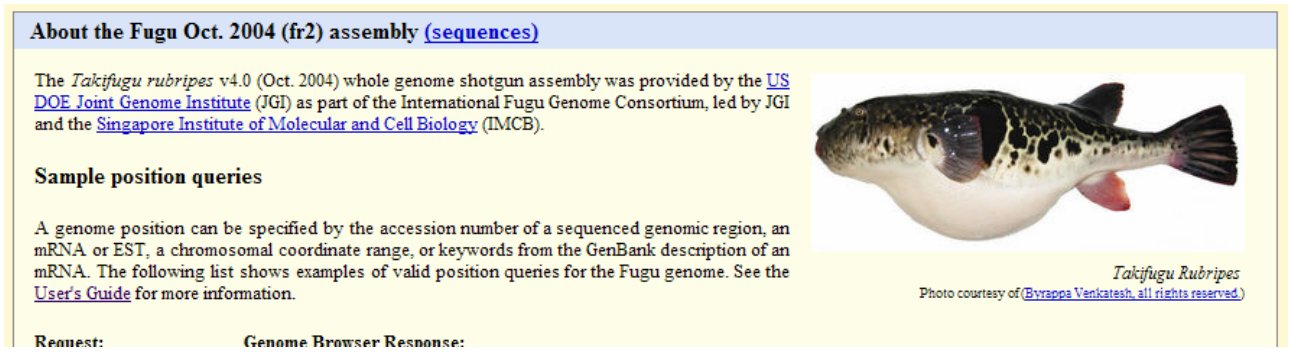
Medaka Chain	Medaka Net	Tetraodon Chain	Tetraodon Net
hide	hide	hide	hide

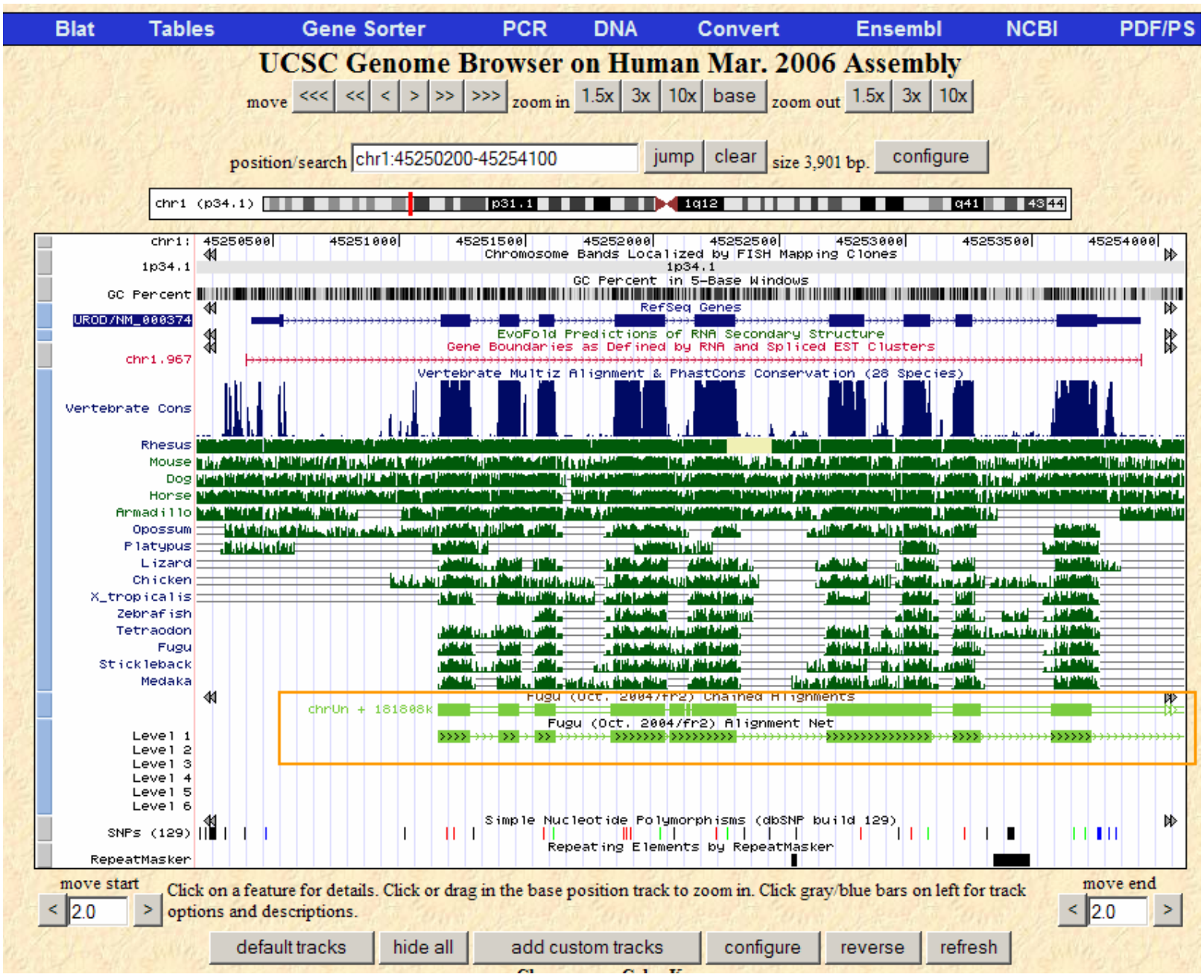
Variation and Repeats refresh

RepeatMasker	WM + SDust	Simple Repeats
dense	hide	dense

refresh

1.3.13.2 Quien es el Fugu?. Afable pececillo

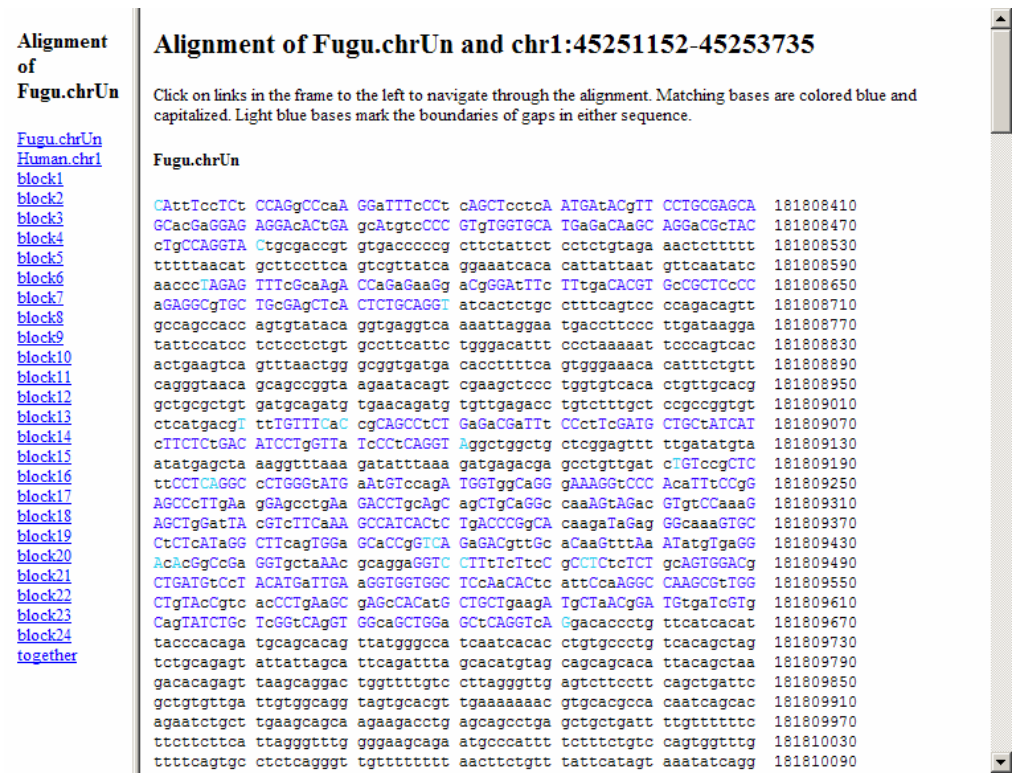




1.3.13.3 Alineamiento Fugu Chain

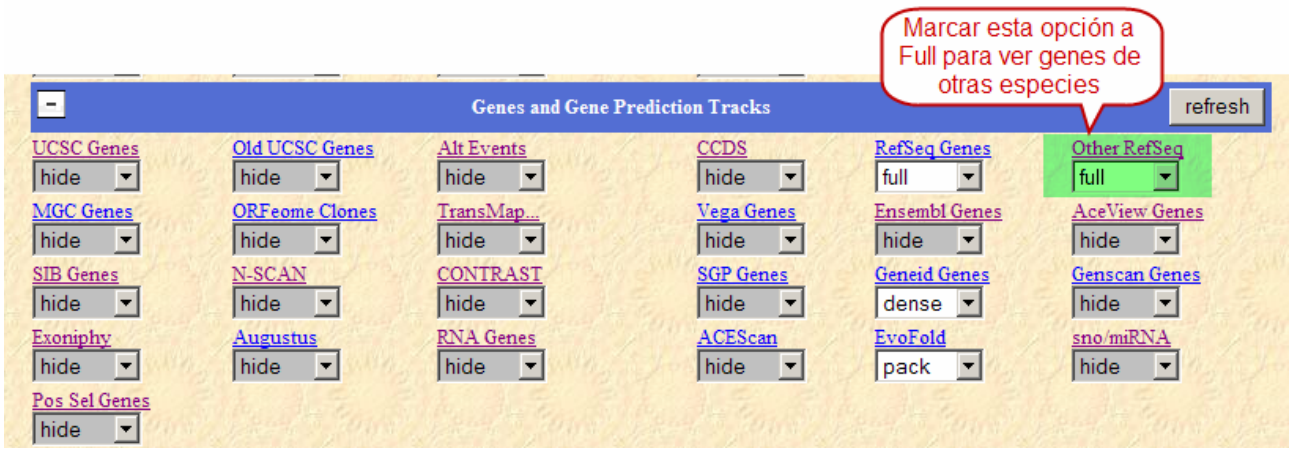
La pista chain muestra cajas separadas o unidas por simples o dobles líneas. Las cajas encuadran las regiones alineadas. Las líneas simples indican huecos que son principalmente debido a una deleción o una inserción en la cadena. Las líneas dobles representan huecos más complejos posiblemente resultado de inversiones, solapamiento, deleciones, o mutaciones acumuladas, o una parte del genoma no secuenciada en una de los genomas comparados. En nuestro caso parece indicar que es un resultado evolutivo complejo

Navegando entre las opciones podemos visualizar el alineamiento.

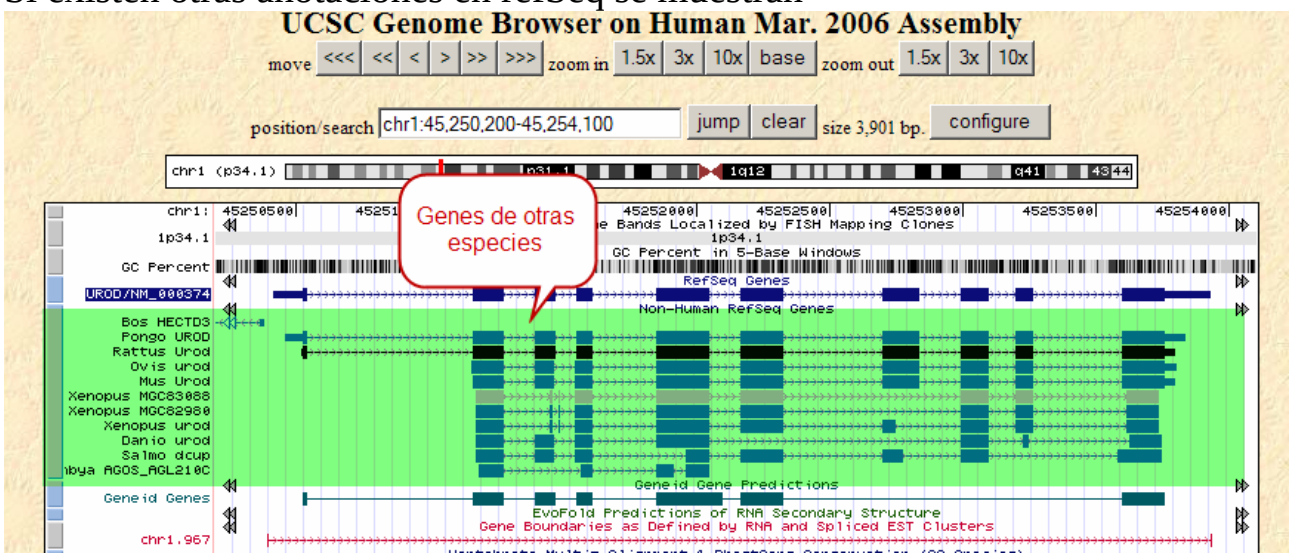


1.3.13.4 Alineamiento Fugu Net

1.3.14 Buscad la opción para visualizar las anotaciones disponibles del gen UROD en otras especies.

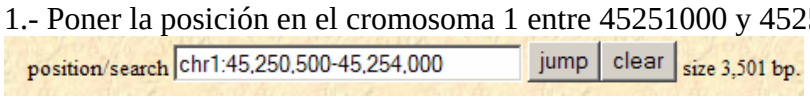


Si existen otras anotaciones en refSeq se muestran

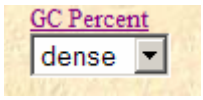


1.3.15 Finalmente, debéis intentar reproducir exactamente la figura que se incluye en este enunciado. Enumerad una por una cada opción/pista activada y qué significado tiene.

Para reproducir más o menos la imagen hay que:
1.- Poner la posición en el cromosoma 1 entre 45251000 y 452535000, que corresponde a chr1:45,250,500-45,254,000



2.- Mostrar la banda de contenido porcentual de CG (CG percent)

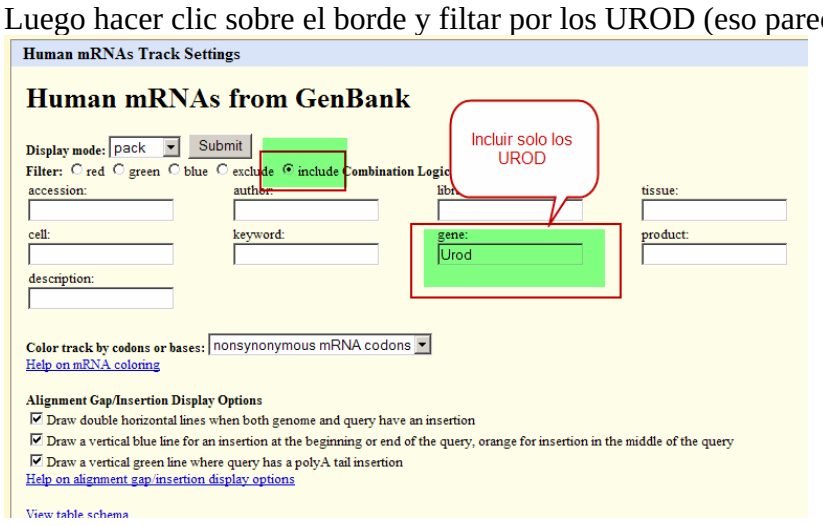


3.- Mostar los genes RefSeq.

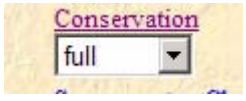
4.- Mostar los other Refseq. (entradas de otros organismos registradas en refSeq)

5.- Mostrar los human mRNAs del genbank. Muestra las entradas de RNA mensajeros del genbank

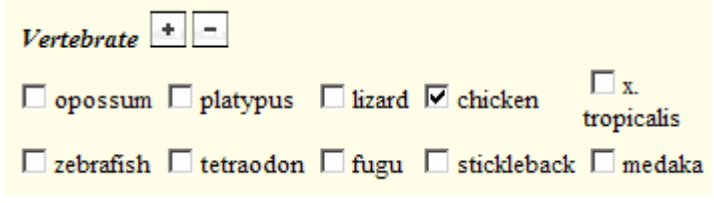
Luego hacer clic sobre el borde y filtar por los UROD (eso parece en el la imagen)



6.- Conservation a FULL. Para mostrar los gráficos para cada especie.



Pulsamos en el borde izquierdo y deseleccionamos los que no están en la imagen :
(dejamos estos)



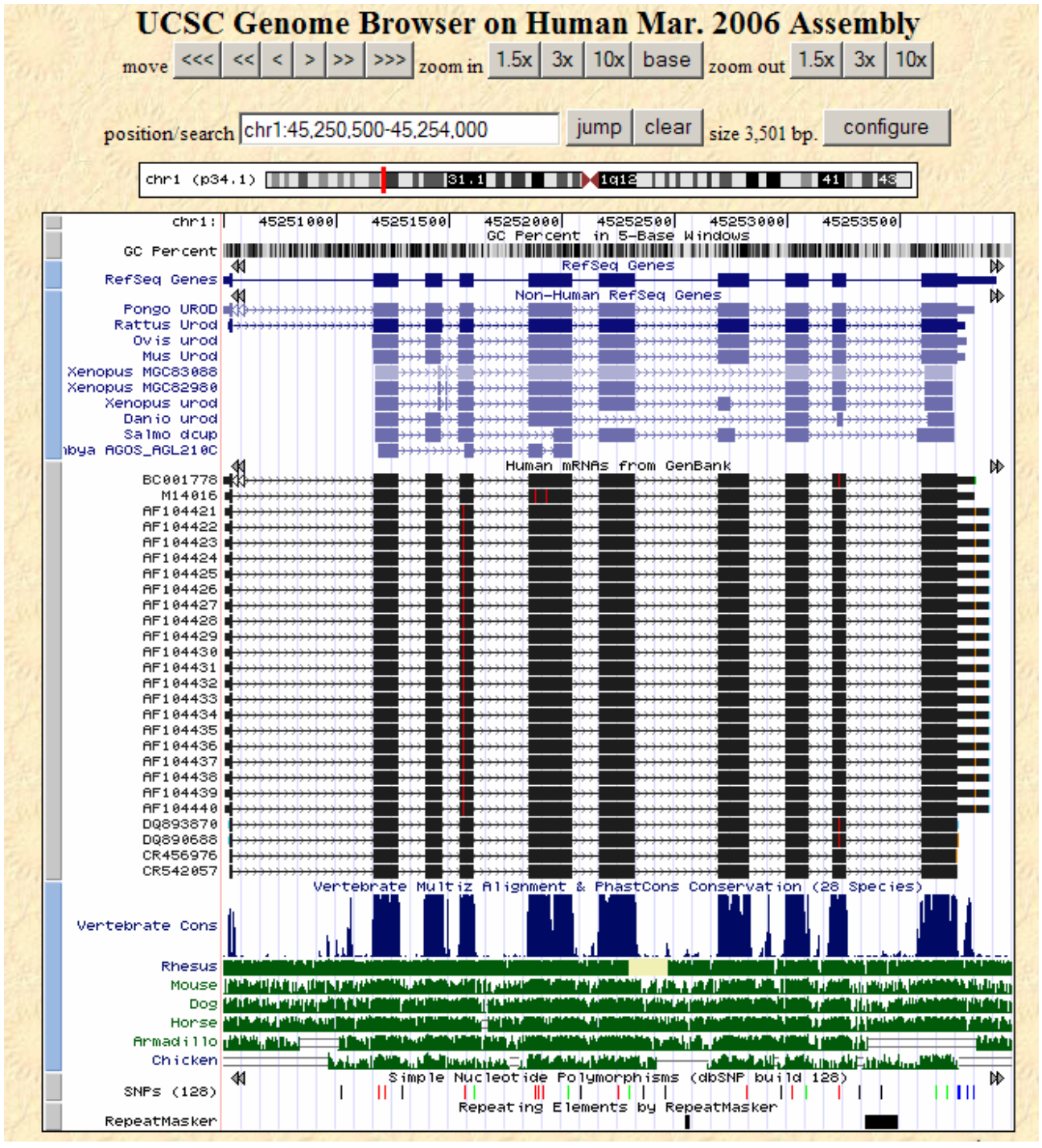
7.- SNPs 128 (cambiamos esta por defecto 129)

http://es.wikipedia.org/wiki/Polimorfismo_de_nucle%C3%B3tido_simple
Los SNP en la wiki:
SNP (Single Nucleotide Polymorphism, pronunciado esnip) es una variación en la secuencia de ADN que afecta a una sola base (adenina (A), timina (T), citosina (C) o guanina (G)) de un a secuencia del genoma. Sin embargo, algunos autores consideran que cambios de unos pocos nucleotidos, como también pequeñas inserciones y deleciones pueden ser consideradas como SNP, donde el término Polimorfismo de nucleótido simple es más adecuado.[1] Una de estas variaciones debe darse al menos en un 1% de la población para ser considerada como un SNP.
Los SNP son de gran utilidad para la investigación médica en el desarrollo de fármacos. Debido a que los SNP no cambian mucho de una generación a otra, es sencillo seguir su evolución en estudios de poblaciones. También se utilizan en algunos tipos de pruebas genéticas.
Los SNP se consideran una forma de mutación puntual que ha sido lo suficientemente exitosa evolutivamente para fijarse en una parte significativa de la población de una especie.

8.- Repeater master. Busca features que estan en distintas librerias y los anota en la banda.

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). On average, almost 50% of a human genomic DNA sequence currently will be masked by the program. Sequence comparisons in RepeatMasker are performed by the program cross_match, an efficient implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green.

Al final queda tal y como esta:

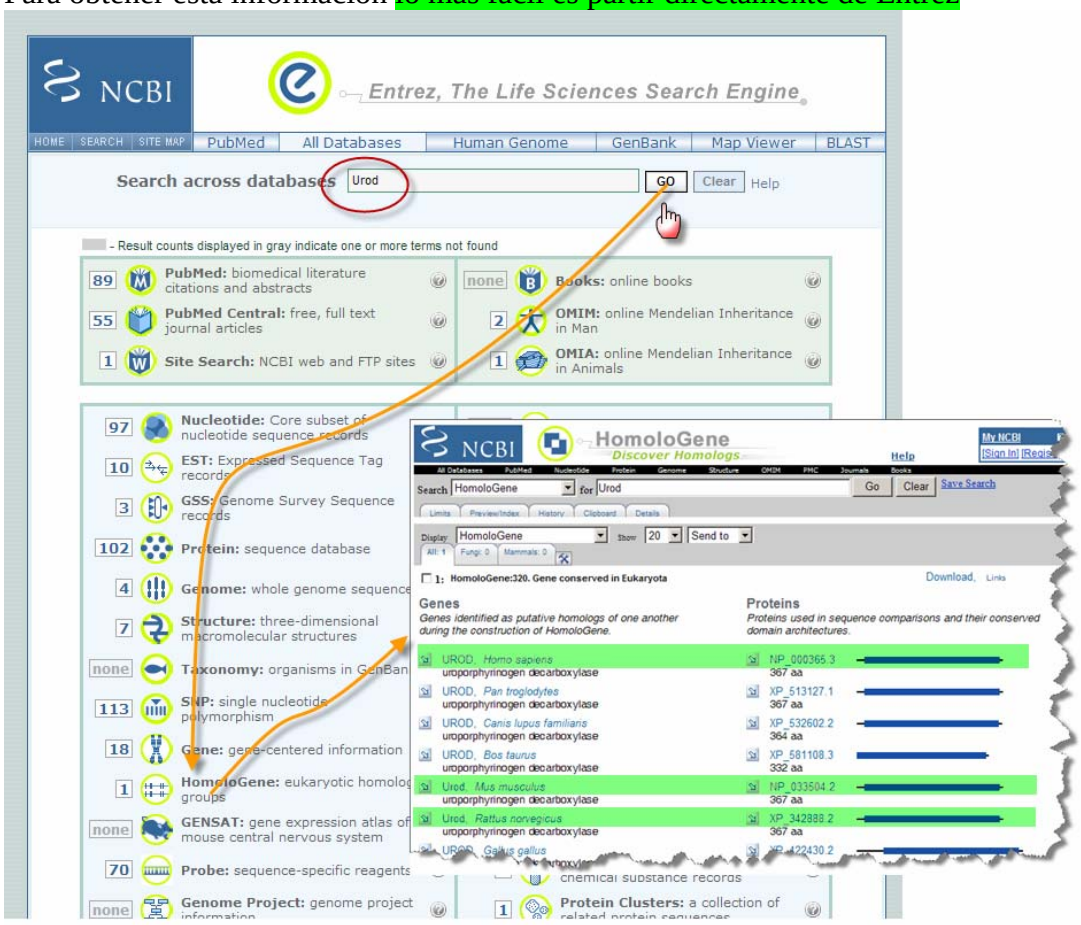


1.4 EJERCICIO 3

1.4.1 Repetid la búsqueda del gen UROD ahora en ratón (*Mus musculus*). Comentad brevemente cuáles son las diferencias más relevantes (si hay) entre el gen anotado en humano y el mismo gen anotado ahora en ratón.

Aspectos importantes a comparar: número de exones, longitud del gen, longitud de la proteína codificada, su secuencia y otras características sencillas. Recordad que para acceder a los datos del genoma de ratón se debe volver a la página principal del Genome browser. Si todo va bien, podéis probar el mismo experimento con el mismo gen en la rata (Rattus norvegicus).

Para obtener esta información lo más fácil es partir directamente de Entrez



1.4.1.1 Información recopilada de Homo Sapiens

```
LOCUS       NM_000374               1383 bp    mRNA    linear   PRI 21-DEC-2008
DEFINITION  Homo sapiens uroporphyrinogen decarboxylase (UROD), mRNA.
ACCESSION   NM_000374
VERSION     NM_000374.3  GI:71051615
KEYWORDS    .
SOURCE      Homo sapiens (human)
     source          1..1383
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="1"
                     /map="1p34"
     gene            1..1383
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /note="uroporphyrinogen decarboxylase"
                     /db_xref="GeneID:7389"
                     /db_xref="HGNC:12591"
                     /db_xref="HPRD:01441"
                     /db_xref="MIM:176100"
     CDS             109..1212
                     /gene="UROD"
                     /gene_synonym="PCT"
                     /EC_number="4.1.1.37"
                     /note="uroporphyrinogen III decarboxylase; fifth enzyme of
                     the heme biosynthetic pathway; fifth enzyme of heme
                     biosynthetic pathway"
                     /codon_start=1
                     /product="uroporphyrinogen decarboxylase"
                     /protein_id="NP_000365.3"
                     /db_xref="GI:71051616"
                     /db_xref="CCDS:CCDS518.1"
                     /db_xref="GeneID:7389"
                     /db_xref="HGNC:12591"
                     /db_xref="HPRD:01441"
                     /db_xref="MIM:176100"
                     /translation="MEANGLGPQGFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLP
                     EFRETRAAQDFFSTCRSPEACCELTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVP
                     GKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRVPLIGFAGAPWT
                     LMTYMVEGGGSSTMAQAKRWLYQRPQASHQLLRILTDALVPYLVGQVVAGAQALQLFE
```

```
                     SHAGHLGPQLFNKFALPYIRDVAKQVKARLREAGLAPVPMIIFAKDGHFALEELAQAG
                     YEVVGLDWTVAPKKARECVGKTVTLQGNLDPCALYASEEEIGQLVKQMLDDFGPHRYI
                     ANLGHGLYPDMDPEHVGAFVDAVHKHSRLLRQN"

>gi|71051615:109-1212 Homo sapiens uroporphyrinogen decarboxylase (UROD), mRNA
ATGGAAGCGAATGGGTTGGGACCTCAGGGTTTTCCGGAGCTGAAGAATGACACATTCCTGCGAGCAGCCT
GGGGAGAGGAAACAGACTACACTCCCGTTTGGTGCATGCGCCAGGCAGGCCGTTACTTACCAGAGTTTAG
GGAAACCCGGGCTGCCCAGGACTTTTTCAGCACGTGTCGCTCTCCTGAGGCCTGCTGTGAACTGACTCTG
CAGCCACTGCGTCGCTTCCCTCTGGATGCTGCCATCATTTTCTCCGACATCCTTGTTGTACCCCAGGCAC
TGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTCCCAGAGCCATTAAGAGAAGAGCAGGA
CCTAGAACGCCTACGGGATCCAGAAGTGGTAGCCTCTGAGCTAGGCTATGTGTTCCAAGCCATCACCCTT
ACCCGACAACGACTGGCTGGACGTGTGCCGCTGATTGGCTTTGCTGGTGCCCCATGGACCCTGATGACAT
ACATGGTTGAGGGTGGTGGCTCAAGCACCATGGCTCAGGCCAAGCGCTGGCTCTATCAGAGACCTCAGGC
TAGTCACCAGCTGCTTCGCATCCTCACTGATGCTCTGGTCCCATATCTGGTAGGACAAGTGGTGGCTGGT
GCCCAGGCATTGCAGCTGTTTGAGTCCCATGCAGGGCATCTTGGCCCACAGCTCTTCAACAAGTTTGCAC
TGCCTTACATCCGTGATGTGGCCAAGCAAGTGAAGGCCAGGTTGCGGGAGGCAGGCCTGGCACCAGTGCC
CATGATCATCTTTGCTAAGGATGGGCATTTTGCCCTGGAGGAGCTGGCCCAAGCTGGCTATGAGGTGGTT
GGGCTTGACTGGACAGTGGCCCCAAAGAAAGCCCGGGAGTGTGTGGGGAAGACGGTGACATTGCAGGGCA
ACCTGGACCCCTGTGCCTTGTATGCATCTGAGGAGGAGATCGGGCAGTTGGTGAAGCAGATGCTGGATGA
CTTTGGACCACATCGCTACATTGCCAACCTGGGCCATGGGCTTTATCCTGACATGGACCCAGAACATGTG
GGCGCCTTTGTGGATGCTGTGCATAAACACTCACGTCTGCTTCGACAGAACTGA
```

```
LOCUS       NP_000365               367 aa            linear   PRI 21-DEC-2008
DEFINITION  uroporphyrinogen decarboxylase [Homo sapiens].
ACCESSION   NP_000365
VERSION     NP_000365.3  GI:71051616
DBSOURCE    REFSEQ: accession NM_000374.3
KEYWORDS    .
SOURCE      Homo sapiens (human)
```

1.4.1.2 Información recipilada de Mus Musculus

Official Symbol	Urod — provided by MGI
Official Full Name	uroporphyrinogen decarboxylase — provided by MGI
Primary source	MGI:98916
Locus tag	RP23-419D6.3
See related	Ensembl:ENSMUSG00000028684
Gene type	protein coding
RefSeq status	PROVISIONAL
Organism	*Mus musculus*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus*
Also known as	Uro-d; AI323803; Urod

Genomic regions, transcripts, and products

(minus strand) Go to reference sequence details — Try our new Sequence Viewer


```
LOCUS       NM_009478               1199 bp    mRNA    linear   ROD 04-JAN-2009
DEFINITION  Mus musculus uroporphyrinogen decarboxylase (Urod), mRNA.
ACCESSION   NM_009478
VERSION     NM_009478.3  GI:145301611
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
     source          1..1199
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="C57BL/6"
                     /db_xref="taxon:10090"
                     /chromosome="4"
                     /map="4 50.6 cM"
     gene            1..1199
                     /gene="Urod"
                     /gene_synonym="Uro-d"
                     /gene_synonym="AI323803"
                     /note="uroporphyrinogen decarboxylase"
                     /db_xref="GeneID:22275"
                     /db_xref="MGI:98916"
     CDS             26..1129
                     /gene="Urod"
                     /gene_synonym="Uro-d"
                     /gene_synonym="AI323803"
                     /EC_number="4.1.1.37"
                     /codon_start=1
                     /product="uroporphyrinogen decarboxylase"
                     /protein_id="NP_033504.2"
                     /db_xref="GI:110347606"
                     /db_xref="CCDS:CCDS18521.1"
                     /db_xref="GeneID:22275"
```

```
                     /db_xref="MGI:98916"
                     /translation="MEANGFGLQNFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLP
                     EFRETRAAQDFFSTCRSPEACCELTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVP
                     GKGPSFPEPLREERDLERLRDPAAAASELGYVFQAITLTRQRLAGRVPLIGFAGAPWT
                     LMTYMVEGGSSSTMAQAKRWLYQRPQASHKLLGILTDVLVPYLIGQVAAGAQALQLFE
                     SHAGHLGTELFSKFALPYIRDVAKRVKAGLQKAGLAPVPMIIFAKDGHFALEELAQAG
                     YEVVGLDWTVAPKKARERVGKAVTLQGNLDPCALYASEEEIGRLVQQMLDDFGPQRYI
                     ANLGHGLYPDMDPERVGAFVDAVHKHSRLLRQN"

LOCUS       NP_033504               367 aa            linear   ROD 04-JAN-2009
DEFINITION  uroporphyrinogen decarboxylase [Mus musculus].
ACCESSION   NP_033504
VERSION     NP_033504.2  GI:110347606
DBSOURCE    REFSEQ: accession NM_009478.3
KEYWORDS    .
SOURCE      Mus musculus (house mouse)

>gi|145301611:26-1129 Mus musculus uroporphyrinogen decarboxylase (Urod), mRNA
ATGGAGGCGAACGGGTTCGGACTCCAGAATTTCCCGGAGCTGAAGAATGACACGTTCCTGAGAGCAGCCT
GGGGAGAGGAAACAGACTATACTCCCGTTTGGTGCATGAGACAGGCAGGCCGCTACTTACCAGAGTTTAG
GGAAACCAGGGCTGCCCAGGACTTCTTCAGCACCTGCCGATCTCCCGAGGCTTGCTGTGAACTGACTCTA
CAGCCACTACGAAGGTTTCCTCTGGATGCTGCCATAATTTTCTCTGACATCCTTGTTGTACCCCAGGCAT
TGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTTCCAGAGCCATTAAGAGAAGAGCGGGA
CTTAGAGCGTCTACGGGATCCAGCAGCAGCGGCTTCAGAGTTAGGCTATGTGTTCCAAGCCATCACCCTT
ACTCGACAACGGCTGGCCGGACGTGTGCCACTAATTGGCTTTGCTGGTGCTCCGTGGACCCTAATGACAT
ACATGGTTGAAGGCGGCAGTTCAAGCACCATGGCTCAGGCCAAACGATGGCTCTACCAAAGGCCACAGGC
CAGTCACAAGCTGCTTGGCATACTCACTGATGTTCTGGTCCCATACCTAATAGGACAAGTGGCTGCTGGT
GCTCAGGCATTGCAGCTCTTTGAGTCCCACGCAGGACATCTTGGCACCGAGCTCTTCAGCAAGTTTGCAC
TGCCCTACATTCGTGATGTGGCCAAGCGAGTGAAGGCTGGGTTGCAGAAGGCAGGCCTGGCACCAGTGCC
CATGATCATCTTTGCTAAGGATGGACATTTTGCCCTGGAAGAGCTGGCCCAGGCTGGCTATGAGGTAGTT
GGACTTGACTGGACAGTGGCTCCAAAGAAAGCCCGGGAACGTGTCGGGAAGGCAGTGACCCTGCAGGGGA
ACCTGGATCCCTGTGCCTTGTATGCATCTGAGGAAGAGATCGGTCGGCTGGTGCAGCAAATGCTGGATGA
CTTTGGGCCTCAACGCTACATTGCCAACCTAGGGCATGGGCTTTACCCTGACATGGACCCAGAACGTGTA
GGAGCCTTTGTGGATGCTGTACACAAACATTCACGCCTGCTTCGACAGAATTGA
```

1.4.1.3 Información Recopilada de Rattus norvegicus

Official Symbol	Urod — provided by RGD
Official Full Name	uroporphyrinogen decarboxylase — provided by RGD
Primary source	RGD:3946
See related	Ensembl:ENSRNOG00000018211
Gene type	protein coding
RefSeq status	VALIDATED
Organism	*Rattus norvegicus*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Rattus*
Also known as	Urod
Summary	enzyme involved in catalyzing the conversion of uroporphyrinogen to coproporphyrinogen [RGD]

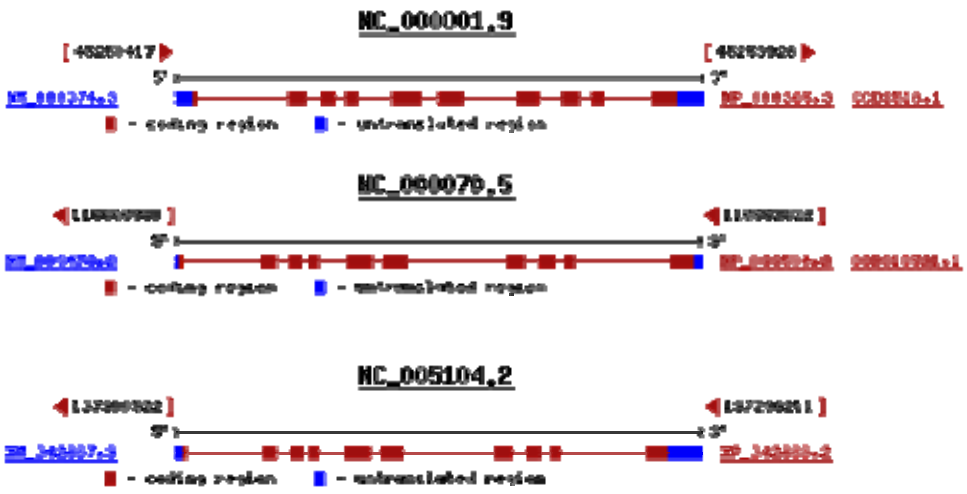
Genomic regions, transcripts, and products

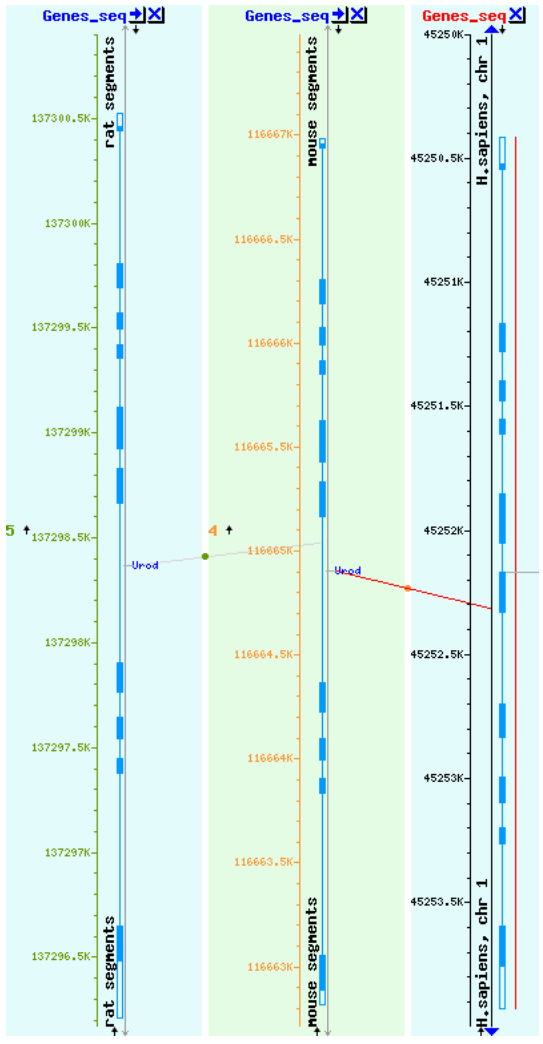
(minus strand) Go to reference sequence details — Try our new Sequence Viewer


```
LOCUS       NM_019209               1232 bp    mRNA    linear   ROD 22-OCT-2008
DEFINITION  Rattus norvegicus uroporphyrinogen decarboxylase (Urod), mRNA.
ACCESSION   NM_019209 XM_001067008 XM_342887
VERSION     NM_019209.1  GI:158749612
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
     source          1..1104
                     /organism="Rattus norvegicus"
                     /mol_type="mRNA"
                     /strain="Sprague-Dawley"
                     /db_xref="taxon:10116"
                     /chromosome="5"
                     /map="5q36"
     gene            <1..>1104
                     /gene="Urod"
                     /note="uroporphyrinogen decarboxylase"
                     /db_xref="GeneID:29421"
                     /db_xref="RGD:3946"
     CDS             1..1104
                     /gene="Urod"
                     /EC_number="4.1.1.37"
                     /codon_start=1
                     /product="uroporphyrinogen decarboxylase"
                     /protein_id="NP_062082.1"
                     /db_xref="GI:158749613"
                     /db_xref="GeneID:29421"
                     /db_xref="RGD:3946"
```

```
                     /translation="MEANGLGLQNFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLP
                     EFRETRAAQDFFSTCRSPEACCELTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVP
                     GKGPSFPEPLREERDLERLRDPAAVASELGYVFQAITLTRQQLAGRVPLIGFAGAPWT
                     LMTYMVEGGSSSTMAQAKRWLYQKPLASHKLLGILTDALVPYLIGQVAAGAQALQLFE
                     SHAGHLGSELFSKFALPYIRDVAKRVKAGLQKAGLAPVPMIIFAKDGHFALEELAQAG
                     YEVVGLDWTVAPKKARERVGKTVTLQGNLDPCALYASEEEIGRLVQQMLNDFGPQRYI
                     ANLGHGLYPDMDPEHVGAFVDAVHKHSRLLRQN"

LOCUS       NP_062082               367 aa            linear   ROD 22-OCT-2008
DEFINITION  uroporphyrinogen decarboxylase [Rattus norvegicus].
ACCESSION   NP_062082 XP_001067008 XP_342888
VERSION     NP_062082.1  GI:158749613
DBSOURCE    REFSEQ: accession NM_019209.1
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)

>gi|158749612:38-1141 Rattus norvegicus uroporphyrinogen decarboxylase (Urod), mRNA
ATGGAGGCGAACGGCTTGGGACTCCAGAATTTCCCGGAGCTGAAGAATGACACGTTCTTGAGAGCAGCCT
GGGGAGAGGAAACAGACTATACTCCTGTTTGGTGCATGAGACAAGCAGGCCGCTACTTACCAGAGTTTAG
GGAAACCAGGGCTGCCCAGGACTTCTTCAGCACCTGTCGATCTCCTGAGGCTTGCTGTGAACTGACTCTG
CAGCCACTGCGAAGGTTTCCTCTGGATGCTGCTATAATTTTCTCTGACATCCTTGTTGTACCCCAGGCAC
TGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTTCCAGAGCCATTAAGAGAAGAGCGGGA
CTTAGAGCGTCTACGGGATCCAGCAGCAGTGGCTTCAGAGTTAGGCTATGTGTTCCAAGCCATCACCCTT
ACCCGACAACAGCTGGCTGGACGTGTGCCACTGATTGGCTTTGCTGGTGCTCCGTGGACCCTGATGACGT
ACATGGTTGAAGGCGGCAGTTCAAGTACCATGGCTCAGGCCAAGCGATGGCTCTATCAGAAGCCACTGGC
CAGTCACAAGCTGCTTGGCATACTCACTGATGCTCTGGTCCCATATCTAATAGGACAAGTAGCTGCTGGT
GCTCAGGCATTGCAGCTCTTTGAGTCCCACGCAGGACATCTTGGCTCCGAGCTCTTCAGCAAGTTTGCAC
TGCCTTACATCCGTGATGTGGCCAAGCGAGTGAAGGCTGGGTTGCAGAAGGCGGGCCTGGCACCAGTGCC
CATGATCATCTTTGCCAAGGATGGACATTTTGCCCTGGAAGAGCTGGCCCAGGCTGGCTATGAGGTAGTT
GGACTTGACTGGACAGTGGCTCCAAAGAAAGCCCGGGAACGTGTTGGAAAGACAGTGACTCTGCAGGGGA
ACCTGGATCCCTGTGCCCTGTATGCATCTGAGGAAGAGATTGGTCGACTGGTGCAGCAGATGCTGAATGA
CTTTGGGCCACAGCGCTACATTGCTAACCTAGGGCATGGGCTTTACCCTGACATGGACCCAGAACACGTA
GGAGCCTTTGTGGATGCAGTACACAAACACTCACGCCTGCTTCGACAGAATTGA
```

1.4.1.4 Comparación visual de los tres genes:



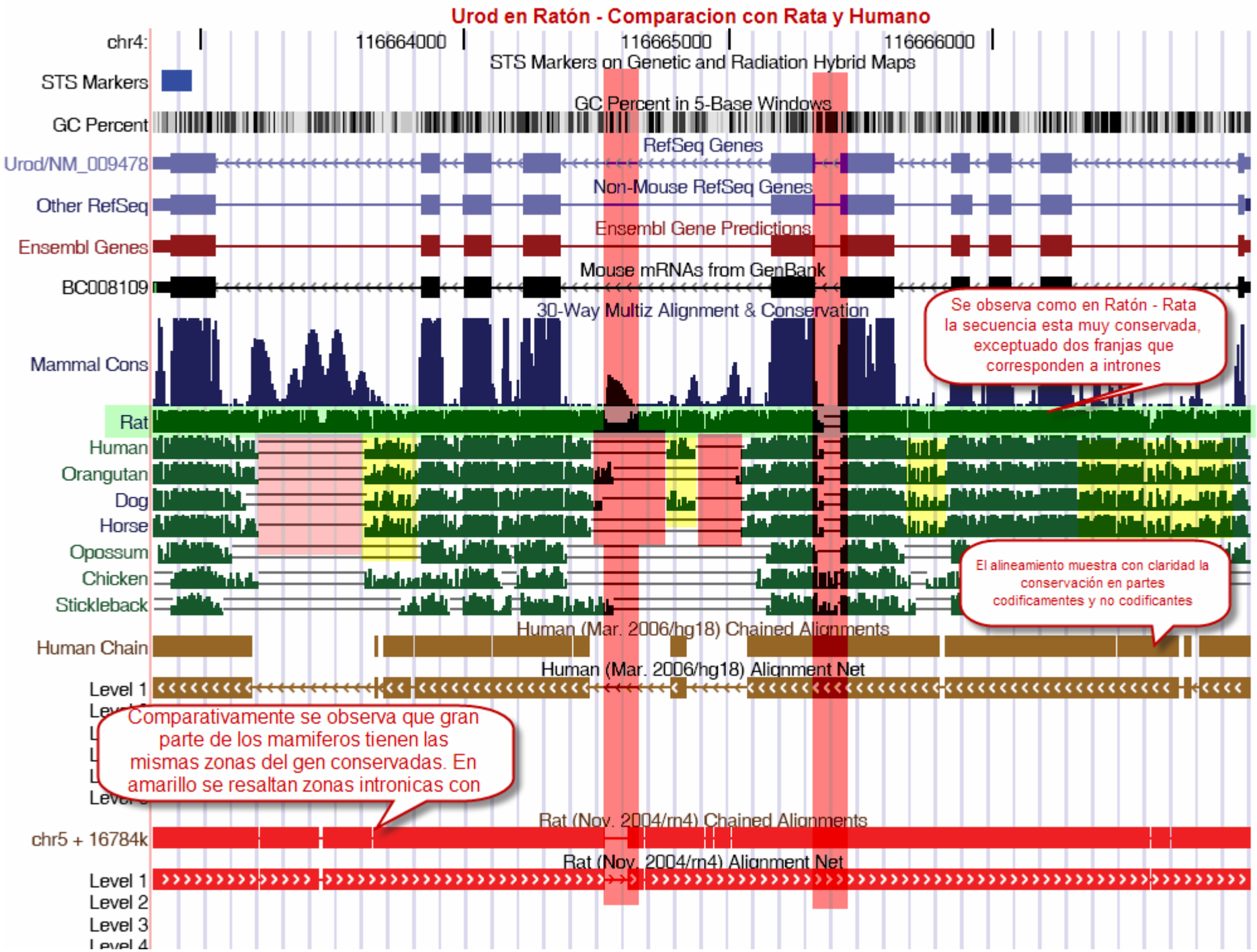


Se observa

- Tanto en humano, ratón como en rata, la estructura del gen es similar, 10 exones.
- Los UTRs tanto los 5’ como los 3’, son de diferente tamaño.
- Parece que la función de la proteína es muy específica, ya que esta muy conservada.
- Como comprobaremos en el siguiente ejercicio, las mayor similitud se da entre ratón-rata. Aunque paradójicamente están en cromosomas distintos… *No se muy bien como se come esto?¿?.*

1.4.1.5 Tabla resumen de los resultados

Organismo	Homo sapiens	Mus musculus	Rattus norvegicus
Gen			
Secuencia	LOCUS NM_000374 1383 bp mRNA linear PRI 21-DEC-2008 DEFINITION Homo sapiens uroporphyrinogen decarboxylase (UROD), mRNA. ACCESSION NM_000374 VERSION NM_000374.3 GI:71051615 KEYWORDS . SOURCE Homo sapiens (human)	LOCUS NM_009478 1199 bp mRNA linear ROD 04-JAN-2009 DEFINITION Mus musculus uroporphyrinogen decarboxylase (Urod), mRNA. ACCESSION NM_009478 VERSION NM_009478.3 GI:145301611 KEYWORDS . SOURCE Mus musculus (house mouse)	LOCUS NM_019209 1232 bp mRNA linear ROD 22-OCT-2008 DEFINITION Rattus norvegicus uroporphyrinogen decarboxylase (Urod), mRNA. ACCESSION NM_019209 XM_001067008 XM_342887 VERSION NM_019209.1 GI:158749612 KEYWORDS . SOURCE Rattus norvegicus (Norway rat)
Proteina	LOCUS NP_000365 367 aa linear PRI 21-DEC-2008 DEFINITION uroporphyrinogen decarboxylase [Homo sapiens]. ACCESSION NP_000365 VERSION NP_000365.3 GI:71051616 DBSOURCE REFSEQ: accession NM_000374.3 KEYWORDS . SOURCE Homo sapiens (human)	LOCUS NP_033504 367 aa linear ROD 04-JAN-2009 DEFINITION uroporphyrinogen decarboxylase [Mus musculus]. ACCESSION NP_033504 VERSION NP_033504.2 GI:110347606 DBSOURCE REFSEQ: accession NM_009478.3 KEYWORDS . SOURCE Mus musculus (house mouse)	LOCUS NP_062082 367 aa linear ROD 22-OCT-2008 DEFINITION uroporphyrinogen decarboxylase [Rattus norvegicus]. ACCESSION NP_062082 XP_001067008 XP_342888 VERSION NP_062082.1 GI:158749613 DBSOURCE REFSEQ: accession NM_019209.1 KEYWORDS . SOURCE Rattus norvegicus (Norway rat)


Urod en Ratón - Comparacion con Rata y Humano
chr4:
116664000
116665000
116666000
STS Markers on Genetic and Radiation Hybrid Maps
STS Markers
GC Percent in 5-Base Windows
GC Percent
RefSeq Genes
Urod/NM_009478
Non-Mouse RefSeq Genes
Other RefSeq
Ensembl Gene Predictions
Ensembl Genes
Mouse mRNAs from GenBank
BC008109
30-Way Multiz Alignment & Conservation
Mammal Cons
Se observa como en Ratón - Rata la secuencia esta muy conservada, exceptuado dos franjas que corresponden a intrones
Rat
Human
Orangutan
Dog
Horse
Opossum
Chicken
Stickleback
El alineamiento muestra con claridad la conservación en partes codificamentes y no codificantes
Human (Mar. 2006/hg18) Chained Alignments
Human Chain
Human (Mar. 2006/hg18) Alignment Net
Level 1
Comparativamente se observa que gran parte de los mamiferos tienen las mismas zonas del gen conservadas. En amarillo se resaltan zonas intronicas con
Rat (Nov. 2004/rn4) Chained Alignments
chr5 + 16784k
Rat (Nov. 2004/rn4) Alignment Net
Level 1
Level 2
Level 3

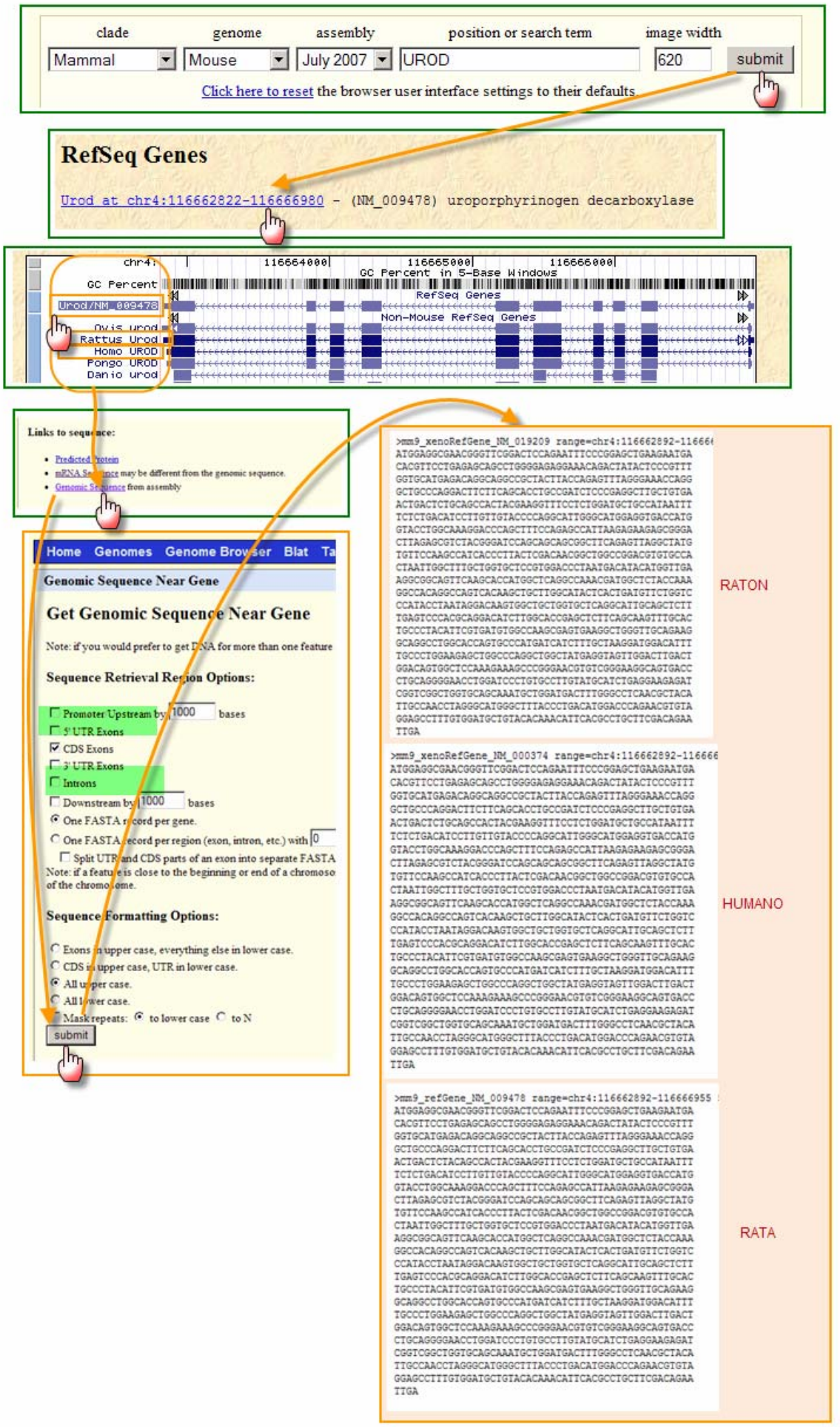
2 PARTE II

2.1 EJERCICIO 1: GLOBAL contra LOCAL

2.1.1 Extraed del browser genómico UCSC http://genome.ucsc.edu/ la región codificante (CDS, solo exones) del gen URO-D en humano y en ratón.

Recordad la forma de hacerlo correctamente (pregunta 21). También tened en cuenta que la secuencia CDS comienza por ATG y acaba con un codón STOP.

Tal y como habíamos hecho en problemas anteriores. Recuperamos las secuencias siguiendo el esquema de la figura:

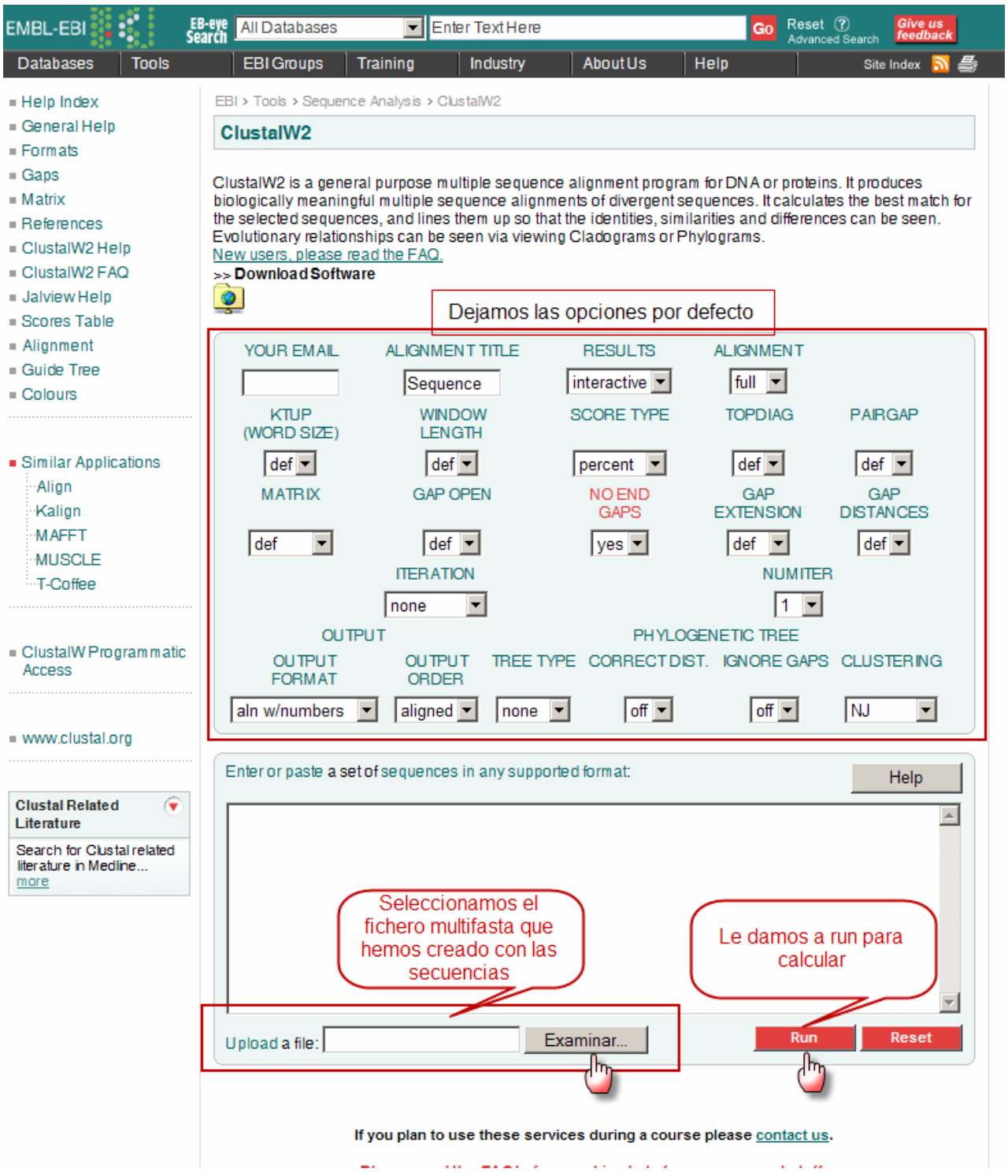


Preparamos un fichero con las tres secuencias, y ficheros separados para cada una de ellas.

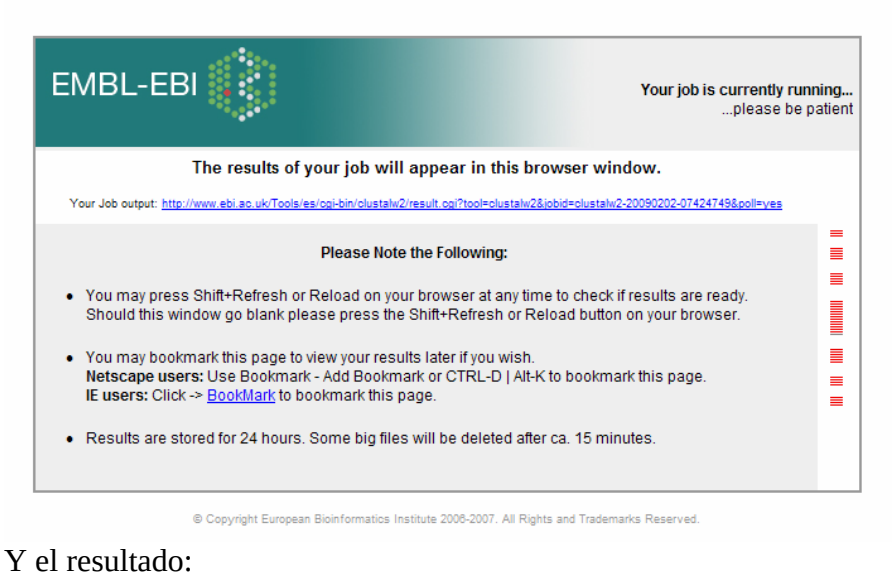
```
>Homo sapiens UROD
ATGGAAGCGAATGGGTTGGGACCTCAGGGTTTTCCGGAGCTGAAGAATGACACATTCCTGCGAGCAGCCT
GGGGAGAGGAAACAGACTACACTCCCGTTTGGTGCATGCGCCAGGCAGGCCGTTACTTACCAGAGTTTAG
GGAAACCCGGGCTGCCCAGGACTTTTTCAGCACGTGTCGCTCTCCTGAGGCCTGCTGTGAACTGACTCTG
CAGCCACTGCGTCGCTTCCCTCTGGATGCTGCCATCATTTTCTCCGACATCCTTGTTGTACCCCAGGCAC
TGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTCCCAGAGCCATTAAGAGAAGAGCAGGA
CCTAGAACGCCTACGGGATCCAGAAGTGGTAGCCTCTGAGCTAGGCTATGTGTTCCAAGCCATCACCCTT
ACCCGACAACGACTGGCTGGACGTGTGCCGCTGATTGGCTTTGCTGGTGCCCCATGGACCCTGATGACAT
ACATGGTTGAGGGTGGTGGCTCAAGCACCATGGCTCAGGCCAAGCGCTGGCTCTATCAGAGACCTCAGGC
TAGTCACCAGCTGCTTCGCATCCTCACTGATGCTCTGGTCCCATATCTGGTAGGACAAGTGGTGGCTGGT
GCCCAGGCATTGCAGCTGTTTGAGTCCCATGCAGGGCATCTTGGCCCACAGCTCTTCAACAAGTTTGCAC
TGCCTTACATCCGTGATGTGGCCAAGCAAGTGAAGGCCAGGTTGCGGGAGGCAGGCCTGGCACCAGTGCC
CATGATCATCTTTGCTAAGGATGGGCATTTTGCCCTGGAGGAGCTGGCCCAAGCTGGCTATGAGGTGGTT
GGGCTTGACTGGACAGTGGCCCCAAAGAAAGCCCGGGAGTGTGTGGGGAAGACGGTGACATTGCAGGGCA
ACCTGGACCCCTGTGCCTTGTATGCATCTGAGGAGGAGATCGGGCAGTTGGTGAAGCAGATGCTGGATGA
CTTTGGACCACATCGCTACATTGCCAACCTGGGCCATGGGCTTTATCCTGACATGGACCCAGAACATGTG
GGCGCCTTTGTGGATGCTGTGCATAAACACTCACGTCTGCTTCGACAGAACTGA
>Mus musculus UROD
ATGGAGGCGAACGGGTTCGGACTCCAGAATTTCCCGGAGCTGAAGAATGACACGTTCCTGAGAGCAGCCT
GGGGAGAGGAAACAGACTATACTCCCGTTTGGTGCATGAGACAGGCAGGCCGCTACTTACCAGAGTTTAG
GGAAACCAGGGCTGCCCAGGACTTCTTCAGCACCTGCCGATCTCCCGAGGCTTGCTGTGAACTGACTCTA
CAGCCACTACGAAGGTTTCCTCTGGATGCTGCCATAATTTTCTCTGACATCCTTGTTGTACCCCAGGCAT
TGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTTCCAGAGCCATTAAGAGAAGAGCGGGA
CTTAGAGCGTCTACGGGATCCAGCAGCAGCGGCTTCAGAGTTAGGCTATGTGTTCCAAGCCATCACCCTT
ACTCGACAACGGCTGGCCGGACGTGTGCCACTAATTGGCTTTGCTGGTGCTCCGTGGACCCTAATGACAT
ACATGGTTGAAGGCGGCAGTTCAAGCACCATGGCTCAGGCCAAACGATGGCTCTACCAAAGGCCACAGGC
CAGTCACAAGCTGCTTGGCATACTCACTGATGTTCTGGTCCCATACCTAATAGGACAAGTGGCTGCTGGT
GCTCAGGCATTGCAGCTCTTTGAGTCCCACGCAGGACATCTTGGCACCGAGCTCTTCAGCAAGTTTGCAC
TGCCCTACATTCGTGATGTGGCCAAGCGAGTGAAGGCTGGGTTGCAGAAGGCAGGCCTGGCACCAGTGCC
CATGATCATCTTTGCTAAGGATGGACATTTTGCCCTGGAAGAGCTGGCCCAGGCTGGCTATGAGGTAGTT
GGACTTGACTGGACAGTGGCTCCAAAGAAAGCCCGGGAACGTGTCGGGAAGGCAGTGACCCTGCAGGGGA
ACCTGGATCCCTGTGCCTTGTATGCATCTGAGGAAGAGATCGGTCGGCTGGTGCAGCAAATGCTGGATGA
CTTTGGGCCTCAACGCTACATTGCCAACCTAGGGCATGGGCTTTACCCTGACATGGACCCAGAACGTGTA
GGAGCCTTTGTGGATGCTGTACACAAACATTCACGCCTGCTTCGACAGAATTGA
>Rattus norvegicus UROD
ATGGAGGCGAACGGCTTGGGACTCCAGAATTTCCCGGAGCTGAAGAATGACACGTTCTTGAGAGCAGCCT
GGGGAGAGGAAACAGACTATACTCCTGTTTGGTGCATGAGACAAGCAGGCCGCTACTTACCAGAGTTTAG
GGAAACCAGGGCTGCCCAGGACTTCTTCAGCACCTGTCGATCTCCTGAGGCTTGCTGTGAACTGACTCTG
CAGCCACTGCGAAGGTTTCCTCTGGATGCTGCTATAATTTTCTCTGACATCCTTGTTGTACCCCAGGCAC
TGGGCATGGAGGTGACCATGGTACCTGGCAAAGGACCCAGCTTTCCAGAGCCATTAAGAGAAGAGCGGGA
CTTAGAGCGTCTACGGGATCCAGCAGCAGTGGCTTCAGAGTTAGGCTATGTGTTCCAAGCCATCACCCTT
ACCCGACAACAGCTGGCTGGACGTGTGCCACTGATTGGCTTTGCTGGTGCTCCGTGGACCCTGATGACGT
ACATGGTTGAAGGCGGCAGTTCAAGTACCATGGCTCAGGCCAAGCGATGGCTCTATCAGAAGCCACTGGC
CAGTCACAAGCTGCTTGGCATACTCACTGATGCTCTGGTCCCATATCTAATAGGACAAGTAGCTGCTGGT
```

2.1.2 El programa CLUSTALW realiza alineamientos globales de dos o más secuencias. Usad el servidor de CLUSTALW implementado en el EBI http://www.ebi.ac.uk/clustalw/ para alinear las dos secuencias codificantes.

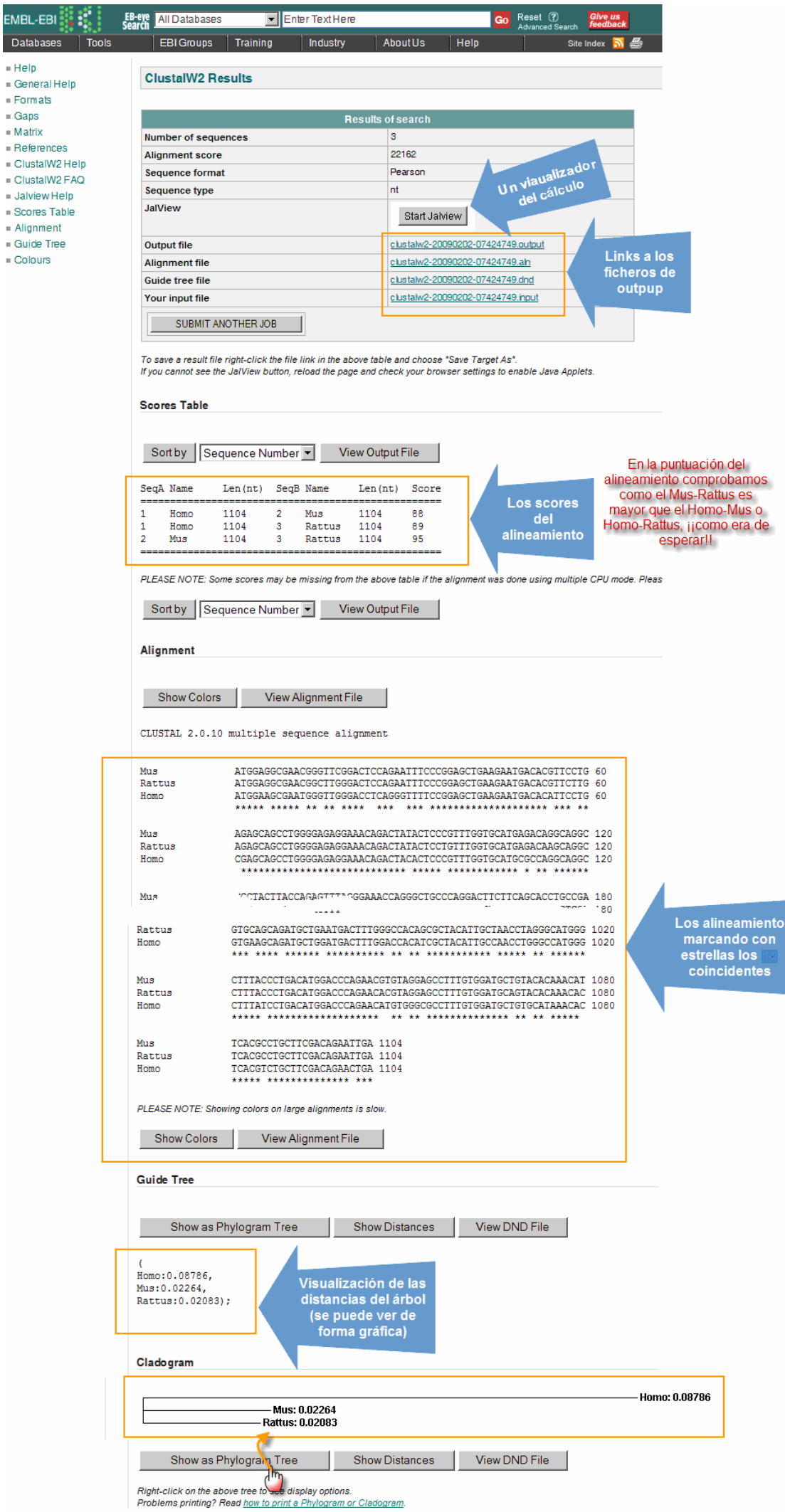
En principio, si no se dice lo contrario durante la PEC, se recomienda usar los parámetros por defecto de los programas utilizados. Recordad que CLUSTALW es un programa que calcula alineamientos globales.



Aparece un cuadro de dialogo para que esperemos a que acabe el cálculo:



Y el resultado:

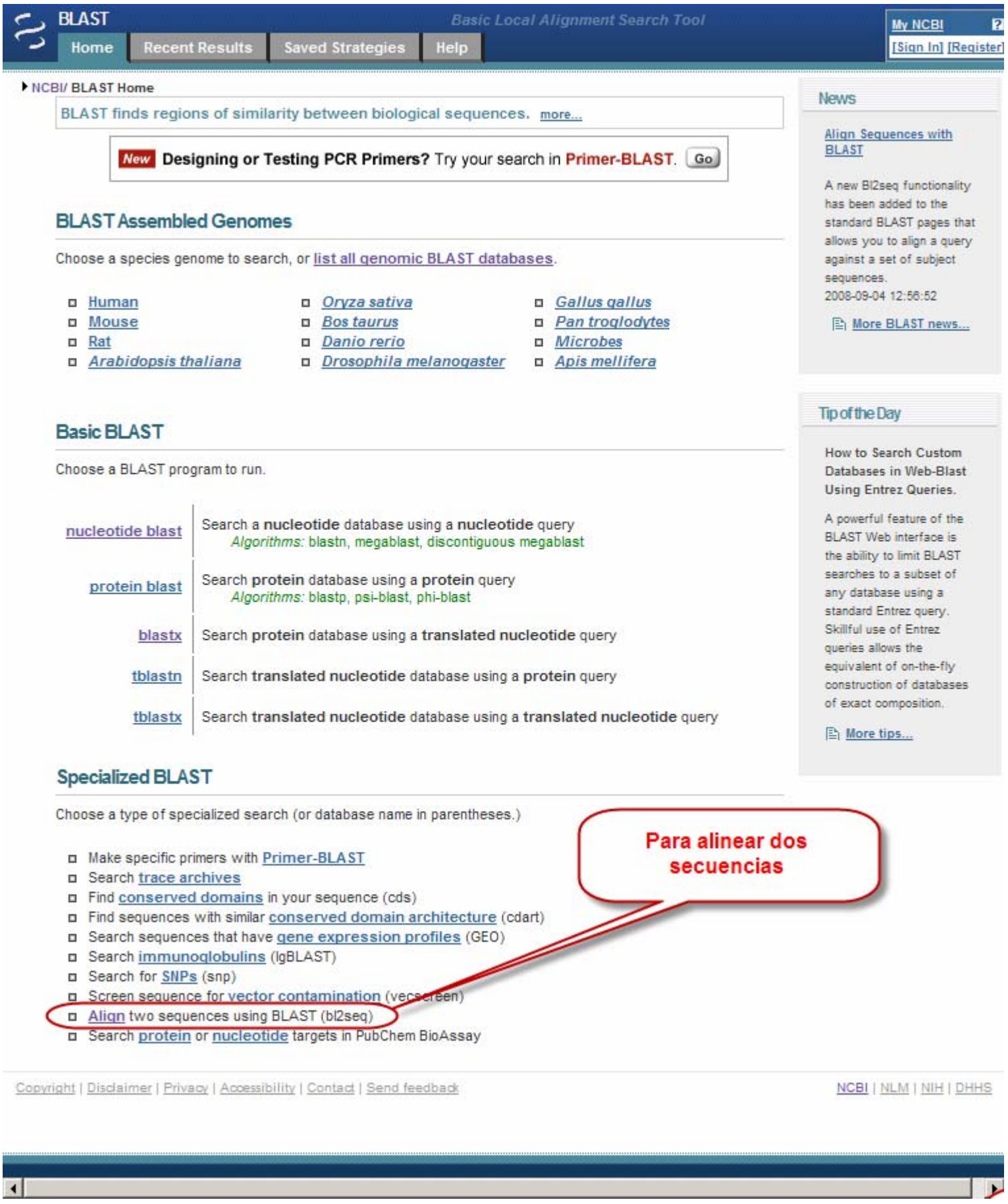


Si únicamente alineamos Huma y Mouse obtenemos:

SeqA	Name	Len(nt)	SeqB	Name	Len(nt)	Score
1	Homo	1104	2	Mus	1104	88

2.1.3 El programa BLAST realiza alineamientos locales. Debéis acceder a la página principal del programa (NCBI) http://www.ncbi.nlm.nih.gov/blast/ y encontrar que versión de BLAST se debe usar para alinear 2 secuencias. Con esta versión, debéis calcular el alineamiento local de las dos regiones CDS del apartado anterior.

Recordad que la mayoría de versiones de BLAST están preparadas para alinear 1 secuencia contra una base de datos de secuencias almacenadas en su servidor. En nuestro caso particular, queremos alinear 2 secuencias, una contra la otra, no deseamos usar ninguna base de datos.



NCBI/ BLAST/ blastn suite-2sequences/ **Formatting Results - SDAPK3N3113**

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download

Blast 2 sequences

Homo - Mouse UROD

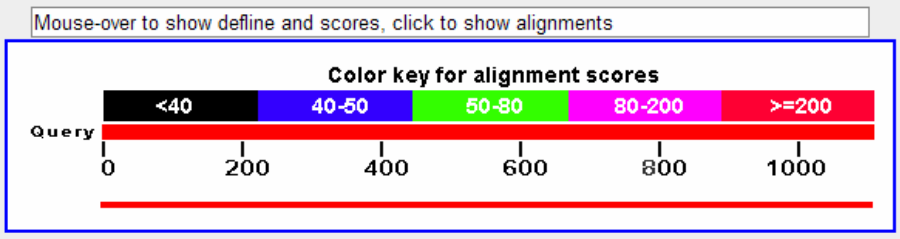
Query ID	lcl|6713	**Subject ID**	6715
Description	Homo sapiens UROD	**Description**	Mus musculus UROD
Molecule type	nucleic acid	**Molecule type**	nucleic acid
Query Length	1104	**Subject Length**	1104
		Program	BLASTN 2.2.19+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports]

▼ Graphic Summary

Distribution of 1 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



▶ Dot Matrix View

▼ Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
6715	Mus musculus UROD	1363	1363	100%	0.0	88%	

▶ Alignments

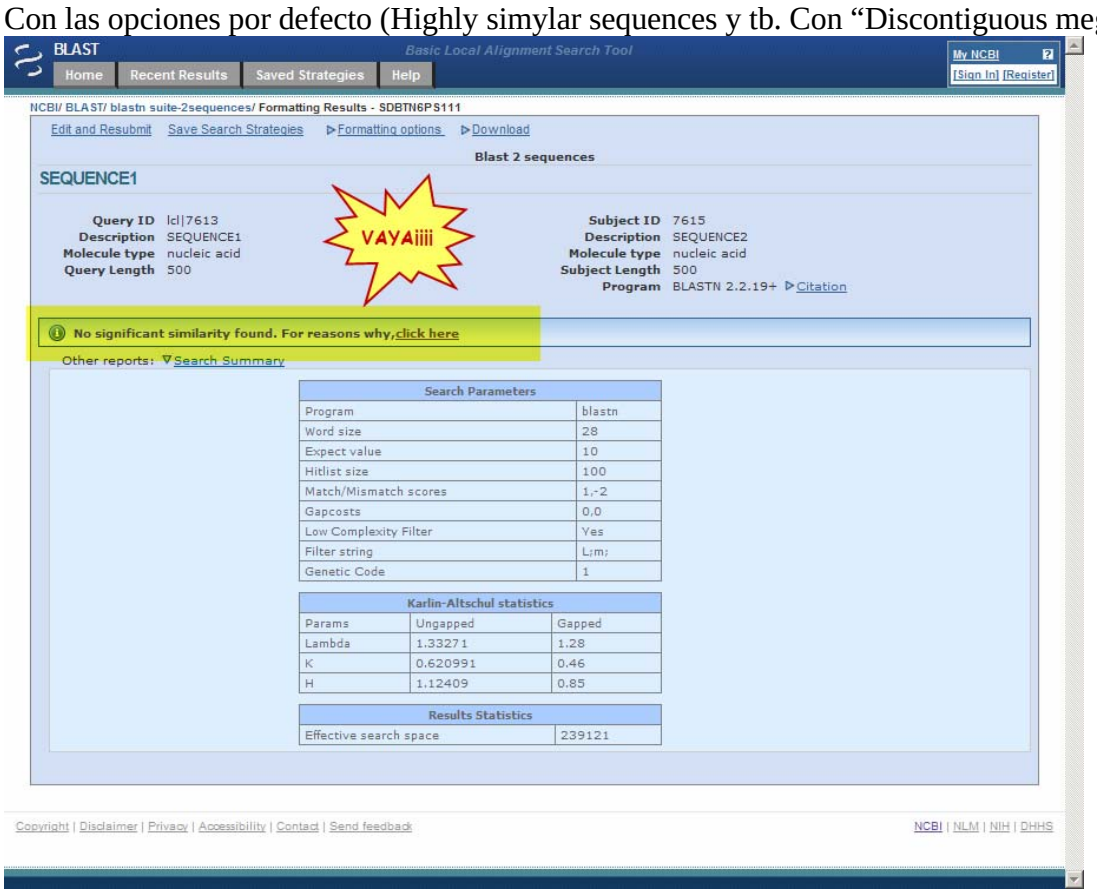
2.1.4 Ahora, usad el servidor de CLUSTALW para alinear la secuencia Ex1-seqA.fa y la secuencia Ex1-seqB.fa adjuntadas en este enunciado.

Alineamiento Global:
El alineamiento resultante es:

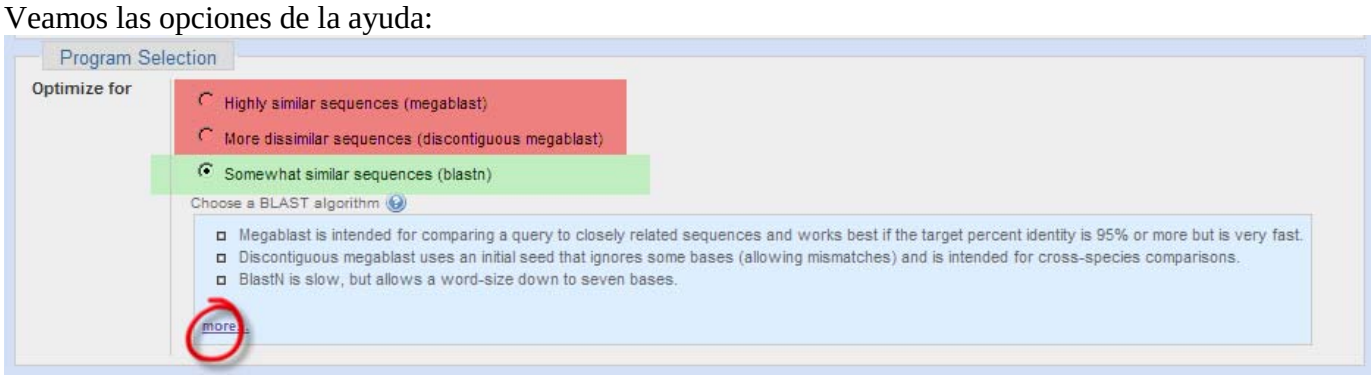
```
SeqA Name        Len(nt)  SeqB Name        Len(nt)  Score
=========================================================
1    SEQUENCE1   500      2    SEQUENCE2   500      5
=========================================================


CLUSTAL 2.0.10 multiple sequence alignment


SEQUENCE1       GGTGTCCATATCCACCAAAAGCATAAAGGGGCCTTTCGAGGTGATCTAACACAACTCCAT 60
SEQUENCE2       ------------CTTGTAAGCTGCAGAGGTGTAATACG-GAAGAGCGCATGCGGCACC-C 46
                            *    **     * *** *   * ** *  ** *  *  *  * **

SEQUENCE1       CAGAGGGCAATGTGTCTGTTTTACAAAGAGAAAACCTCTTGGCAGATCGGACACACTATA 120
SEQUENCE2       CAGCGTACG-CATAGACATCCCCCTGTCGCAATGTTGCCTGGAGTGTGGCCTACACTAGC 105
                *** *  *    *     *    *      **     * ***    * *   ******

SEQUENCE1       GGCAGGCTTGATACCCCACTTTACACAGTCATATGGCAGTACTTCG-TACGAAGATAGTT 179
SEQUENCE2       GGTAAGATTTAGTCTGTAG---ACGTAATTATAATCATATATACCGATTTCATTACGGTG 162
                ** * * ** *  *   *    **  * * ***      **   ** *   *  *  **

SEQUENCE1       CCCCAATCCAGCCAGGTGCGCGTGTATTCCCCATCGAACTTGCTCTGCCTCGTGGCTTCG 239
SEQUENCE2       CTGGAACGTTGCGTTGTACGTGTCTCCTGTCTACTGAGGTTGA--CACCTCGCAGTGCAA 220
                *   **    **   ** ** ** *  *  * *  **  ***     *****  *

SEQUENCE1       CTTTGCCAACACACGCCTCAGTATATGGCGAACAATAATAAAGGGAGGC---CTCCATTG 296
SEQUENCE2       GAATTCCATTGAACCCATAGTCGTGGTGTGAGAAGCACATGACGGACATTAGCCTCATCA 280
                   * ***    ** * *     *   * **  *  *    * ***      *  ***

SEQUENCE1       AAGACTAGCA---TAGGTTCTGAGAGATTGCAGTGATTGGTTATGCGGCACCCCAGCGTA 353
SEQUENCE2       ACACTTAGCAACTCACGCTTCGCAGGTTATATGGGATGAGATGCCAGGGAGCTCATATCA 340
                *    *****    * * *  *   * *    * ***  * *    ** * * **    *

SEQUENCE1       CGCATAGTCCACATAGTCATCCCCAAGAGGCCTGAGACTGGTATTGTACAAGGCCCGTAA 413
SEQUENCE2       TGCG--GTAGATCGGTTAAGTCGCAATCTATCCTGGATAATCAGGGAAATACGCGCTAGT 398
                 **   **  *     * *  * ***     *   **     *  * *  * ** *

SEQUENCE1       TTGGGTTGACCACTAGCGACCCATTC--TTCAGCCCAAGTGCAACATGGCGCGGATATAT 471
SEQUENCE2       TTCTACGAAACCGAGGCGGCGAAAGCATTTGTGCCGAGGCACTATTTAAGACAAGTATTC 458
                **      * *    *** *  *  *  **  *** * *  * *  *    *   ***

SEQUENCE1       AACTAGTATCTAATGGAATCCCTGCTCTC------------ 500
SEQUENCE2       AGCTGTTTGATAAACGGCCTGCTGAGCGTACACAATTTTCTT 500
                * **  *   ***  *     ***  *
```

2.1.5 Como en la pregunta 3, realizad el alineamiento local de las dos secuencias Ex1-seqA.fa y Ex1-seqB.fa adjuntadas en este enunciado.

Con las opciones por defecto (Highly simylar sequences y tb. Con “Discontiguous megablast”) obtenemos :



Veamos las opciones de la ayuda:



http://www.ncbi.nlm.nih.gov/blast/producttable.shtml#tab31

4. Explanation for the program choices given in Tables 3.1 and 3.2

4.1 MEGABLAST is the tool of choice to identify a nucleotide sequence.

The best way to identify an unknown sequence is to see if that sequence already exists in a public database. If the database sequence is a well-characterized sequence, then one will have access to a wealth of biological information. MEGABLAST, discontiguous-megablast, and blastn all can be used to accomplish this goal. However, MEGABLAST is specifically designed to efficiently find long alignments between very similar sequences and thus is the best tool to use to find the identical match to your query sequence. In addition to the expect value significance cut-off, MEGABLAST also provides an adjustable percent identity cut-off for the alignment, which provides cut-off in addition to the significance cut-off threshold set by Expect value.
Web MEGABLAST and discontiguous megablast pages can also accept batch queries, the only web BLAST pages with this capability. Please refer to the "Batch Search" section for details.

4.2 Discontiguous MEGABLAST is better at finding nucleotide sequences similar, but not identical, to your nucleotide query.

The BLAST nucleotide algorithm finds similar sequences by breaking the query into short subsequences called words. The program identifies the exact matches to the query words first (word hits). BLAST program then extends these word hits in multiple steps to generate the final gapped alignments.
One of the important parameters governing the sensitivity of BLAST searches is the length of the initial words, or word size as it is called. The most important reason that blastn is more sensitive than MEGABLAST is that it uses a shorter default word size (11). Because of this, blastn is better than MEGABLAST at finding alignments to related nucleotide sequences from other organisms. The word size is adjustable in blastn and can be reduced from the default value to a minimum of 7 to increase search sensitivity.
A more sensitive search can be achieved by using the newly introduced discontiguous megablast page. This page uses an algorithm with the same name, which is similar to that reported by *Ma et.al.* Rather than requiring exact word matches as seeds for alignment extension, discontiguous megablast uses non-contiguous word within a longer window of template. In coding mode, the third base wobbling is taken into consideration by focusing on finding matches at the first and second codon positions while ignoring the mismatches in the third position. Searching in discontiguous MEGABLAST using the same word size is more sensitive and efficient than standard blastn using the same word size. For this reason, it is now the recommended tool for this type of search. Alternative non-coding patterns can also be specified if desired. Additional details on discontiguous are available at:
www.ncbi.nlm.nih.gov/blast/discontiguous.html
www.ncbi.nlm.nih.gov/Web/Newsltr/FallWinter02/blastlab.html
Parameters unique for discontiguous megablast are:

- word size: retricted to two options, i.e., 11 or 12
- template: only three options are available, 16, 18, or 21
- template type: coding (0), non-coding (1), or both (2)

It is important to point out that nucleotide-nucleotide searches are not the best method for finding homologous protein coding regions in other organisms. That task is better accomplished by performing searches at the protein level, by direct protein-protein BLAST searches or by translated BLAST searches.

This is because of the codon degeneracy, the greater information available in amino acid sequence, and the more sophisticated algorithm and scoring matrix used in protein-protein BLAST.

4.3 "Search for short nearly exact matches" is useful for primer or short nucleotide searches.

Short sequences (less than 20 bases) will often not find any significant matches to the database entries under the standard nucleotide-nucleotide BLAST settings. The usual reasons for this are that the significance threshold governed by the Expect value parameter is set too stringently and the default word size parameter is set too high.

You can adjust both the word size and the expect value on the standard BLAST pages to work with short sequences. NCBI provides a BLAST page with these values preset to give optimal results with short sequences. This page ("Search for short nearly exact matches") is linked under the nucleotide BLAST section of the main BLAST page.

Table 4.3.1 Parameter settings for standard blastn and "Search for short and nearly exact matches"

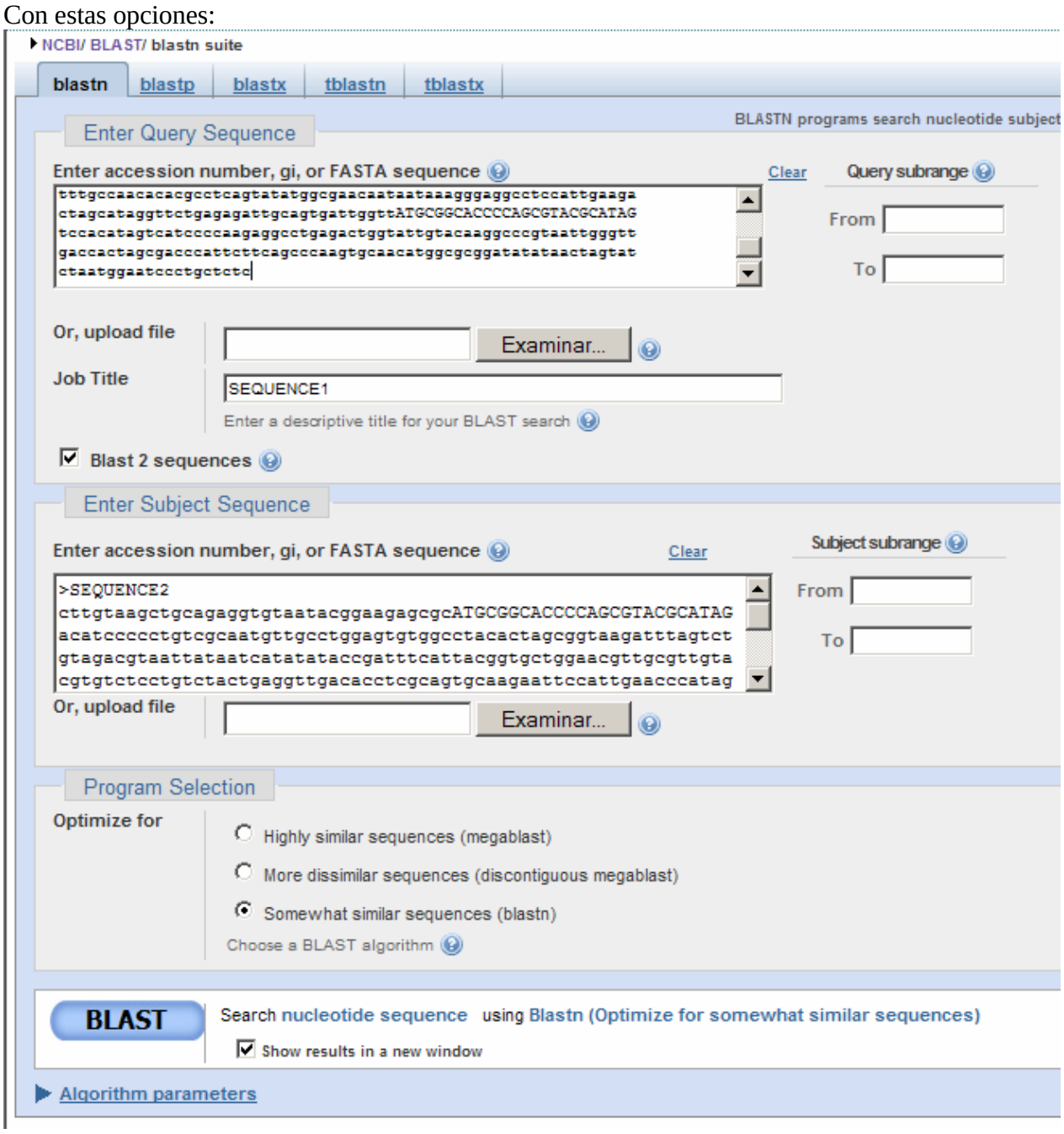
Program	Word Size	DUST Filter Setting	Expect Value
Standard blastn	11	On	10
Search for short nearly exact matches	7	Off	1000

A common use of this page is to check the specificity of PCR or hybridization primers. A useful way to check a pair of PCR primers is to first concatenate them by inserting string of 20 or more N's in between the two primers, and then search the concatenated pair as one sequence. Since BLAST looks for local alignments and automatically searches both strands, there is no need to reverse complement the reverse primer before doing the concatenation or the search.

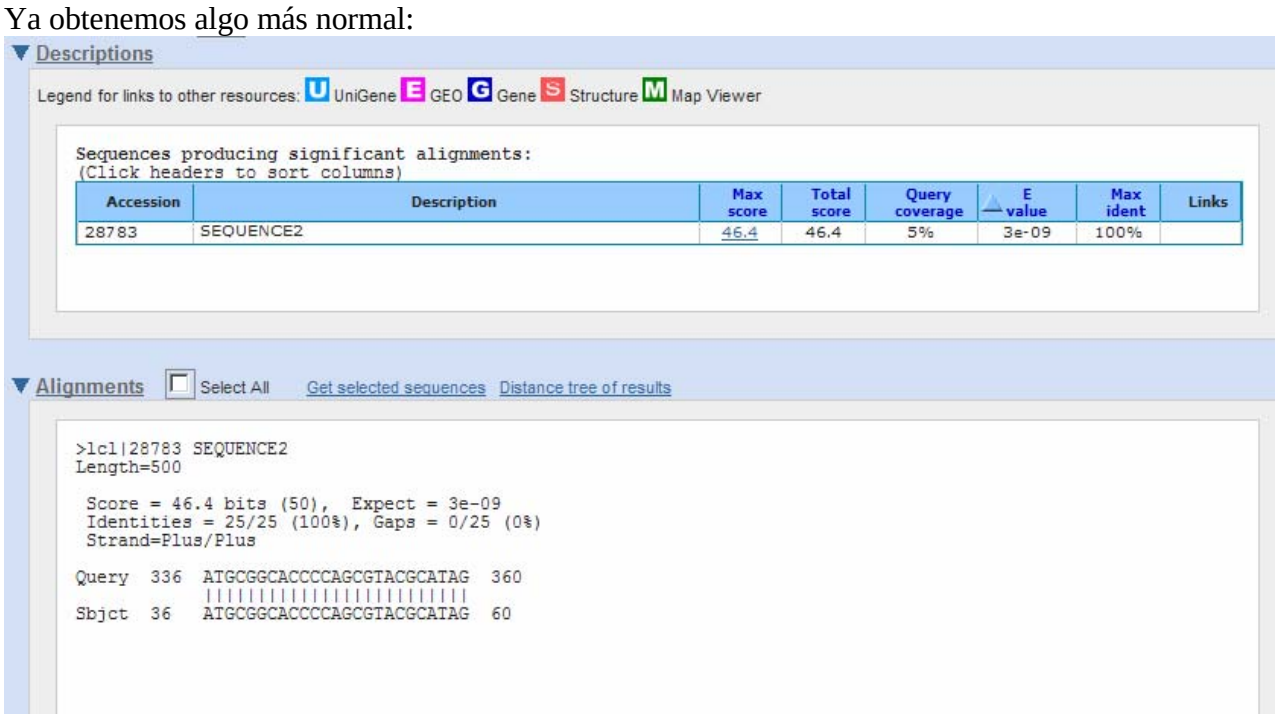
The query sequence should contain no ambiguous bases. Consensus motifs with degenerate bases, such as **AACNNNNNNRTAYG** (StySQI recognition site) or **TGGNNNNNNGCCAA** (NF-1 binding site) will not work for this type of search.

En definitiva, tenemos que usar blastn

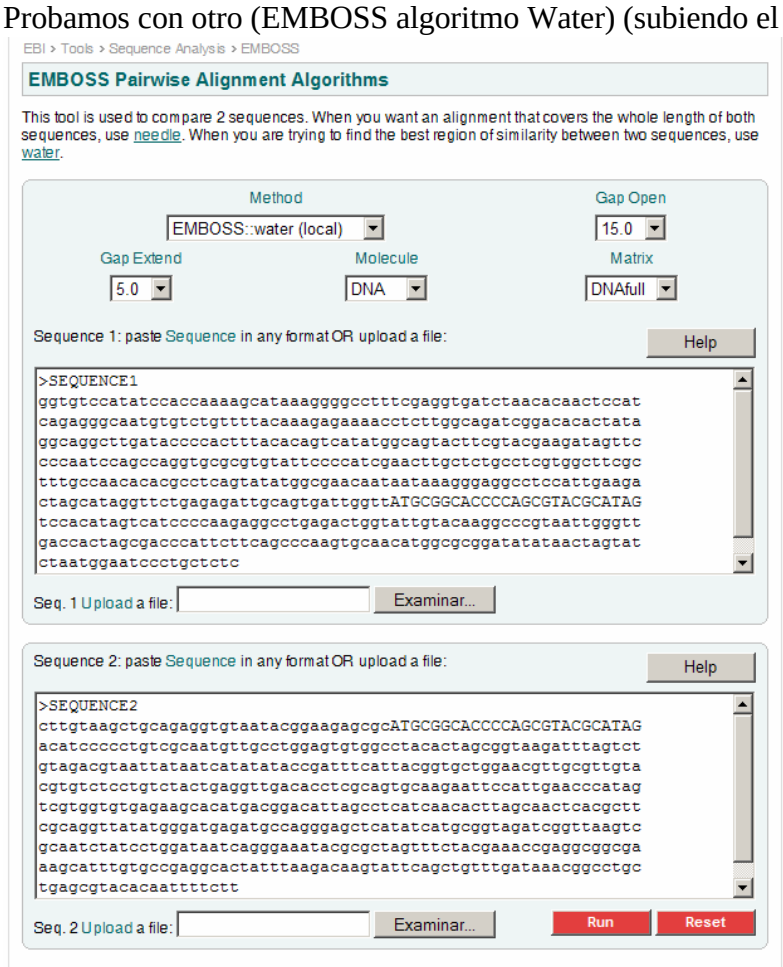
Con estas opciones:

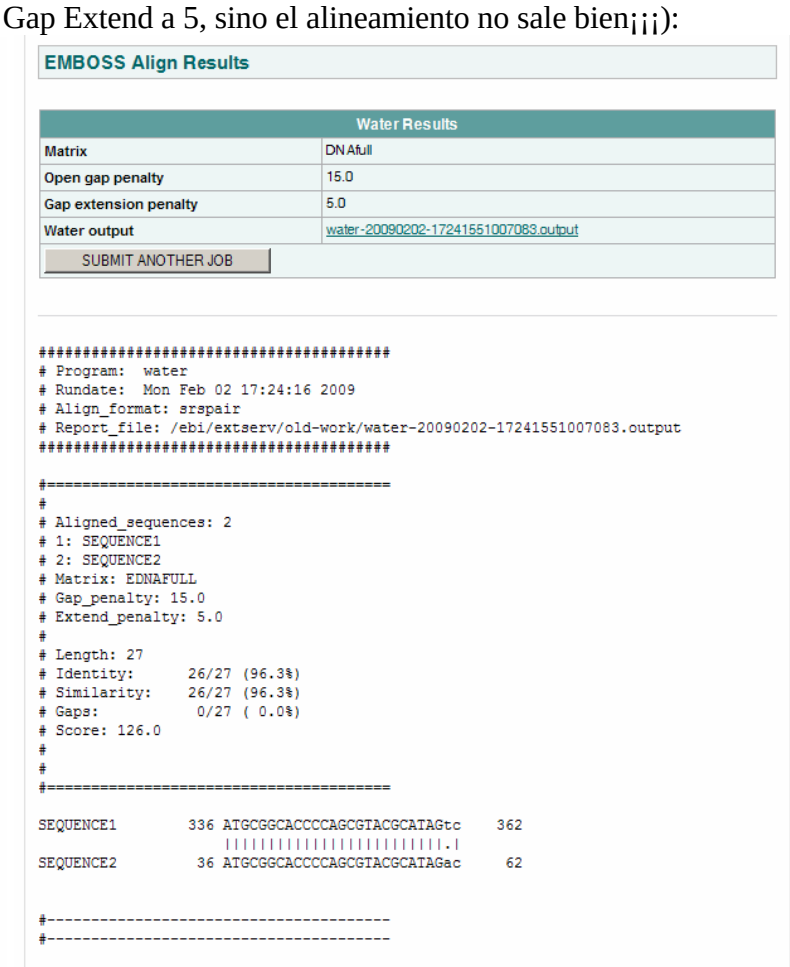


Ya obtenemos algo más normal:



Probamos con otro (EMBOSS algoritmo Water) (subiendo el Gap Extend a 5, sino el alineamiento no sale bien¡¡¡):





Los resultados son similares , pero no iguales.

2.1.6 Comparad los resultados del alineamiento global y local en los dos pares de secuencia: los 2 CDSs y las dos secuencias adjuntas. Decidid cual de los dos programas es el más adecuado para analizar cada uno de los dos casos.

El alineamiento global es el más adecuado para secuencias muy similares, tanto en su longitud como en su contenido. El local, en cambio, es más adecuado para localizar fragmentos cortos conservados en secuencias muy diferentes. Observad la salida de cada programa y juzgad que programa es el más adecuado desde el punto de vista de la interpretación biológica de los resultados.

En el caso de los dos CDS de Humano y Ratón es programa más adecuado es CLUSTALW ya que:

1. Son secuencias de homólogos, por lo tanto parecidas,
2. Son del mismo tamaño (aproximado).
3. Nos interesa una alineación global

En el caso de las secuencias problema" adjuntas" es mejor el BLASTN,

1. Aunque se ve a simple vista no tenemos claro donde está la parte codificante, o si la tiene, o si corresponde a algo.
2. Son secuencias cortas,
3. Nos interesa conocer si hay fragmentos comunes.

2.2 EJERCICIO 2: BLAST

2.2.1 Extraed del browser genómico UCSC http://genome.ucsc.edu/ la región codificante (CDS, solo exones) del gen URO-D humano.

Es la misma secuencia del bloque anterior.

2.2.2 Usando el programa BLAST para alinear 2 secuencias del bloque anterior, seleccionad la versión de BLAST (BLASTN,…) ideal para alinear el CDS del gen URO-D contra la proteína adjunta Ex2-prot.fa . Analizad el resultado en términos biológicos: estamos alineando una región codificante humana contra una proteína de otra especie. ¿Qué podéis decir de este alineamiento?

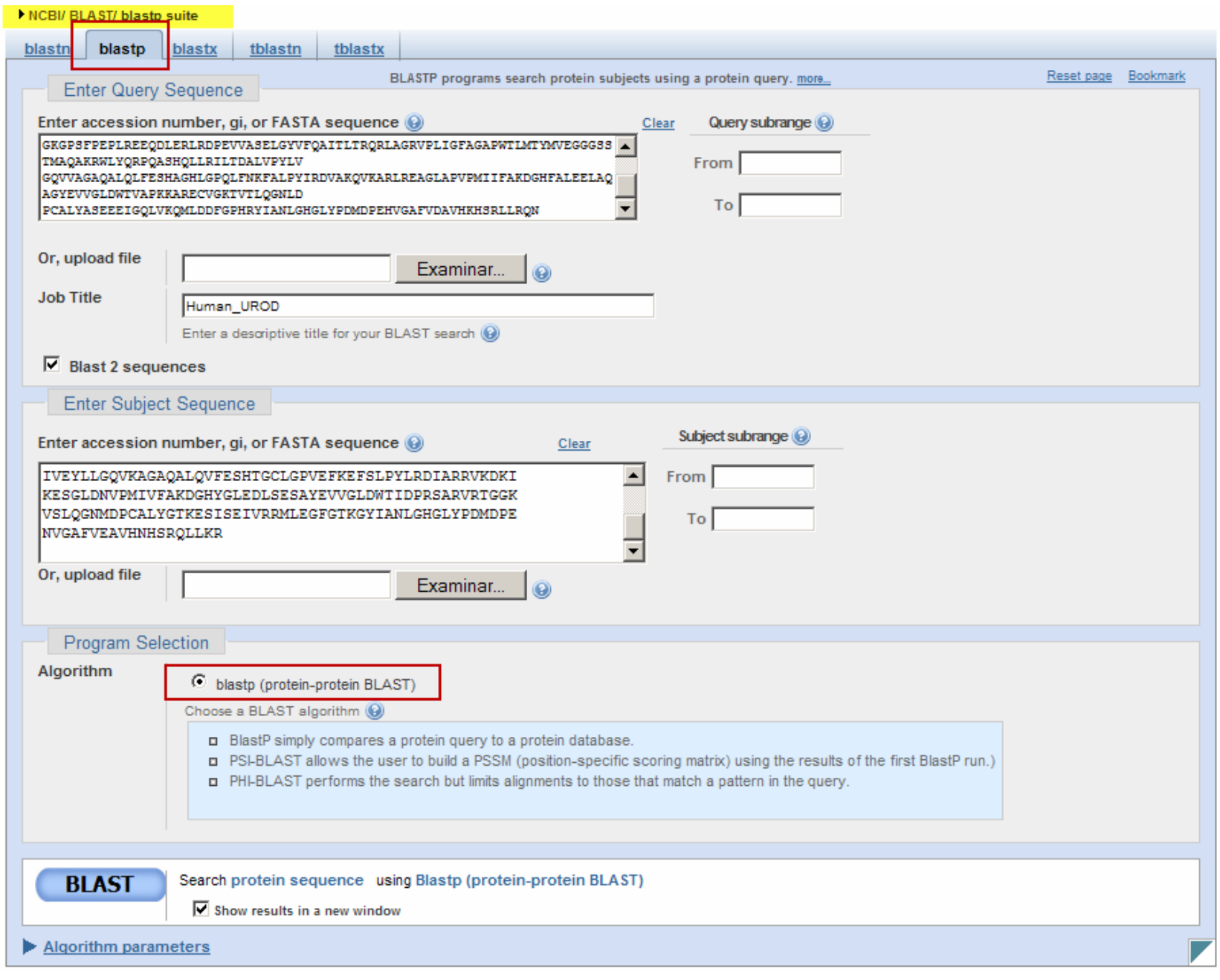
NOTA: la primera caja del formulario de BLAST representa la secuencia y la segunda caja representa la base de datos (que contiene solo la otra secuencia).

Dado que Ex2-prot.fa es una secuencia de aminoácidos, usaremos la versión de blast optimizada para proteínas.

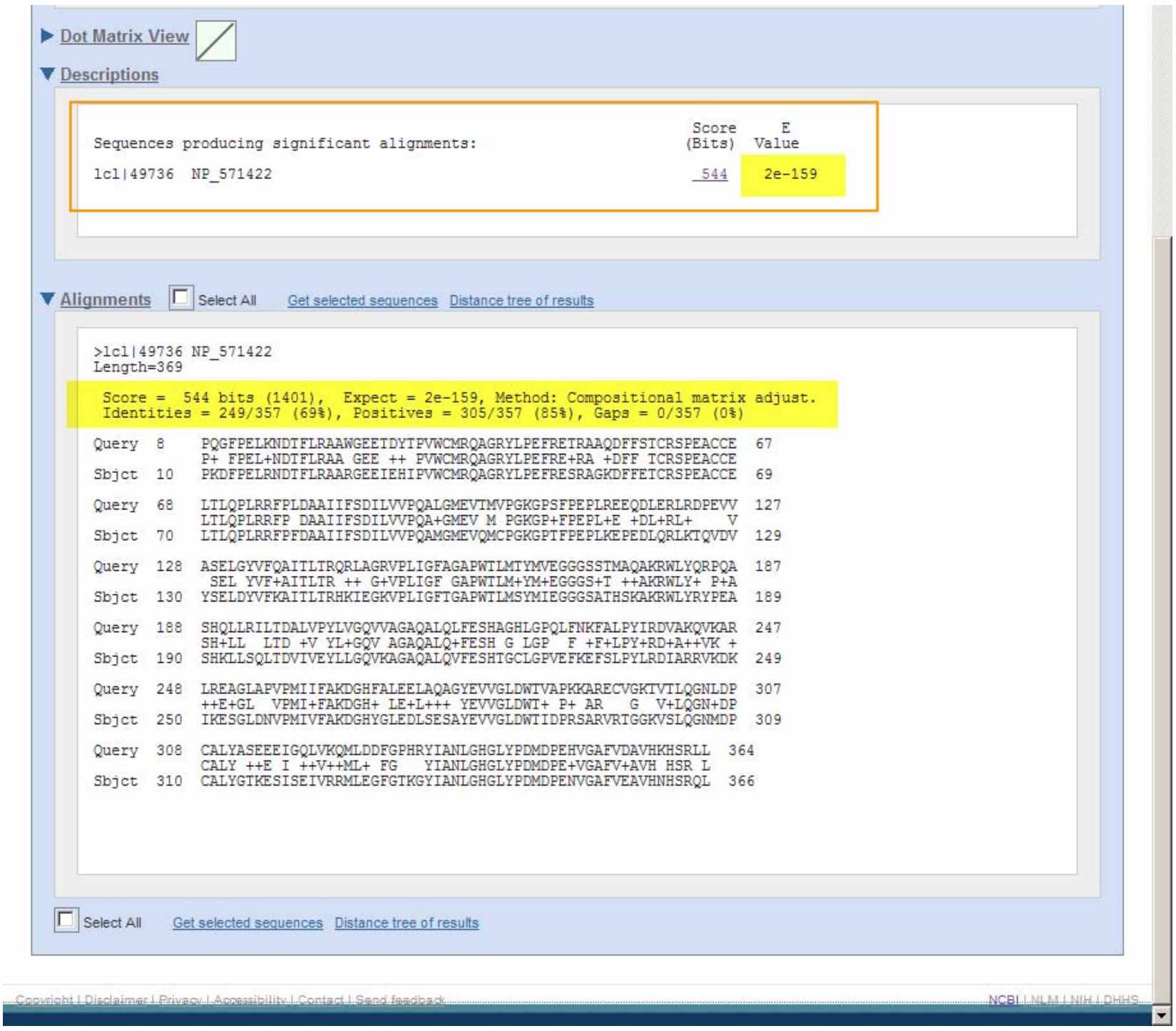
Primero creamos un fasta con la traducción a proteína del UROD humano. Ya la hemos obtenido en un apartado anterior.

>Human_UROD
MEANGLGPQGFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLPEFRETRAAQDFFSTCRSPEACCELTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVP
GKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRVPLIGFAGAPWTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLLRILTDALVPYLV
GQVVAGAQALQLFESHAGHLGPQLFNKFALPYIRDVAKQVKARLREAGLAPVPMIIFAKDGHFALEELAQAGYEVVGLDWTVAPKKARECVGKTVTLQGNLD
PCALYASEEEIGQLVKQMLDDFGPHRYIANLGHGLYPDMDPEHVGAFVDAVHKHSRLLRQN

Y usamos BLASTP para alinear las dos secuencias:



El alineamiento obtenido es el siguiente:



▶ Dot Matrix View

▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value
lcl|49736 NP_571422	544	2e-159

▼ Alignments Select All Get selected sequences Distance tree of results

```
>lcl|49736 NP_571422
Length=369

 Score =  544 bits (1401),  Expect = 2e-159, Method: Compositional matrix adjust.
 Identities = 249/357 (69%), Positives = 305/357 (85%), Gaps = 0/357 (0%)

Query  8    PQGFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLPEFRETRAAQDFFSTCRSPEACCE  67
            P+ FPEL+NDTFLRAA GEE ++ PVWCMRQAGRYLPEFRE+RA +DFF TCRSPEACCE
Sbjct  10   PKDFPELRNDTFLRAARGEEIEHIPVWCMRQAGRYLPEFRESRAGKDFFETCRSPEACCE  69

Query  68   LTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVPGKGPSFPEPLREEQDLERLRDPEVV  127
            LTLQPLRRFP DAAIIFSDILVVPQA+GMEV M PGKGP+FPEPL+E +DL+RL+    V
Sbjct  70   LTLQPLRRFPFDAAIIFSDILVVPQAMGMEVQMCPGKGPTFPEPLKEPEDLQRLKTQVDV  129

Query  128  ASELGYVFQAITLTRQRLAGRVPLIGFAGAPWTLMTYMVEGGGSSTMAQAKRWLYQRPQA  187
             SEL YVF+AITLTR ++ G+VPLIGF GAPWTLM+YM+EGGGS+T ++AKRWLY+ P+A
Sbjct  130  YSELDYVFKAITLTRHKIEGKVPLIGFTGAPWTLMSYMIEGGGSATHSKAKRWLYRYPEA  189

Query  188  SHQLLRILTDALVPYLVGQVVAGAQALQLFESHAGHLGPQLFNKFALPYIRDVAKQVKAR  247
            SH+LL  LTD +V YL+GQV AGAQALQ+FESH G LGP  F +F+LPY+RD+A++VK +
Sbjct  190  SHKLLSQLTDVIVEYLLGQVKAGAQALQVFESHTGCLGPVEFKEFSLPYLRDIARRVKDK  249

Query  248  LREAGLAPVPMIIFAKDGHFALEELAQAGYEVVGLDWTVAPKKARECVGKTVTLQGNLDP  307
            ++E+GL  VPMI+FAKDGH+ LE+L+++ YEVVGLDWT+ P+ AR   G  V+LQGN+DP
Sbjct  250  IKESGLDNVPMIVFAKDGHYGLEDLSESAYEVVGLDWTIDPRSARVRIGGKVSLQGNMDP  309

Query  308  CALYASEEEIGQLVKQMLDDFGPHRYIANLGHGLYPDMDPEHVGAFVDAVHKHSRLL  364
            CALY ++E I ++V++ML+ FG   YIANLGHGLYPDMDPE+VGAFV+AVH HSR L
Sbjct  310  CALYGTKESISEIVRRMLEGFGTKGYIANLGHGLYPDMDPENVGAFVEAVHNHSRQL  366
```

Select All Get selected sequences Distance tree of results

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback NCBI | NLM | NIH | DHHS

Podemos observar que:

- E-value extremadamente bajo: Seguramente son homologas.
- No hay gaps en el alineamiento y tiene un 69% de identidad, digamos que la conservación es media. Seguramente las zonas comunes serán los centros activos de la proteina.

Para corroborar las observaciones, veamos de quien es el Ex2_prot.fa

Hacemos BLASTP sobre la base de datos de proteinas NR (no redundante) y listamos los resultados por especie.
En amarillo resalto los que voy a comentar.

Organism Report

Danio rerio (leopard danio, ...) [bony fishes] taxid 7955		
ref|NP_571422.1| uroporphyrinogen decarboxylase [Danio rerio]	766	0.0
sp|Q9PTS2.1|DCUP_DANRE RecName: Full=Uroporphyrinogen deca...	766	0.0
gb|AAF14346.1|AF095639_1 uroporphyrinogen decarboxylase [D...	766	0.0
emb|CAX12678.1| uroporphyrinogen decarboxylase [Danio rerio]	766	0.0
gb|AAI08076.1| Uroporphyrinogen decarboxylase [Danio rerio]	761	0.0
gb|AAH92696.1| Urod protein [Danio rerio]	758	0.0
Salmo salar [bony fishes] taxid 8030		
ref|NP_001133714.1| Uroporphyrinogen decarboxylase [Salmo ...	654	0.0
gb|ACI33757.1| Uroporphyrinogen decarboxylase [Salmo salar]	654	0.0
Tetraodon nigroviridis [bony fishes] taxid 99883		
emb|CAG10903.1| unnamed protein product [Tetraodon nigrovi...	650	0.0
Xenopus (Silurana) tropicalis [frogs & toads] taxid 8364		
gb|AAH88815.1| Urod protein [Xenopus tropicalis]	602	2e-170
ref|NP_001011486.2| uroporphyrinogen decarboxylase [Xenopu...	601	3e-170

emb|CAJ82884.1| uroporphyrinogen decarboxylase [Xenopus tr...	601	3e-170
Xenopus laevis (common platanna, ...) [frogs & toads] taxid 8355		
ref|NP_001085981.1| MGC82980 protein [Xenopus laevis]	601	4e-170
gb|AAH73643.1| MGC82980 protein [Xenopus laevis]	601	4e-170
ref|NP_001084556.1| hypothetical protein LOC414506 [Xenopu...	598	4e-169
gb|AAH68896.1| MGC83088 protein [Xenopus laevis]	598	4e-169
Gallus gallus (bantam, ...) [birds] taxid 9031		
ref|XP_422430.2| PREDICTED: similar to LOC496978 protein [...	583	1e-164
Ovis aries (domestic sheep, ...) [even-toed ungulates] taxid 9940		
ref|NP_001012341.1| uroporphyrinogen decarboxylase [Ovis a...	548	2e-154
sp|Q8HY31.1|DCUP_SHEEP RecName: Full=Uroporphyrinogen deca...	548	2e-154
emb|CAC82649.1| uroporphyrinogen decarboxylase [Ovis aries]	548	2e-154
Bos taurus (cow, ...) [even-toed ungulates] taxid 9913		
ref|XP_581108.4| PREDICTED: similar to uroporphyrinogen de...	548	2e-154
Pan troglodytes [primates] taxid 9598		
ref|XP_513127.1| PREDICTED: uroporphyrinogen decarboxylase...	545	3e-153
ref|XP_001154880.1| PREDICTED: uroporphyrinogen decarboxyl...	545	3e-153
ref|XP_001154766.1| PREDICTED: uroporphyrinogen decarboxyl...	492	3e-137
ref|XP_001154586.1| PREDICTED: uroporphyrinogen decarboxyl...	470	8e-131
ref|XP_001154132.1| PREDICTED: similar to Chain A, Phe232l...	463	9e-129
ref|XP_001153755.1| PREDICTED: similar to Chain A, Phe232l...	456	2e-126
ref|XP_001154358.1| PREDICTED: uroporphyrinogen decarboxyl...	449	3e-124
ref|XP_001154711.1| PREDICTED: uroporphyrinogen decarboxyl...	433	1e-119
synthetic construct [other sequences] taxid 32630		
gb|AAX37109.1| uroporphyrinogen decarboxylase [synthetic c...	544	5e-153
gb|AAP36644.1| Homo sapiens uroporphyrinogen decarboxylase...	541	4e-152
gb|AAX43947.1| uroporphyrinogen decarboxylase [synthetic c...	541	4e-152
gb|AAX32348.1| uroporphyrinogen decarboxylase [synthetic c...	541	4e-152
gb|AAX32349.1| uroporphyrinogen decarboxylase [synthetic c...	541	4e-152
gb|ABM81614.1| uroporphyrinogen decarboxylase [synthetic c...	541	4e-152
gb|ABM84796.1| uroporphyrinogen decarboxylase [synthetic c...	541	4e-152
Mus musculus (mouse) [rodents] taxid 10090		
ref|NP_033504.2| uroporphyrinogen decarboxylase [Mus muscu...	544	5e-153
sp|P70697.2|DCUP_MOUSE RecName: Full=Uroporphyrinogen deca...	544	5e-153
gb|AAH08109.1| Uroporphyrinogen decarboxylase [Mus musculus]	544	5e-153
dbj|BAE43002.1| unnamed protein product [Mus musculus]	544	5e-153
emb|CAM23809.1| uroporphyrinogen decarboxylase [Mus musculus]	544	5e-153
gb|EDL30576.1| mCG14438 [Mus musculus]	543	9e-153
gb|AAB18294.1| uroporphyrinogen decarboxylase	540	1e-151
Homo sapiens (man) [primates] taxid 9606		
ref|NP_000365.3| uroporphyrinogen decarboxylase [Homo sapi...	544	6e-153
sp|P06132.2|DCUP_HUMAN RecName: Full=Uroporphyrinogen deca...	544	6e-153
pdb|1URO|A Chain A, Uroporphyrinogen Decarboxylase	544	6e-153
pdb|1R3Q|A Chain A, Uroporphyrinogen Decarboxylase In Comp...	544	6e-153
pdb|1R3Y|A Chain A, Uroporphyrinogen Decarboxylase In Comp...	544	6e-153
gb|AAC03563.1| uroporphyrinogen decarboxylase [Homo sapiens]	544	6e-153
emb|CAG33257.1| UROD [Homo sapiens]	544	6e-153
emb|CAG46854.1| UROD [Homo sapiens]	544	6e-153
emb|CAI16440.1| uroporphyrinogen decarboxylase [Homo sapiens]	544	6e-153
gb|EAX07007.1| uroporphyrinogen decarboxylase, isoform CRA...	544	6e-153
dbj|BAF84566.1| unnamed protein product [Homo sapiens]	544	6e-153
pdb|1JPH|A Chain A, Ile260thr Mutant Of Human Urod, Human ...	543	1e-152
pdb|1R3R|A Chain A, Uroporphyrinogen Decarboxylase With Mu...	543	1e-152
pdb|1R3V|A Chain A, Uroporphyrinogen Decarboxylase Single ...	543	1e-152
pdb|1R3W|A Chain A, Uroporphyrinogen Decarboxylase Y164f M...	542	1e-152
pdb|1JPI|A Chain A, Phe232leu Mutant Of Human Urod, Human ...	542	2e-152
pdb|1R3S|A Chain A, Uroporphyrinogen Decarboxylase Single ...	541	4e-152
pdb|1R3T|A Chain A, Uroporphyrinogen Decarboxylase Single ...	541	4e-152
gb|AAH01778.1| Uroporphyrinogen decarboxylase [Homo sapiens]	541	4e-152
gb|AAP35383.1| uroporphyrinogen decarboxylase [Homo sapiens]	541	4e-152
pdb|1JPK|A Chain A, Gly156asp Mutant Of Human Urod, Human ...	541	4e-152
gb|AAC50482.1| uroporphyrinogen decarboxylase	540	6e-152
emb|CAA61540.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04571.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04572.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04573.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04574.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04575.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04576.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04577.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04578.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04579.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04580.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04581.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04582.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04583.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04584.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04585.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04586.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04587.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04588.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04589.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAD04590.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAP44118.1| uroporphyrinogen decarboxylase [Homo sapiens]	540	6e-152
gb|AAA61258.1| uroporphyrinogen decarboxylase (EC 4.1.1.37)	540	1e-151
pdb|2Q71|A Chain A, Uroporphyrinogen Decarboxylase G168r S...	538	3e-151
pdb|2Q6Z|A Chain A, Uroporphyrinogen Decarboxylase G168r S...	538	3e-151
Equus caballus (equine, ...) [odd-toed ungulates] taxid 9796		
ref|XP_001496410.2| PREDICTED: similar to Uroporphyrinogen...	541	3e-152

Pongo abelii (Orang-utan, ...) [primates] taxid 9601		
ref|NP_001129018.1| uroporphyrinogen decarboxylase [Pongo ...	541	3e-152
sp|Q5RDK5.1|DCUP_PONAB RecName: Full=Uroporphyrinogen deca...	541	3e-152
emb|CAH90152.1| hypothetical protein [Pongo abelii]	541	3e-152
Rattus norvegicus (brown rat, ...) [rodents] taxid 10116		
ref|NP_062082.1| uroporphyrinogen decarboxylase [Rattus no...	541	4e-152
gb|EDL90239.1| uroporphyrinogen decarboxylase, isoform CRA...	541	4e-152
gb|AAI58691.1| Uroporphyrinogen decarboxylase [Rattus norv...	541	4e-152
prf||1310344A decarboxylase,uroporphyrinogen	526	2e-147
sp|P32362.1|DCUP_RAT RecName: Full=Uroporphyrinogen decarb...	523	1e-146
emb|CAB50784.1| uroporphyrinogen decarboxylase [Rattus nor...	523	1e-146
Monodelphis domestica [marsupials] taxid 13616		
ref|XP_001375996.1| PREDICTED: similar to uroporphyrinogen...	535	2e-150
Canis lupus familiaris (dogs) [carnivores] taxid 9615		
ref|XP_532602.2| PREDICTED: similar to Uroporphyrinogen de...	530	6e-149
ref|XP_861812.1| PREDICTED: similar to Uroporphyrinogen de...	500	8e-140
ref|XP_861894.1| PREDICTED: similar to Uroporphyrinogen de...	488	3e-136
ref|XP_861922.1| PREDICTED: similar to Uroporphyrinogen de...	458	3e-127

Nuestro nuevo amigo:

About the Zebrafish July 2007 (danRer5) assembly (sequences)

The July 2007 zebrafish (*Danio rerio*) Zv7 assembly was produced by The Wellcome Trust Sanger Institute in collaboration with the Max Planck Institute for Developmental Biology in Tuebingen, Germany, and the Netherlands Institute for Developmental Biology (Hubrecht Laboratory), Utrecht, The Netherlands.

Danio rerio
Photo courtesy of NHGRI (Press Photos)

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic region, an mRNA or EST, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the zebrafish genome. Note that some position queries (e.g. "huntington") may return matches to the mRNA records of other species. In these cases, the mRNAs are mapped to their homologs in zebrafish. See the User's Guide for more information.

Vemos que el alineamiento con UROD de Danio rerio coincide al 100%

GENE ID: 30617 urod | uroporphyrinogen decarboxylase [Danio rerio]
(10 or fewer PubMed links)

Score = 766 bits (1978), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 369/369 (100%), Positives = 369/369 (100%), Gaps = 0/369 (0%)

```
Query  1    MMDKDSFILPKDFPELRNDTFLRAARGEEIEHIPVWCMRQAGRYLPEFRESRAGKDFFET  60
            MMDKDSFILPKDFPELRNDTFLRAARGEEIEHIPVWCMRQAGRYLPEFRESRAGKDFFET
Sbjct  1    MMDKDSFILPKDFPELRNDTFLRAARGEEIEHIPVWCMRQAGRYLPEFRESRAGKDFFET  60

Query  61   CRSPEACCELTLQPLRRFPFDAAIIFSDILVVPQAMGMEVQMCPGKGPTFPEPLKEPEDL  120
            CRSPEACCELTLQPLRRFPFDAAIIFSDILVVPQAMGMEVQMCPGKGPTFPEPLKEPEDL
Sbjct  61   CRSPEACCELTLQPLRRFPFDAAIIFSDILVVPQAMGMEVQMCPGKGPTFPEPLKEPEDL  120

Query  121  QRLKTQVDVYSELDYVFKAITLTRHKIEGKVPLIGFTGAPWTLMSYMIEGGGSATHSKAK  180
            QRLKTQVDVYSELDYVFKAITLTRHKIEGKVPLIGFTGAPWTLMSYMIEGGGSATHSKAK
Sbjct  121  QRLKTQVDVYSELDYVFKAITLTRHKIEGKVPLIGFTGAPWTLMSYMIEGGGSATHSKAK  180

Query  181  RWLYRYPEASHKLLSQLTDVIVEYLLGQVKAGAQALQVFESHTGCLGPVEFKEFSLPYLR  240
            RWLYRYPEASHKLLSQLTDVIVEYLLGQVKAGAQALQVFESHTGCLGPVEFKEFSLPYLR
Sbjct  181  RWLYRYPEASHKLLSQLTDVIVEYLLGQVKAGAQALQVFESHTGCLGPVEFKEFSLPYLR  240

Query  241  DIARRVKDKIKESGLDNVPMIVFAKDGHYGLEDLSESAYEVVGLDWTIDPRSARVRTGGK  300
            DIARRVKDKIKESGLDNVPMIVFAKDGHYGLEDLSESAYEVVGLDWTIDPRSARVRTGGK
Sbjct  241  DIARRVKDKIKESGLDNVPMIVFAKDGHYGLEDLSESAYEVVGLDWTIDPRSARVRTGGK  300

Query  301  VSLQGNMDPCALYGTKESISEIVRRMLEGFGTKGYIANLGHGLYPDMDPENVGAFVEAVH  360
            VSLQGNMDPCALYGTKESISEIVRRMLEGFGTKGYIANLGHGLYPDMDPENVGAFVEAVH
Sbjct  301  VSLQGNMDPCALYGTKESISEIVRRMLEGFGTKGYIANLGHGLYPDMDPENVGAFVEAVH  360

Query  361  NHSRQLLKR  369
            NHSRQLLKR
Sbjct  361  NHSRQLLKR  369
```

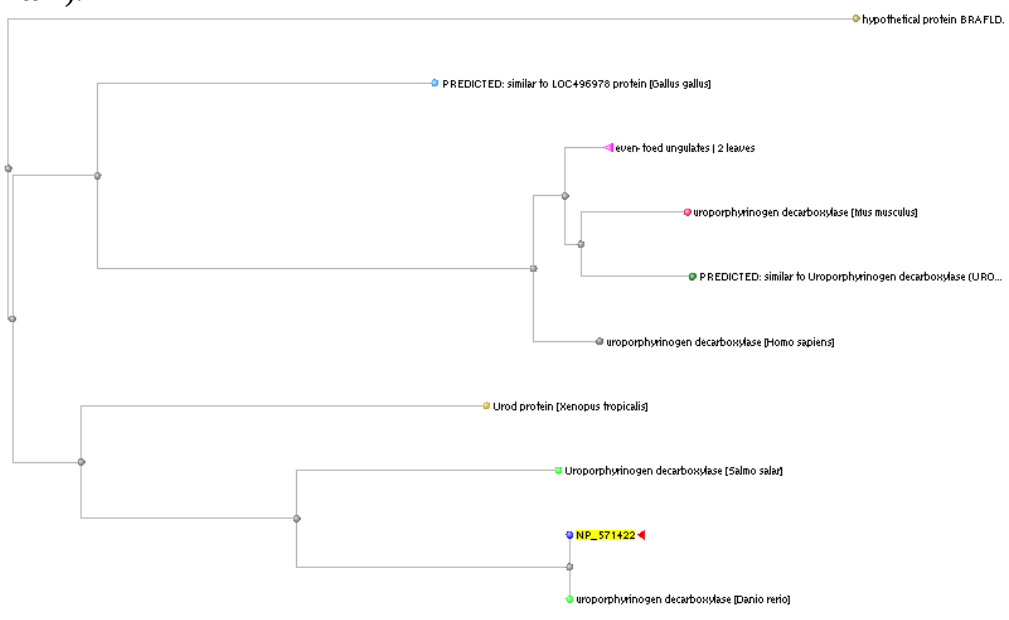
Este es alineamiento con Homo sapiens

```
 GENE ID: 7389 UROD | uroporphyrinogen decarboxylase [Homo sapiens]
(Over 10 PubMed links)

 Score =  544 bits (1401),  Expect = 6e-153, Method: Compositional matrix adjust.
 Identities = 249/357 (69%), Positives = 305/357 (85%), Gaps = 0/357 (0%)

Query  10   PKDFPELRNDTFLRAARGEEIEHIPVWCMRQAGRYLPEFRESRAGKDFFETCRSPEACCE  69
            P+ FPEL+NDTFLRAA GEE ++ PVWCMRQAGRYLPEFRE+RA +DFF TCRSPEACCE
Sbjct  8    PQGFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLPEFRETRAAQDFFSTCRSPEACCE  67

Query  70   LTLQPLRRFPFDAAIIFSDILVVPQAMGMEVQMCPGKGPTFPEPLKEPEDLQRLKTQVDV  129
            LTLQPLRRFP DAAIIFSDILVVPQA+GMEV M PGKGP+FPEPL+E +DL+RL+    V
Sbjct  68   LTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVPGKGPSFPEPLREEQDLERLRDPEVV  127

Query  130  YSELDYVFKAITLTRHKIEGKVPLIGFTGAPWTLMSYMIEGGGSATHSKAKRWLYRYPEA  189
             SEL YVF+AITLTR ++ G+VPLIGF GAPWTLM+YM+EGGGS+T ++AKRWLY+ P+A
Sbjct  128  ASELGYVFQAITLTRQRLAGRVPLIGFAGAPWTLMTYMVEGGGSSTMAQAKRWLYQRPQA  187

Query  190  SHKLLSQLTDVIVEYLLGQVKAGAQALQVFESHTGCLGPVEFKEFSLPYLRDIARRVKDK  249
            SH+LL  LTD +V YL+GQV AGAQALQ+FESH G LGP  F +F+LPY+RD+A++VK +
Sbjct  188  SHQLLRILTDALVPYLVGQVVAGAQALQLFESHAGHLGPQLFNKFALPYIRDVAKQVKAR  247

Query  250  IKESGLDNVPMIVFAKDGHYGLEDLSESAYEVVGLDWTIDPRSARVRTGGKVSLQGNMDP  309
            ++E+GL  VPMI+FAKDGH+ LE+L+++ YEVVGLDWT+ P+ AR   G  V+LQGN+DP
Sbjct  248  LREAGLAPVPMIIFAKDGHFALEELAQAGYEVVGLDWTVAPKKARECVGKTVTLQGNLDP  307

Query  310  CALYGTKESISEIVRRMLEGFGTKGYIANLGHGLYPDMDPENVGAFVEAVHNHSRQL  366
            CALY ++E I ++V++ML+ FG   YIANLGHGLYPDMDPE+VGAFV+AVH HSR L
Sbjct  308  CALYASEEEIGQLVKQMLDDFGPHRYIANLGHGLYPDMDPEHVGAFVDAVHKHSRLL  364
```

Y este con el Ratón.

```
 GENE ID: 22275 Urod | uroporphyrinogen decarboxylase [Mus musculus]
(Over 10 PubMed links)

 Score =  544 bits (1401),  Expect = 5e-153, Method: Compositional matrix adjust.
 Identities = 251/365 (68%), Positives = 306/365 (83%), Gaps = 1/365 (0%)

Query  2    MDKDSFILPKDFPELRNDTFLRAARGEEIEHIPVWCMRQAGRYLPEFRESRAGKDFFETC  61
            M+ + F L ++FPEL+NDTFLRAA GEE ++ PVWCMRQAGRYLPEFRE+RA +DFF TC
Sbjct  1    MEANGFGL-QNFPELKNDTFLRAAWGEETDYTPVWCMRQAGRYLPEFRETRAAQDFFSTC  59

Query  62   RSPEACCELTLQPLRRFPFDAAIIFSDILVVPQAMGMEVQMCPGKGPTFPEPLKEPEDLQ  121
            RSPEACCELTLQPLRRFP DAAIIFSDILVVPQA+GMEV M PGKGP+FPEPL+E  DL+
Sbjct  60   RSPEACCELTLQPLRRFPLDAAIIFSDILVVPQALGMEVTMVPGKGPSFPEPLREERDLE  119

Query  122  RLKTQVDVYSELDYVFKAITLTRHKIEGKVPLIGFTGAPWTLMSYMIEGGGSATHSKAKR  181
            RL+      SEL YVF+AITLTR ++ G+VPLIGF GAPWTLM+YM+EGG S+T ++AKR
Sbjct  120  RLRDPAAAASELGYVFQAITLTRQRLAGRVPLIGFAGAPWTLMTYMVEGGSSSTMAQAKR  179

Query  182  WLYRYPEASHKLLSQLTDVIVEYLLGQVKAGAQALQVFESHTGCLGPVEFKEFSLPYLRD  241
            WLY+ P+ASHKLL  LTDV+V YL+GQV AGAQALQ+FESH G LG   F +F+LPY+RD
Sbjct  180  WLYQRPQASHKLLGILTDVLVPYLIGQVAAGAQALQLFESHAGHLGTELFSKFALPYIRD  239

Query  242  IARRVKDKIKESGLDNVPMIVFAKDGHYGLEDLSESAYEVVGLDWTIDPRSARVRTGGKV  301
            +A+RVK  ++++GL  VPMI+FAKDGH+ LE+L+++ YEVVGLDWT+ P+ AR R G  V
Sbjct  240  VAKRVKAGLQKAGLAPVPMIIFAKDGHFALEELAQAGYEVVGLDWTVAPKKARERVGKAV  299

Query  302  SLQGNMDPCALYGTKESISEIVRRMLEGFGTKGYIANLGHGLYPDMDPENVGAFVEAVHN  361
            +LQGN+DPCALY ++E I  +V++ML+ FG + YIANLGHGLYPDMDPE VGAFV+AVH
Sbjct  300  TLQGNLDPCALYASEEEIGRLVQQMLDDFGPQRYIANLGHGLYPDMDPERVGAFVDAVHK  359

Query  362  HSRQL  366
            HSR L
Sbjct  360  HSRLL  364
```

Vemos que el % de conservación es similar en ambos (tb en el de la Rattus), lo cual indica que las zonas conservadas son posiblemente las activas (harían falta un análisis estructural y de reactividad para confirmarlo, aquí únicamente lo podemos intuir).



Aquí el árbol.

2.2.3 Ahora debéis realizar el alineamiento con el programa BLASTN de las secuencias adjuntas Ex2-genomicA.fa i Ex2-genomicB.fa

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help | My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite-2sequences/ Formatting Results - SDFPJANH113

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download

Blast 2 sequences

HumanSeq (35001 letters)

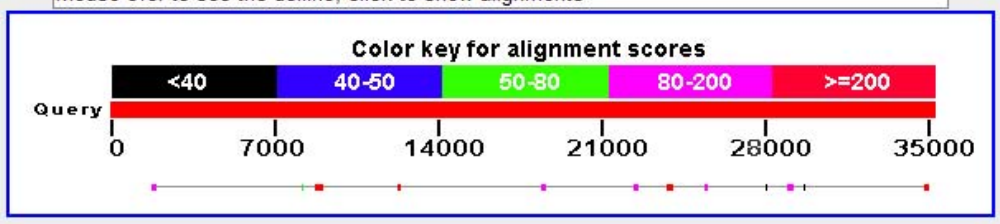
Query ID	lcl|54497	**Subject ID**	54499
Description	HumanSeq	**Description**	MouseSeq
Molecule type	nucleic acid	**Molecule type**	nucleic acid
Query Length	35001	**Subject Length**	26501
		Program	BLASTN 2.2.19+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports]

▼ Graphic Summary

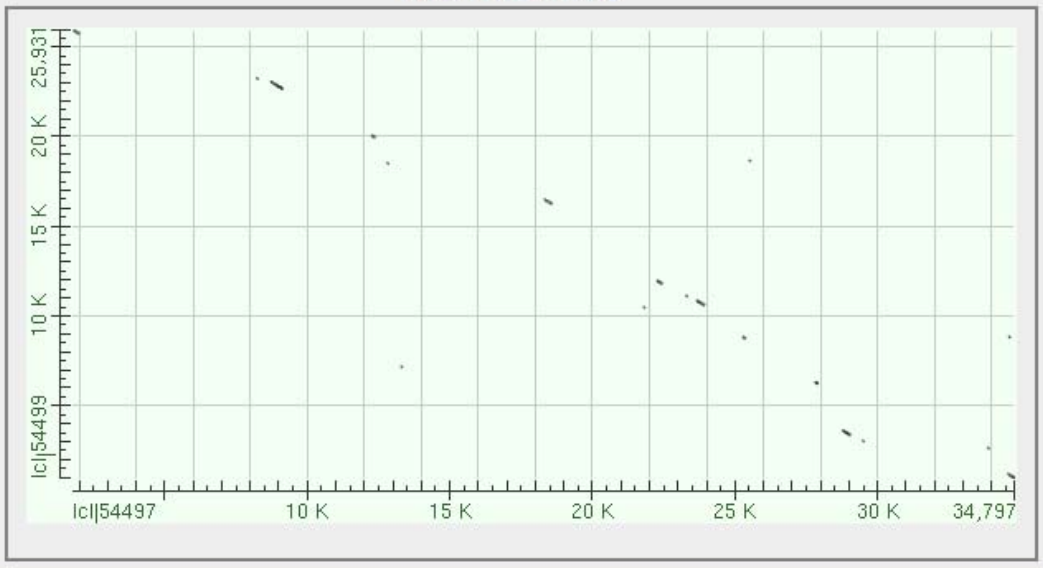
Distribution of 19 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



▼ Dot Matrix View

Plot of lcl|54497 vs 54499



▼ Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
54499	MouseSeq	248	1950	6%	2e-66	100%	

▶ Alignments

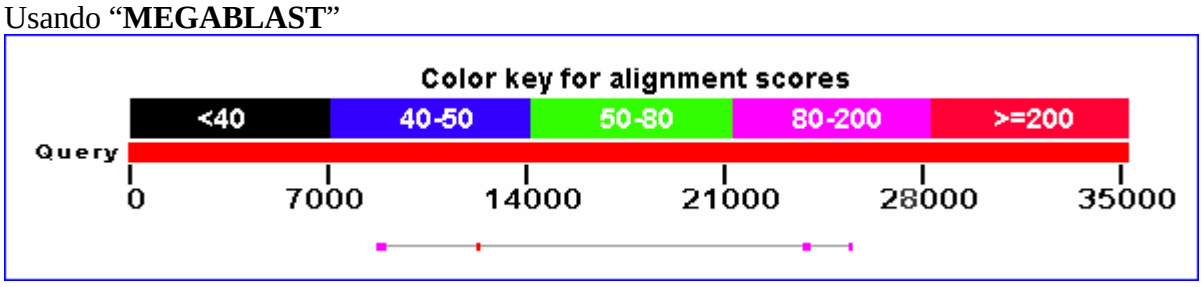
En el grafico de puntos ya observamos que a pesar de no ser del mismo tamaño se forma una diagonal que muestra las zonas alineadas.

Y algunos de los alineamientos locales

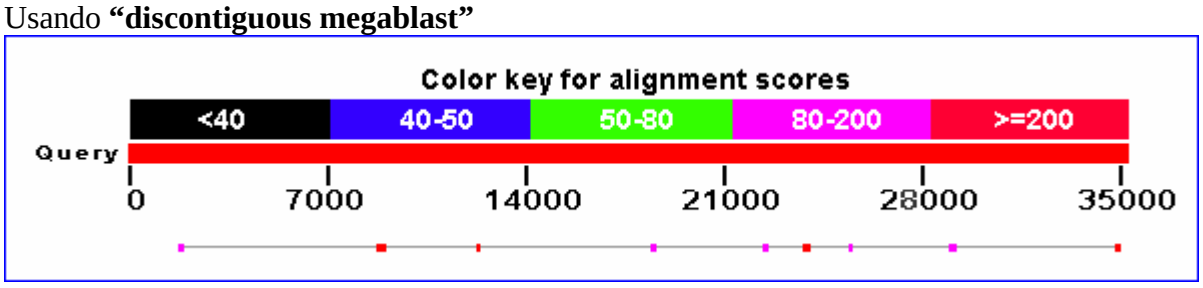
```
>lcl|54499 MouseSeq
Length=26501

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start
position

 Score =  248 bits (274),  Expect = 2e-66
 Identities = 306/409 (74%), Gaps = 32/409 (7%)
 Strand=Plus/Minus

Query  8738   TTTAGTCAGTAttttttGTGCAGCTTGGAAAGTGAAGAGTAACTGTCttttttt-----  8792
              ||| |||||||| | |  |||||| ||||||||| |||| |||||  ||||||||
Sbjct  23044  TTTTGTCAGTATGTGT--GTGCAGTTTGGAAAGTAAAGACTAACTACCTTTTTTTCTTTG  22987

Query  8793   --tCTGTCTGTTACAGGAAACGGGATTATGAAGGTTATTTATGCTCCCTGCTGCTCCCTG  8850
                |||||  ||||  |||||||| ||||||||| | || || || ||| | ||| ||||||
Sbjct  22986  TTTCTGTTTGTTGTAGGAAACGAGATTATGAAAGCTACTTGTGTTCCTTACTGTTCCCTG  22927

Query  8851   CAGAATCCCGAAGCTCTGTTTTTGCACTGAGGGCCTTTAATGTGGAACTGGCTCAGGCTG  8910
              |||| |  | ||| |||| || ||||||| ||||||| ||||||||||||||||||||
Sbjct  22926  CAGAGTGTCAAAGGTCTGCTTCTGCACTGCGGGCCTTCAATGTGGAACTGGCTCAGGCTG  22867

Query  8911   GTATTAAGATACCTTAAAATATTATTTGCGAAATGGTGATATAAGGTGT--TTAATGCTG  8968
              ||||||| ||||||||||||||| |  | | |||||  | || |||||  ||  | |||
Sbjct  22866  GTATTAAAATACCTTAAAATATTA-TGACAAGATGGT--TGTAGGGTGTGCTTGGTACTG  22810

Query  8969   AACAATAAAGAAATATAGTTGTAATTTATATGTTAGATGTGTTTAAGGTCTCTGCTAGTG  9028
              || | |||||||||| |   ||||||   ||| |||||||  |||| ||   ||||||||
Sbjct  22809  AATAGTAAAGAAATAGA---GTAATT--CATGGTAGATGTACTTAAAGTTGTTGCTAGTG  22755

Query  9029   ATTTCTTTTCCCTGCTAACTGTTTAGGCTTAGTAATTCAATACCAA-AAATAAAACTA-T  9086
               |||||||  | ||||| ||  | || ||   ||||| | || ||| |||||||||| |
Sbjct  22754  GTTTCTTTCTCATGCTAGCTACTCAGCCT---TAATT-AGTATCAAGAAATAAAACTAGT  22699

Query  9087   CTA-------ATGAACTTAATGGTTCTAGAAATGCCATTCTGGAATTTT  9128
              | |       ||||| | ||||| ||| | ||||  ||||| || ||||
Sbjct  22698  CAATGGGTCTATGAATTCAATGGGTCTGGGAATGATATTCTAGACTTTT  22650


 Score =  226 bits (250),  Expect = 5e-60
 Identities = 188/226 (83%), Gaps = 8/226 (3%)
 Strand=Plus/Minus

Query  34578  TGATCATTTTATGATCTGGTGATCACTATTCTCTTTTTTCCAGGTTTCTCTAGAGGACTT  34637
              |||||||||||||||| |||| | ||  ||||||| | ||||||||||||||||||||
Sbjct  1196   TGATCATTTTATGATCTAGTGACCGCTGCTCTCTTTCT-CCAGGTTTCTCTAGAGGACTA  1138

Query  34638  TCTAAAGAAAATTCAGCGAGTGGATTTTGATATATTCCACCCATCTTTACAGCAGAAGAA  34697
               ||||||||||| |||||||| || |||||||||||||| || ||||||||||||||||
Sbjct  1137   CTTAAAGAAAATCCAGCGAGTAGACTTTGATATATTCCATCCGTCTTTACAGCAGAAGAA  1078

Query  34698  TACATTACTTCCATTATATTTGTATATTCAGTCATGGAGAAAAACATATTAAAATAATTT  34757
               |  |||||||||||| |||||||||||| |||||||||||||| ||| |||||| ||||
Sbjct  1077   CATGTTACTTCCATTATCTTTGTATATTCAATCATGGAGAAAAAGATACTAAAATGATTT  1018

Query  34758  CATG--GC----ATTGATGTTAATTCTAGTCTATTAGTTTTATAAA  34797
               |||  ||    | |||| ||| || |||| |||| |||| | |||
Sbjct  1017   TATGTTGCTCATACTGAT-TTACTTTTAGTTTATTGGTTTAAAAAA  973


 Score =  224 bits (248),  Expect = 2e-59
 Identities = 225/289 (77%), Gaps = 15/289 (5%)
 Strand=Plus/Minus

Query  23664  TTTCACTCATAAATCCTGTTTTTCAGATGTAAAACAGGATAAATGTCTCCTTTTACATG-  23722
              ||||  ||||||||| |||||| ||||||| | ||||    ||||  | | |||||| |
Sbjct  10855  TTTCCATCATAAATCGTGTTTTCCAGATGTCAGACAG----AATGGATACCTTTACAGGG  10800

Query  23723  TTTAGGTTATTTCTCTATGCTATTTCTATTGATTTATTATGCACTTAG---AATAGAGCA  23779
              ||||||| || |||   ||    | || ||||||| ||||||    ||   || ||||||
Sbjct  10799  TTTAGGTGATGTCT---TG----TGCTCTTGATTTGTTATGCTTCAAGGACAACAGAGCA  10747

Query  23780  GCCCTGTGTAATTTCTGAAATCATAGGTATAAAGGATCTTCATGCAGATCATGCTGCAAG  23839
```

```
             ||||||  |  ||||||   |  ||||  | ||||||||||||||||||||||||||||||
Sbjct  10746  ACCCTGTTTGGTTTCTGTGGTTCTAGGCGTGAAGGATCTTCATGCAGATCATGCTGCAAG  10687

Query  23840  TCATATTGGAAAAGCACAAGGCATTGTCACTTGCTTGAGAGCAACACCATATCATGGGAG  23899
              || ||||||||  |||||||||||||||||  |||||||||||||||||  |||||  | ||
Sbjct  10686  CCACATTGGAAAGGCACAAGGCATTGTCACGTGCTTGAGAGCAACACCCTATCACAGTAG  10627

Query  23900  CAGAAGAAAGGTGTTCCTTCCCATGGATATTTGTATGCTGGTAAGGCTG  23948
              ||| ||| ||||||||||  |||||||||  | ||| ||| |||  ||||||
Sbjct  10626  CAGGAGACAGGTGTTCCTGCCCATGGATGTCTGTGTGCAGGTGAGGCTG  10578
```

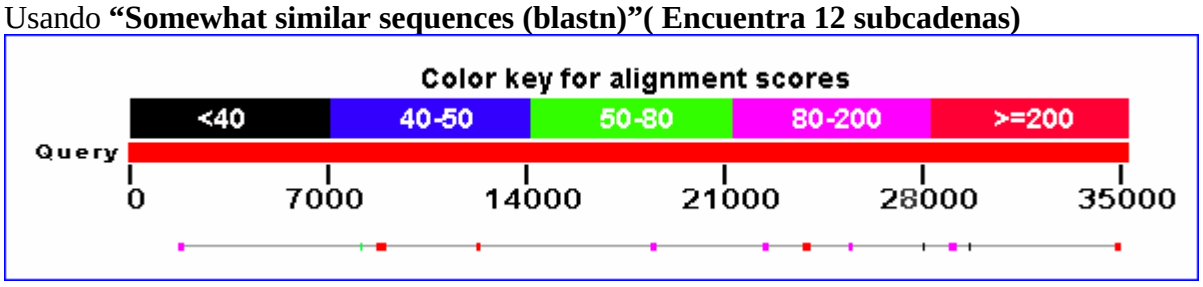
Resumen usando distintas opciones
Usando "**MEGABLAST**"



Usando **"discontiguous megablast"**



Usando **"Somewhat similar sequences (blastn)"(Encuentra 12 subcadenas)**



2.2.4 Ídem pero usando ahora el programa TBLASTX con las misma secuencias adjuntas Ex2-genomicA.fa i Ex2-genomicB.fa

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ tblastx-2sequences/ Formatting Results - SDG1DP9K11R

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download

Blast 2 sequences

HumanSeq

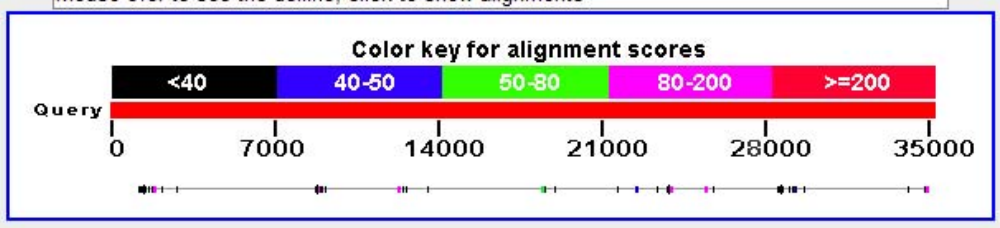
Query ID	lcl|15337	**Subject ID**	15339
Description	HumanSeq	**Description**	MouseSeq
Molecule type	nucleic acid	**Molecule type**	nucleic acid
Query Length	35001	**Subject Length**	26501
		Program	TBLASTX 2.2.19+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports]

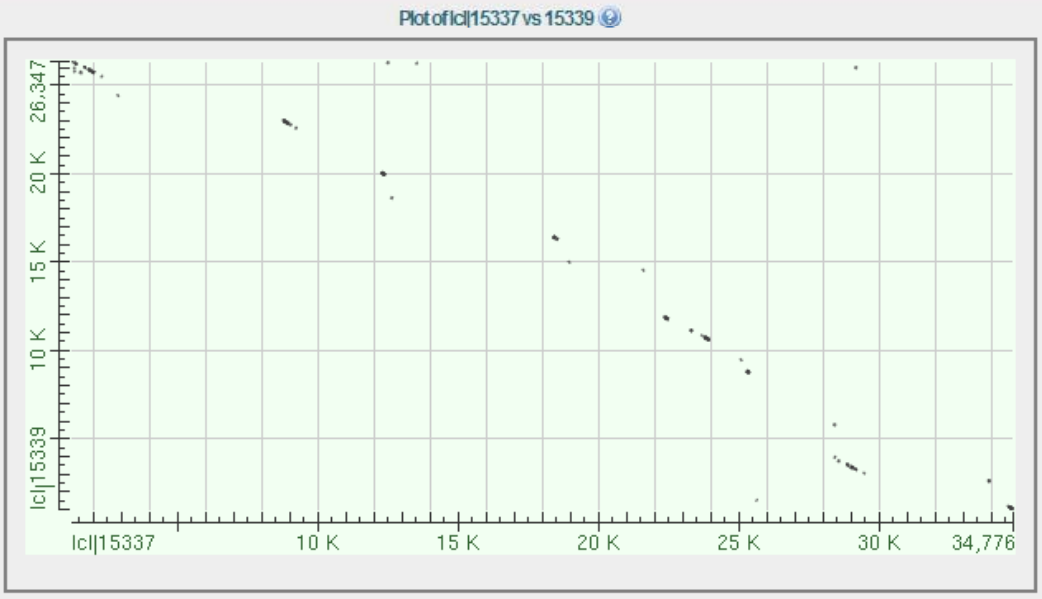
▼ Graphic Summary

Distribution of 104 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



▼ Dot Matrix View



▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	N
lcl|15339 MouseSeq	131	6e-175	23

▶ Alignments

Se observa claramente que BLASTX “encuentra” mas trozos en común que BLASTN

```
>lcl|15339 MouseSeq
Length=26501
```

```
 Score =  131 bits (280),  Expect(23) = 6e-175
 Identities = 49/70 (70%), Positives = 61/70 (87%), Gaps = 0/70 (0%)
 Frame = +3/-1

Query  23739  MLFLLIYYALRIEQPCVISEIIGIKDLHADHAASHIGKAQGIVTCLRATPYHGSRRKVFL  23918
              ++ L+ Y +   EQPC++S ++G+KDLHADHAASHIGKAQGIVTCLRATPYH SRR+VFL
Sbjct  10787  LVLLICYASRTTEQPCLVSVVLGVKDLHADHAASHIGKAQGIVTCLRATPYHSSRRQVFL  10608

Query  23919  PMDICMLVRL  23948
              PMD+C+ VRL
Sbjct  10607  PMDVCVQVRL  10578


 Score =  112 bits (239),  Expect(23) = 6e-175
 Identities = 47/55 (85%), Positives = 50/55 (90%), Gaps = 0/55 (0%)
 Frame = +1/-1

Query  34612  FFQVSLEDFLKKIQRVDFDIFHPSLQQKNTLLPLYLYIQSWRKTY*NNFMALMLI  34776
              F QVSLED+LKKIQRVDFDIFHPSLQQKN LLPL LYIQSWRK Y*N+FM L+LI
Sbjct  1163   FLQVSLEDYLKKIQRVDFDIFHPSLQQKNMLLPLSLYIQSWRKRY*NDFMLLILI  999


 Score =  108 bits (230),  Expect(23) = 6e-175
 Identities = 42/46 (91%), Positives = 43/46 (93%), Gaps = 0/46 (0%)
 Frame = +2/-3

Query  12248  FKPCL*VKDSVSEKTIGLMRMQFWKKTVEDIYCDNPPHQPVAIELW  12385
              FK CL VKDSVSEKTIGLMRMQFWKK VED+YCDNPPHQPVAIELW
Sbjct  20067  FKLCLQVKDSVSEKTIGLMRMQFWKKAVEDMYCDNPPHQPVAIELW  19930


 Score = 97.3 bits (206),  Expect(23) = 6e-175
 Identities = 46/78 (58%), Positives = 52/78 (66%), Gaps = 0/78 (0%)
 Frame = +3/-2

Query  1794   GRRARSAGVMAASAHGSVWGPLRLGIPGLCCRRPPLGLYARMRRLPGPEVSGRSVAAASG  1973
              GRRA  AGVMA S  GSV GP   G+  L  R+PP   + R+RRLPGP    RSVAAASG
Sbjct  25930  GRRAPEAGVMATSMLGSVRGPRPFGLANLFHRQPPRDAWERVRRLPGPSAVRRSVAAASG  25751

Query  1974   PGAWGTDHYCLELLR*AS  2027
              PG  G+  YCLELLR* +
Sbjct  25750  PGIPGSHLYCLELLR*VA  25697


 Score = 95.5 bits (202),  Expect(23) = 6e-175
 Identities = 38/41 (92%), Positives = 40/41 (97%), Gaps = 0/41 (0%)
 Frame = +3/-2

Query  25263  TQHGVSQEDFLRRNQDKNVRDVIYDIASQAHLHLKHVSRLF  25385
              +QHGVSQEDFLRRNQDKNVRDV+YDIASQAHLHLKHVS LF
Sbjct  8845   SQHGVSQEDFLRRNQDKNVRDVVYDIASQAHLHLKHVSSLF  8723


 Score = 88.1 bits (186),  Expect(23) = 6e-175
 Identities = 38/47 (80%), Positives = 40/47 (85%), Gaps = 0/47 (0%)
 Frame = +3/-2

Query  8796   VCYRKRDYEGYLCSLLLPAESRSSVFALRAFNVELAQAGIKIP*NII  8936
              VC RKRDYE YLCSLL PAE + S  ALRAFNVELAQAGIKIP*NI+
Sbjct  22981  VCCRKRDYESYLCSLLFPAECQRSASALRAFNVELAQAGIKIP*NIM  22841
```

Veamos que hace exactamente BLASTX

4.9 "Translated query vs protein database (blastx)" is useful for finding similar proteins to those encoded by a nucleotide query.

Translated BLAST services are useful when trying to find homologous proteins to a nucleotide coding region. Blastx compares translational products of the nucleotide query sequence to a protein database. Because blastx translates the query sequence in all six reading frames and provides combined significance statistics for hits to different frames, it is particularly useful when the reading frame of the query sequence is unknown or it contains errors that may lead to frame shifts or other coding errors. Thus blastx is often the first analysis performed with a newly determined nucleotide sequence and is used extensively in analyzing EST sequences. This search is more sensitive than nucleotide blast since the comparison is performed at the protein level.

2.2.5 Responde lo siguiente ahora sobre (9) y (10): ¿que programa detecta más fragmentos comunes? ¿Qué programa te parece más potente para encontrar estos fragmentos?

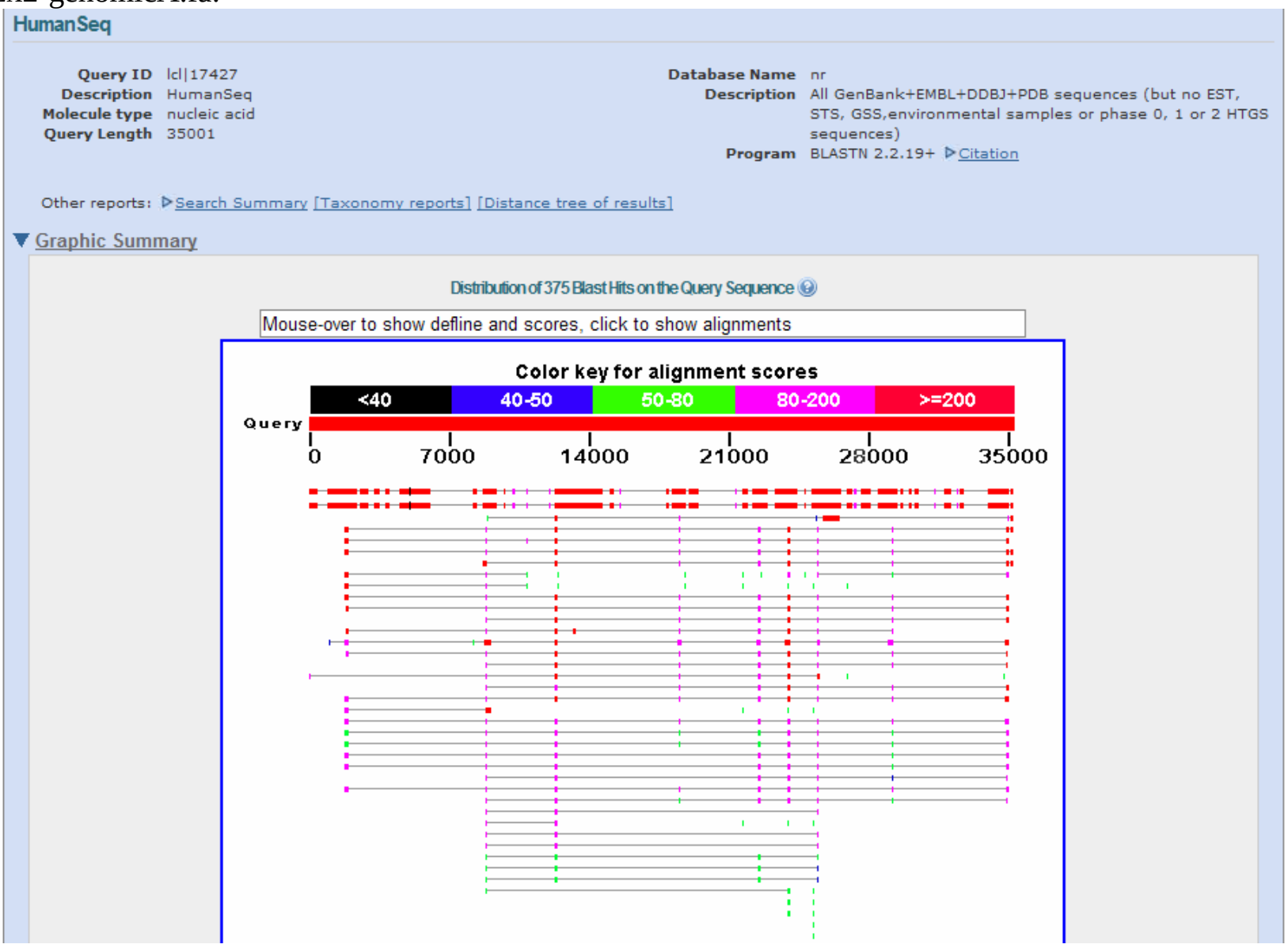
Observamos de las preguntas anteriores que:

- Blastn encuentra 12 fragmentos comunes.
- Blastx encuentra 104 fragmentos.

Reformulo>>>¿Por qué es más potente BLASTX que BLASTN?.

Primero comprobemos con BLAST que son estás secuencias:

Ex2-genomicA.fa:

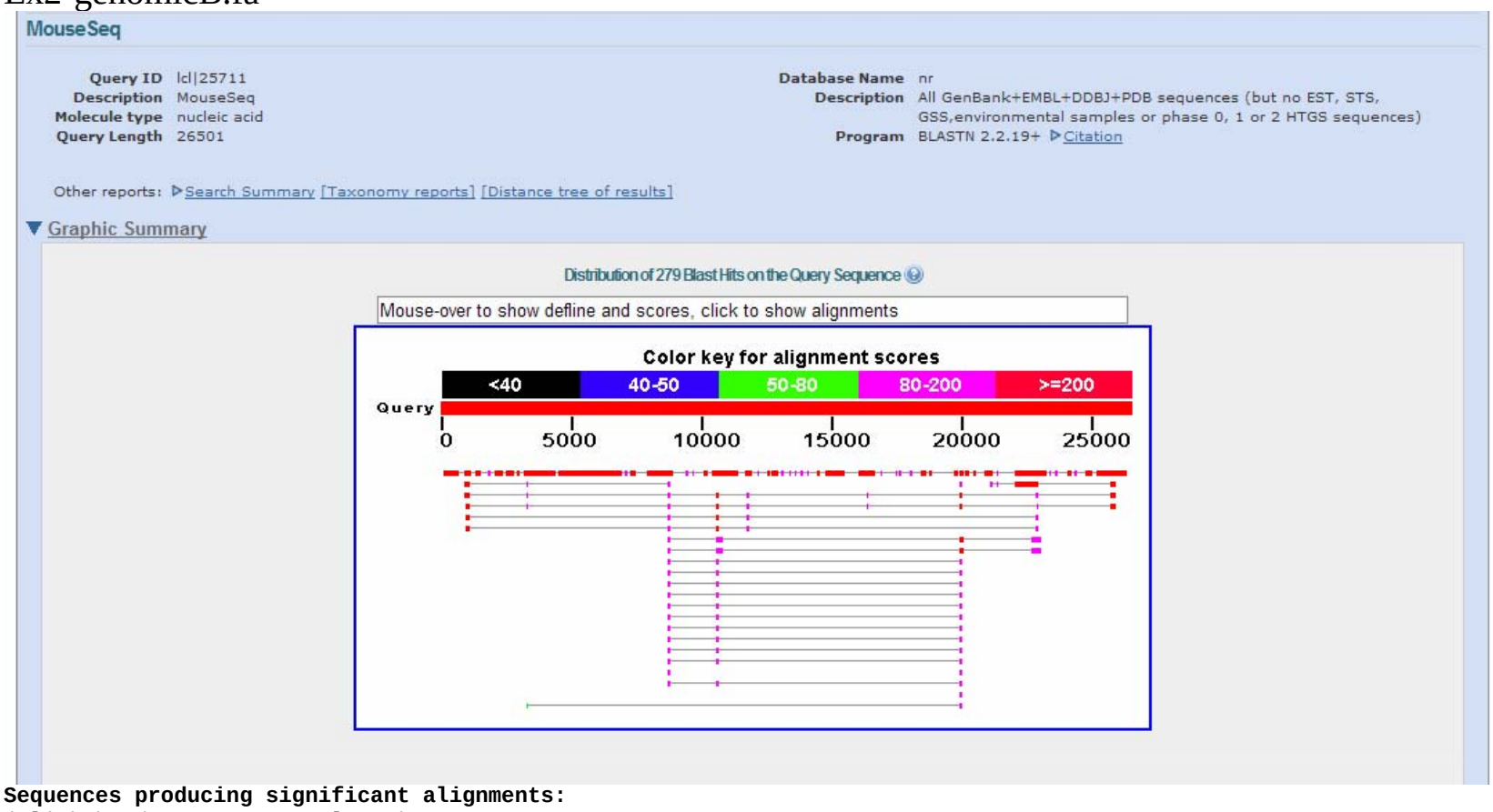


Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
gi|21105002|AC068189.11	Homo sapiens chromosome 8, clone RP11-320N21, complete sequence	4370	3.064e+04	49%	0.0	100%
gi|10716643|AC018801.4	Homo sapiens BAC clone RP11-279O19 from 8, complete sequence	4354	3.058e+04	49%	0.0	100%
gi|10435232|AK023339.1	Homo sapiens cDNA FLJ13277 fis, clone OVARC1001044	1624	2418	3%	0.0	100%
gi|124517690|NM_152416.2	Homo sapiens chromosome 8 open reading frame 38 (C8orf38), mRNA	387	2178	3%	2e-102	100%
gi|20380223|BC028166.1	Homo sapiens chromosome 8 open reading frame 38, mRNA (cDNA clone IMAGE:5245422), complete cds	385	2030	3%	9e-102	100%
gi|114620976|XM_001144499.1	PREDICTED: Pan troglodytes similar to putative phytoene synthase, transcript variant 1 (LOC464293), mRNA	378	2142	3%	1e-99	100%
gi|114620978|XM_519865.2	PREDICTED: Pan troglodytes similar to putative phytoene synthase, transcript variant 2 (LOC464293), mRNA	363	1946	3%	3e-95	100%
gi|123996308|DQ894830.2	Synthetic construct Homo sapiens clone IMAGE:100009290; FLH178894.01L; RZPDo839A06129D chromosome 8 open reading frame 38 (C8orf38) gene, encodes complete protein	360	627	1%	3e-94	100%

Podemos observar que la secuencia problema “Ex2-genomic.fa” parece ser el cromosoma 8 humano, al que se le han sustituido fragmentos, con “n” en la secuencia fasta.

Ex2-genomicB.fa



Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
gi|31096529|AL732497.15	Mouse DNA sequence from clone RP23-130H14 on chromosome 4 Contains a ribosomal protein S11 (Rps11) pseudogene, the Plekhf2 gene for pleckstrin homology domain containing family F (with FYVE domain) member 2, a novel gene and a CpG island, complete sequence	4364	1.841e+05	58%	0.0	100%
gi|74193144|AK041507.1	Mus musculus 3 days neonate thymus cDNA, RIKEN full-length enriched library, clone:A630018A01 product:hypothetical Terpenoid synthase containing protein, full insert sequence	1744	2405	4%	0.0	100%
gi|147906107|NM_001085493.1	Mus musculus RIKEN cDNA 2310030N02 gene (2310030N02Rik), mRNA	418	2046	4%	7e-112	100%
gi|12844408|AK009547.1	Mus musculus adult male tongue cDNA, RIKEN full-length enriched library, clone:2310030N02 product:unclassifiable, full insert sequence	379	1880	3%	3e-100	100%
gi|28913437|BC048539.1	Mus musculus RIKEN cDNA 2310030N02 gene, mRNA (cDNA clone IMAGE:6539873)	337	625	1%	2e-87	100%
gi|109476121|XM_001072347.1	PREDICTED: Rattus norvegicus similar to F23N19.9 (predicted) (RGD1309085_predicted), mRNA	235	932	2%	8e-57	99%
gi|62648625|XM_232684.3	PREDICTED: Rattus norvegicus similar to F23N19.9 (predicted) (RGD1309085_predicted), mRNA	235	932	2%	8e-57	99%
gi|21105002|AC068189.11	Homo sapiens chromosome 8, clone RP11-320N21, complete sequence	209	748	3%	5e-49	96%
gi|10716643|AC018801.4	Homo sapiens BAC clone RP11-279O19 from 8, complete sequence	209	748	3%	5e-49	96%
gi|194376371|AK301414.1	Homo sapiens cDNA FLJ57629 complete cds	187	516	1%	2e-42	96%
gi|194391377|AK298631.1	Homo sapiens cDNA FLJ57417 complete cds	185	514	1%	8e-42	96%
gi|124517690|NM_152416.2	Homo sapiens chromosome 8 open reading frame 38 (C8orf38), mRNA	185	514	1%	8e-42	96%
gi|114620976|XM_001144499.1	PREDICTED: Pan troglodytes similar to putative phytoene synthase, transcript variant 1 (LOC464293), mRNA	185	520	1%	8e-42	97%
gi|45594394|AY444560.1	Homo sapiens putative phytoene synthase mRNA, complete cds	185	514	1%	8e-42	96%
gi|114620978|XM_519865.2	PREDICTED: Pan troglodytes similar to putative phytoene synthase, transcript variant 2 (LOC464293), mRNA	182	516	1%	1e-40	97%

La secuencia problema “Ex2-genomicB.fa” es igualmente el cromosoma 4 del Ratón, con fragmentos sustituidos.

Como vemos estamos antes dos secuencias largas de nucleótidos de dos especies distintas, que pueden potencialmente pueden contener secuencias homologas. Todo parece indicar que los matchs de BLASTX nos están mostrando las zonas en donde puede existir un proteina homologa.

Como dice la ayuda de BLASTX, esta diseñado para:

> Translated BLAST services are useful when trying to find homologous proteins to a nucleotide coding region. Blastx compares translational products of the nucleotide query sequence to a protein database. Because blastx translates the query sequence in all six reading frames and provides combined significance statistics for hits to different frames, it is particularly useful when the reading frame of the query sequence is unknown or it contains errors that may lead to frame shifts or other coding errors. Thus blastx is often the first analysis performed with a newly determined nucleotide sequence and is used extensively in analyzing EST sequences. This search is more sensitive than nucleotide blast since the comparison is performed at the protein level.

En donde dice claramente que cuando hacemos este tipo de comparaciones muestra mucha más sensibilidad que la busqueda por nucleotidos (BLASTN). En definitiva BLASTX es más potente para comparar secuencias largas que pueden contener fragmentos homólogos.

Probemos la funcionalidad de BLASTX

Vemos que los querys de ambas secuencias hacen match común sobre gi|21105002|AC068189.11 (Un clon del cromosoma 8) en amarillo.

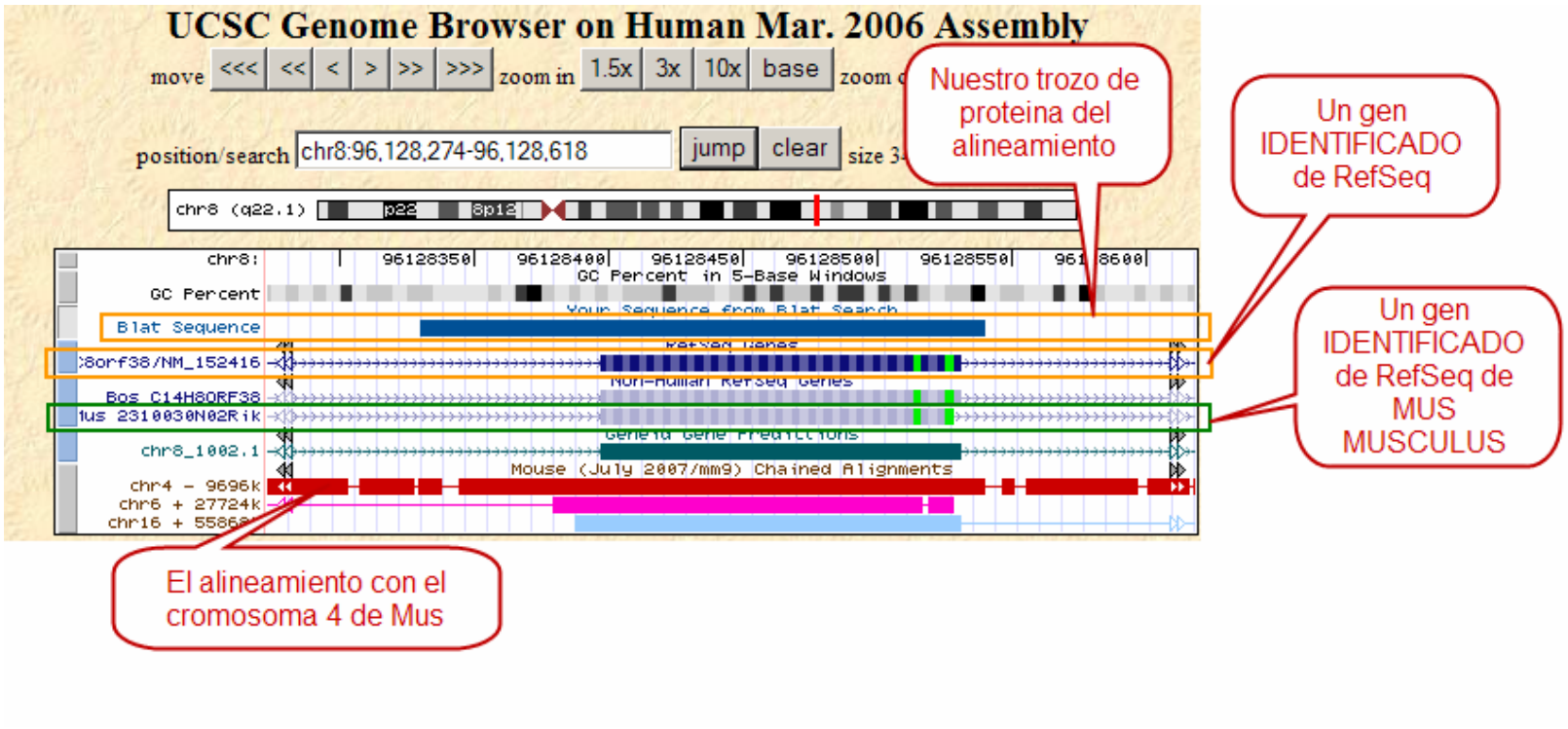
Profundicemos un poco en esto.

- Entramos en el Genome Browser y selecciono el cromosoma 8 (en la casilla de búsqueda escribir chr8)
- Añado con BLAT el primer fragmento coincidente de nuestras dos secuencias problema: (que habíamos obtenido anteriormente, en azul abajo))

```
Score =  131 bits (280),  Expect(23) = 6e-175
 Identities = 49/70 (70%), Positives = 61/70 (87%), Gaps = 0/70 (0%)
 Frame = +3/-1

Query  23739  MLFLLIYYALRIEQPCVISEIIGIKDLHADHAASHIGKAQGIVTCLRATPYHGSRRKVFL  23918
              ++ L+ Y +   EQPC++S ++G+KDLHADHAASHIGKAQGIVTCLRATPYH SRR+VFL
Sbjct  10787  LVLLICYASRTTEQPCLVSVVLGVKDLHADHAASHIGKAQGIVTCLRATPYHSSRRQVFL  10608

Query  23919  PMDICMLVRL  23948
              PMD+C+ VRL
Sbjct  10607  PMDVCVQVRL  10578
```

- Añadimos las pistas de other refseq y activo el Mouse Chain

Curiosamente obtenemos esto:



Si comprobamos quien es NM_152416 vemos que:

```
LOCUS       NM_152416               1808 bp    mRNA    linear   PRI 07-NOV-2008
DEFINITION  Homo sapiens chromosome 8 open reading frame 38 (C8orf38), mRNA.
ACCESSION   NM_152416
VERSION     NM_152416.2  GI:124517690
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1808)
  AUTHORS   Suzuki,Y., Yamashita,R., Shirota,M., Sakakibara,Y., Chiba,J.,
            Mizushima-Sugano,J., Nakai,K. and Sugano,S.
  TITLE     Sequence comparison of human and mouse genes reveals a homologous
            block structure in the promoter regions
  JOURNAL   Genome Res. 14 (9), 1711-1718 (2004)
   PUBMED   15342556
```

Parece que la ayuda de BLASTX esta en lo cierto, sirve para localizar homologos..

2.3 EJERCICIO 3: ANOTACION DE SECUENCIAS

2.3.1 Dadas las 3 secuencias Ex3-unknown1.fa, Ex3-unknown2.fa y Ex3-unknown3.fa adjuntadas al enunciado, probad todas las comparaciones dos a dos o tres a tres con el programa CLUSTALW, para decidir cual de las 3 secuencias NO parece estar relacionada con las otras dos restantes.

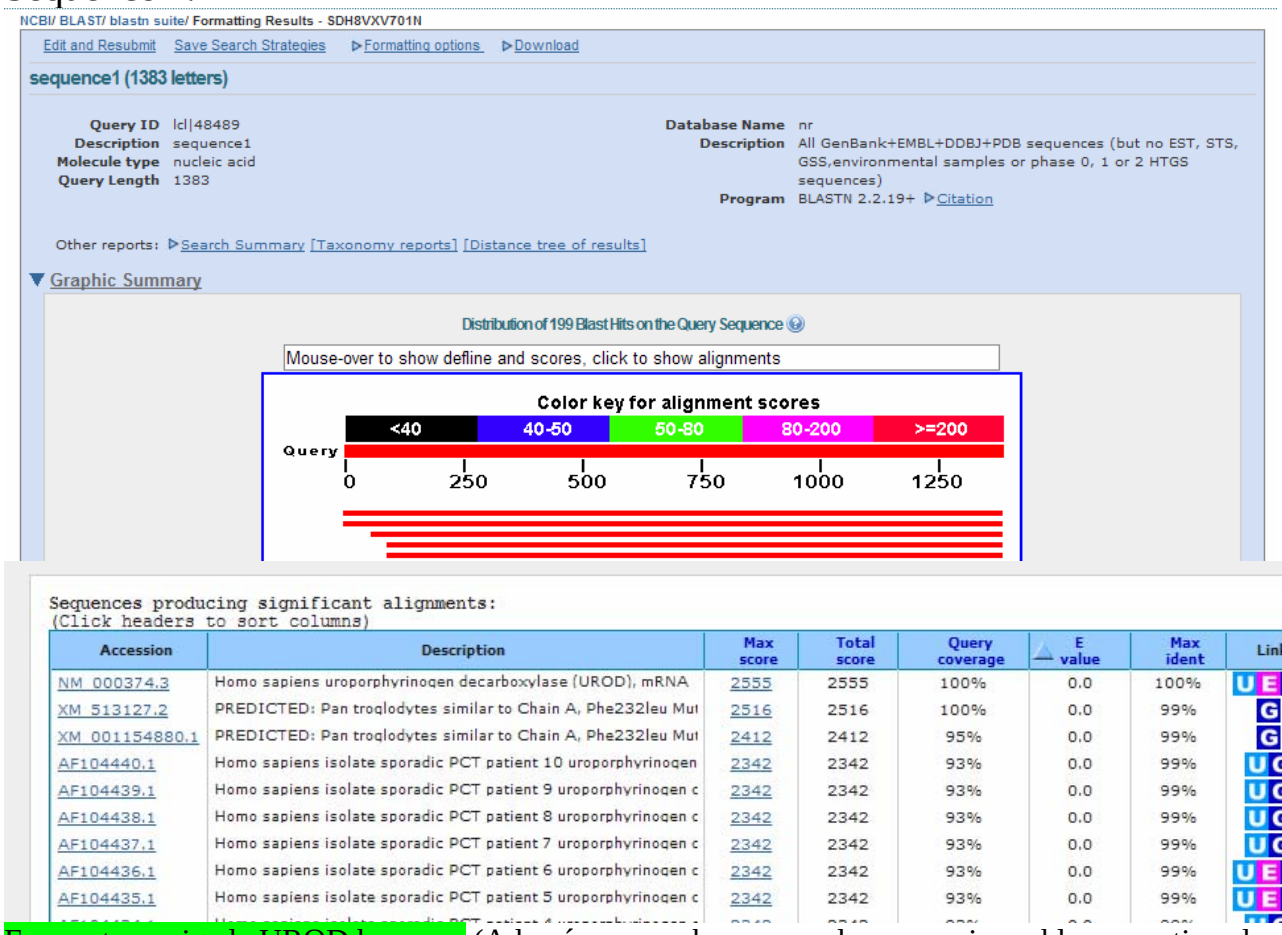
Calculad los alineamientos que consideréis oportunos y reflexionad la respuesta.

2.3.2 Usad el programa BLAST mas adecuado en la pagina Web http://www.ncbi.nlm.nih.gov/blast/ para averiguar que tipo de secuencia es Ex3-unknown1.fa. Esto representa realizar la anotación de la secuencia: debéis tener en cuenta los resultados más significativos.

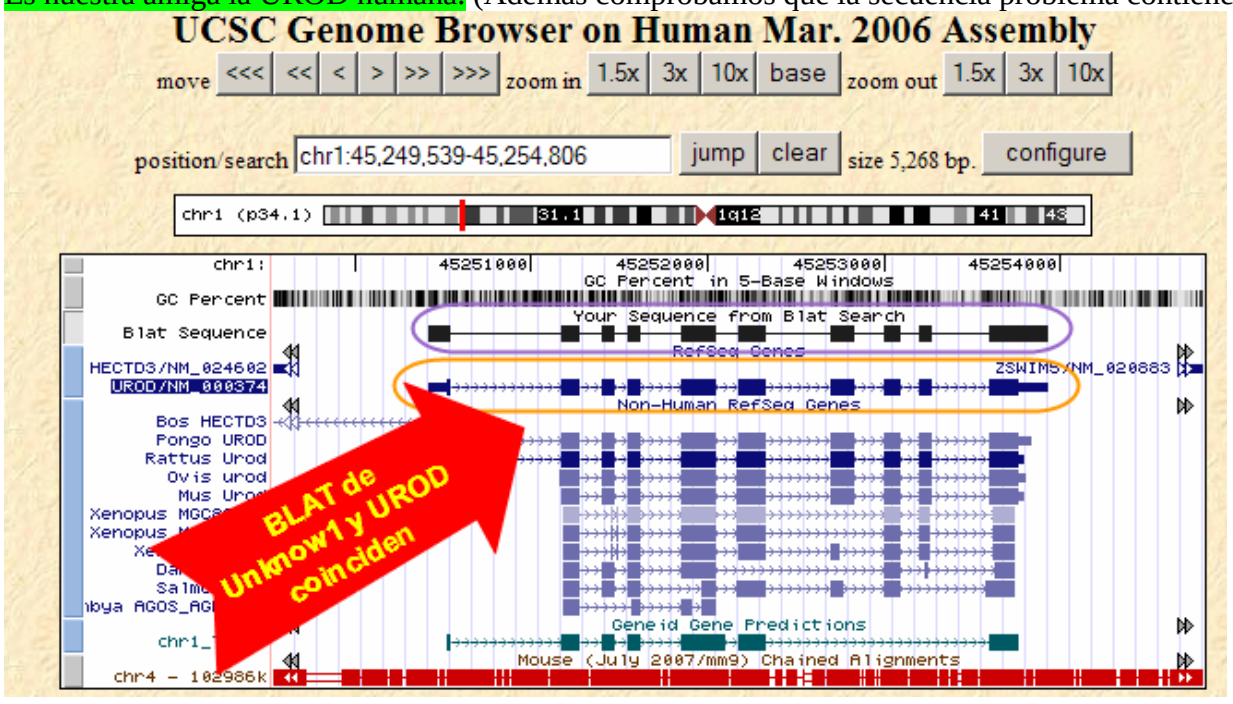
En este caso, no queremos alinear esta secuencia con otra de nuestras secuencias, sino con una base de datos de secuencias anotadas en el NCBI. Debéis escoger el programa y la base de datos que os parezcan mas apropiados según vuestra intuición y lo que habéis aprendido durante esta PEC.

Por la experiencia que he acumulado en la PEC primero creo que es mejor resolver el siguiente apartado y luego este. Como no cuesta mucho investigo primero que es cada una de las secuencias problema:

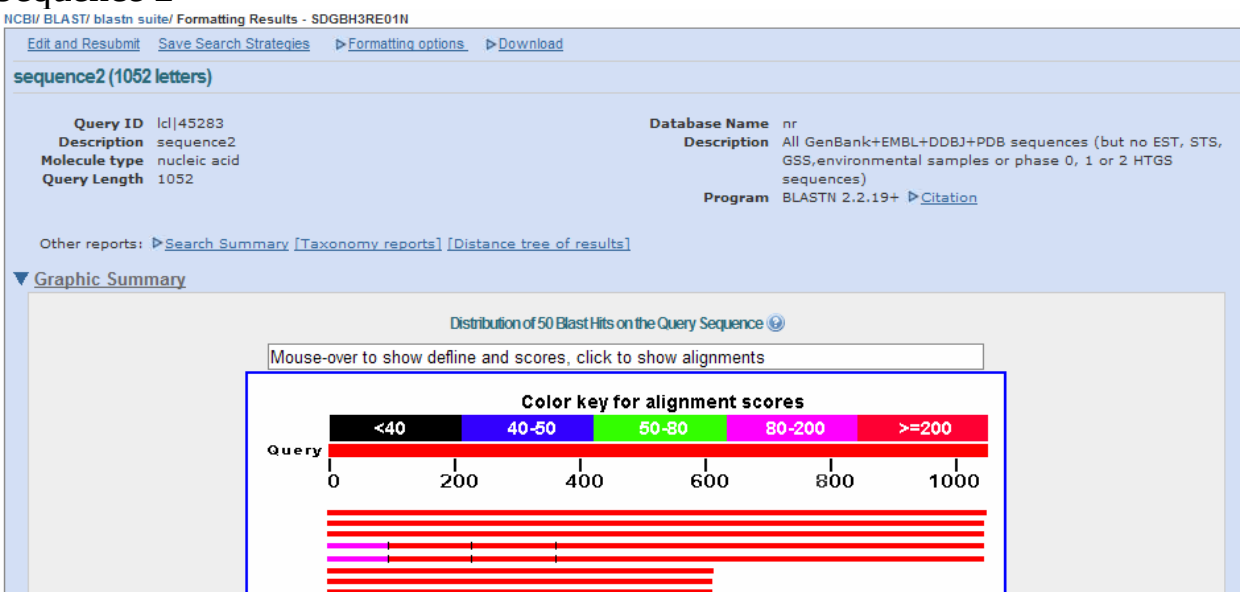
Sequence 1.



Es nuestra amiga la UROD humana. (Además comprobamos que la secuencia problema contiene los UTRs, intrones y exones)



Sequence 2

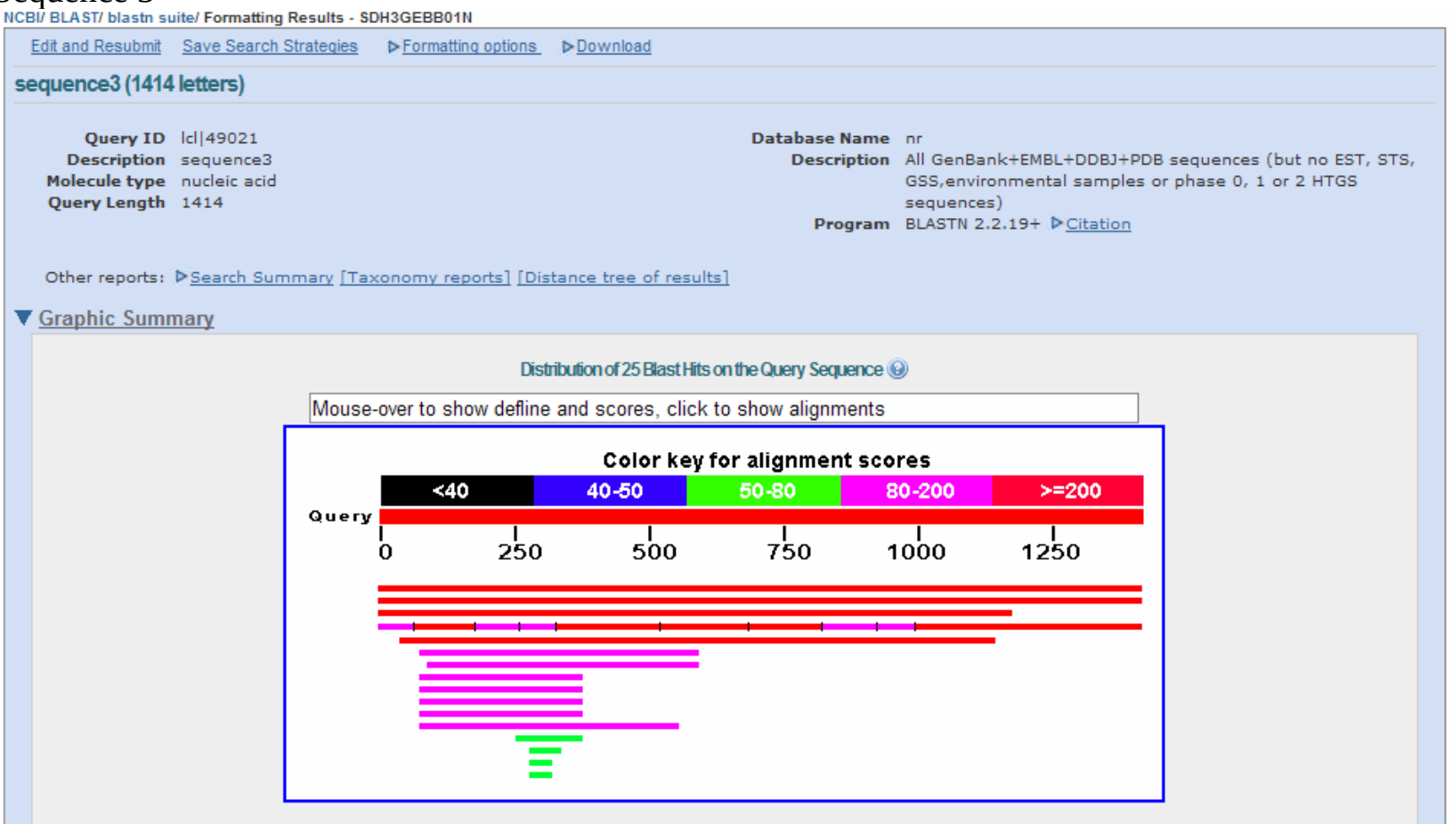


Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
D89076.1	Mus musculus mRNA for prealbumin, complete cds	1943	1943	100%	0.0	100%	U E G
NM_013697.4	Mus musculus transthyretin (Ttr), mRNA	1914	1914	99%	0.0	99%	U E G
AK014454.1	Mus musculus 18 days pregnant adult female placenta and extra	1866	1866	99%	0.0	98%	U E G
AC129078.6	Mus musculus BAC clone RP24-71C5 from chromosome 18, cor	1245	1928	99%	0.0	100%	
AC135290.5	Mus musculus BAC clone RP23-58D13 from 18, complete seque	1245	1928	99%	0.0	100%	
BC086926.1	Mus musculus transthyretin, mRNA (cDNA clone MGC:107649	1140	1140	58%	0.0	100%	U G
AK018701.1	Mus musculus adult male kidney cDNA, RIKEN full-length enrich	1134	1134	58%	0.0	100%	U E G
AK075588.1	Mus musculus adult male kidney cDNA, RIKEN full-length enrich	1134	1134	58%	0.0	100%	U E G
X03351.1	Mouse mRNA for prealbumin	1134	1134	58%	0.0	100%	U E G
M19524.1	Mouse prealbumin (transthyretin) gene, 5' flank	1129	1129	58%	0.0	99%	
AK050155.1	Mus musculus adult male liver tumor cDNA, RIKEN full-length er	1127	1127	57%	0.0	100%	U E G
BC032069.1	Mus musculus transthyretin, mRNA (cDNA clone IMAGE:41643	1122	1122	57%	0.0	100%	U E G
BC024702.1	Mus musculus transthyretin, mRNA (cDNA clone MGC:18651 IM	1090	1090	58%	0.0	98%	U E G
AF479660.1	Rattus norvegicus transthyretin-related protein (TTN) mRNA, co	837	837	58%	0.0	91%	U G
BC086946.1	Rattus norvegicus transthyretin, mRNA (cDNA clone MGC:1087	826	826	56%	0.0	91%	U G
AK080757.1	Mus musculus adult retina cDNA, RIKEN full-length enriched libr	819	819	43%	0.0	99%	U E G
X14876.1	Rat mRNA for transthyretin	817	817	56%	0.0	91%	U G

Es la prealbumina de raton.

Sequence 3



Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AF095639.1	Danio rerio uroporphyrinogen decarboxylase (urod) mRNA, compl	2612	2612	100%	0.0	100%	U E G
BC092696.1	Danio rerio uroporphyrinogen decarboxylase, mRNA (cDNA clone	2516	2516	100%	0.0	98%	U G
BC108075.1	Danio rerio uroporphyrinogen decarboxylase, mRNA (cDNA clone	2152	2152	83%	0.0	99%	U G
CU571330.9	Zebrafish DNA sequence from clone CH73-38P6 in linkage grou	785	2653	100%	0.0	100%	
BT045495.1	Salmo salar clone ssal-rqf-523-063 Uroporphyrinogen decarbox	562	562	78%	2e-156	76%	U G
XM_581108.4	PREDICTED: Bos taurus similar to uroporphyrinogen decarboxyl	180	180	36%	2e-41	74%	U G
NM_001012341.1	Ovis aries uroporphyrinogen decarboxylase (urod), mRNA	176	176	35%	3e-40	74%	U G
XM_532602.2	PREDICTED: Canis familiaris similar to Uroporphyrinogen decarl	156	156	21%	3e-34	76%	U E G
XM_856829.1	PREDICTED: Canis familiaris similar to Uroporphyrinogen decarl	156	156	21%	3e-34	76%	U E G
XM_856777.1	PREDICTED: Canis familiaris similar to Uroporphyrinogen decarl	156	156	21%	3e-34	76%	U E G

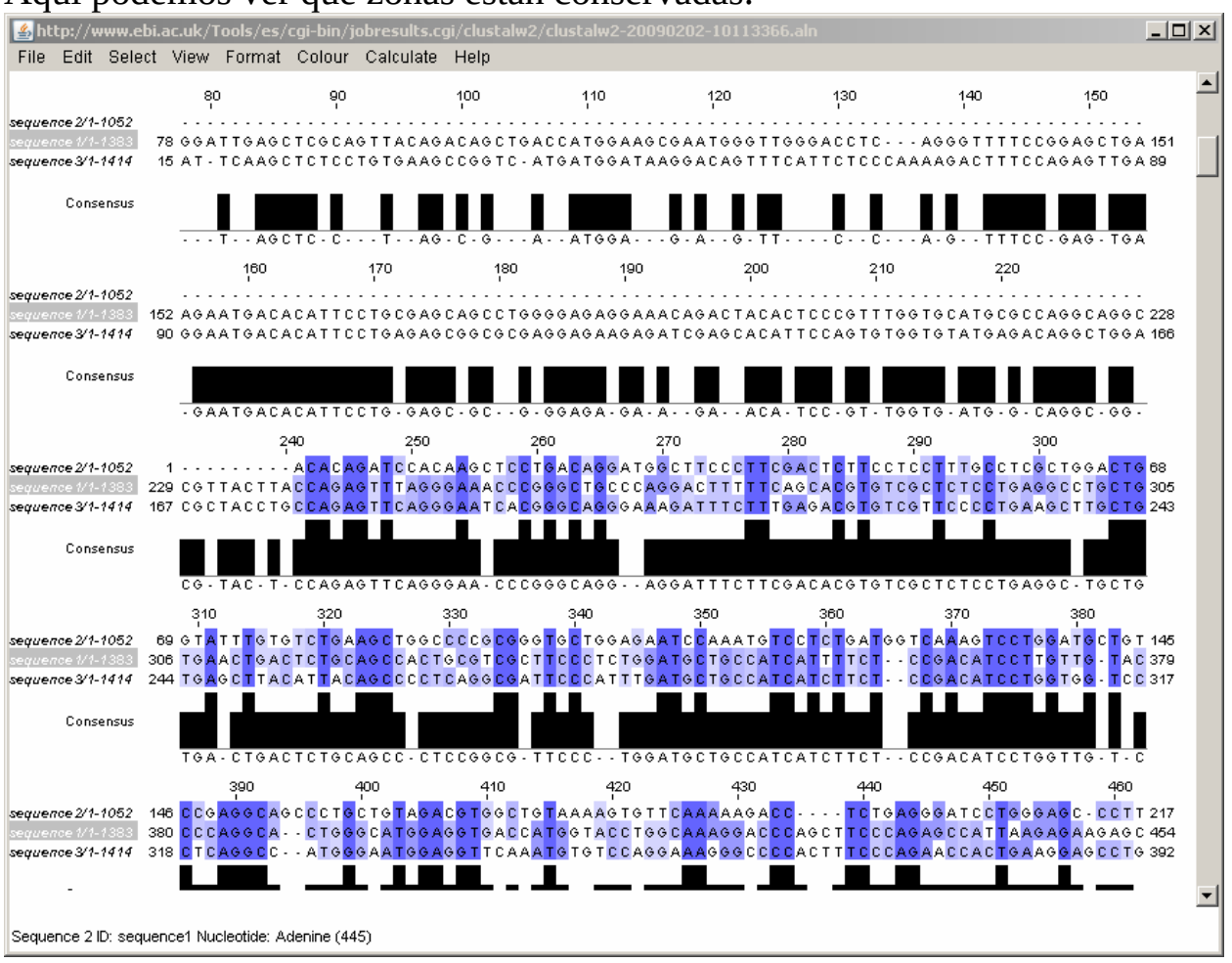
Es la UROD del pez Zebra.

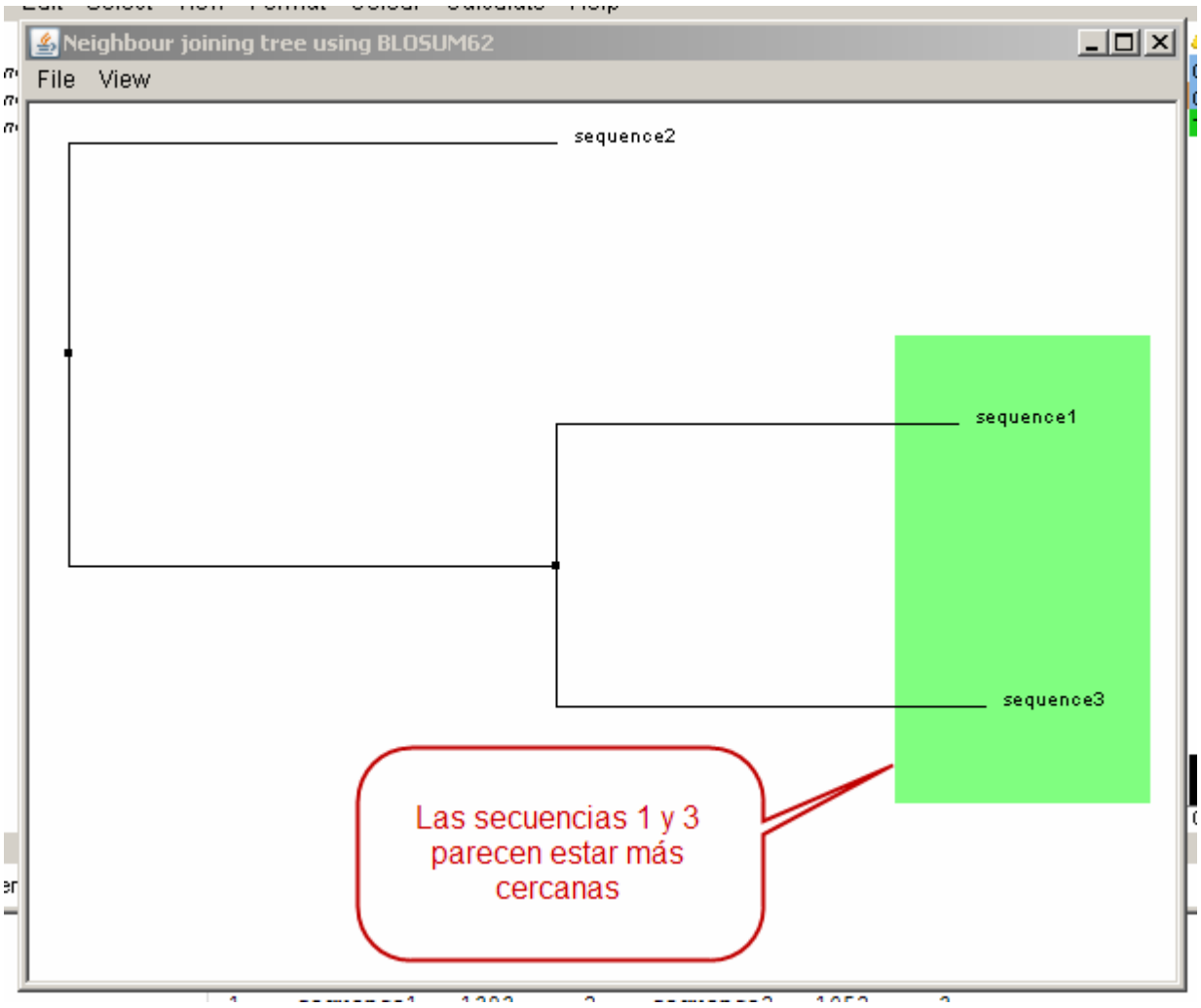
El Alineamiento con CLUSTALW de las tres secuencias, en la ordenación inicial ya muestra las diferencias:

```
SeqA Name        Len(nt)  SeqB Name        Len(nt)  Score
=========================================================
1    sequence1   1383     3    sequence3   1414     51
1    sequence1   1383     2    sequence2   1052     3
2    sequence2   1052     3    sequence3   1414     2
=========================================================
```

No obstante hay que hacer notar que el distinto tamaño de las secuencias puede desvirtuar el score. Estamos haciendo un alineamiento global y los gaps tb puntuan. Creo que lo adecuado seria hacer primero alineamientos locales y recortar las partes de las secuencias menos coincidentes, y luego repetir los alineamientos con los fragmentos más comunes. Igualmente hay que hacer una prueba estadística con cadenas del mismo tamaño y contenido CG, pero aleatorias. De esta manera comprobaríamos la significación de los scores.

Probando jalView
Aquí podemos ver que zonas están conservadas:





Podemos concluir que la sequence2 “Ex3-unknow2.fa” no está relacionada con la 1 y la 3.

Para testear los resultados:

Recupero un fasta de prealbumina de Humano (human_like2.fa)

```
LOCUS       NM_000371                938 bp    mRNA    linear   PRI 01-FEB-2009
DEFINITION  Homo sapiens transthyretin (TTR), mRNA.
ACCESSION   NM_000371
VERSION     NM_000371.3  GI:221136767
KEYWORDS    .
SOURCE      Homo sapiens (human)
```

Y volvemos a alinear

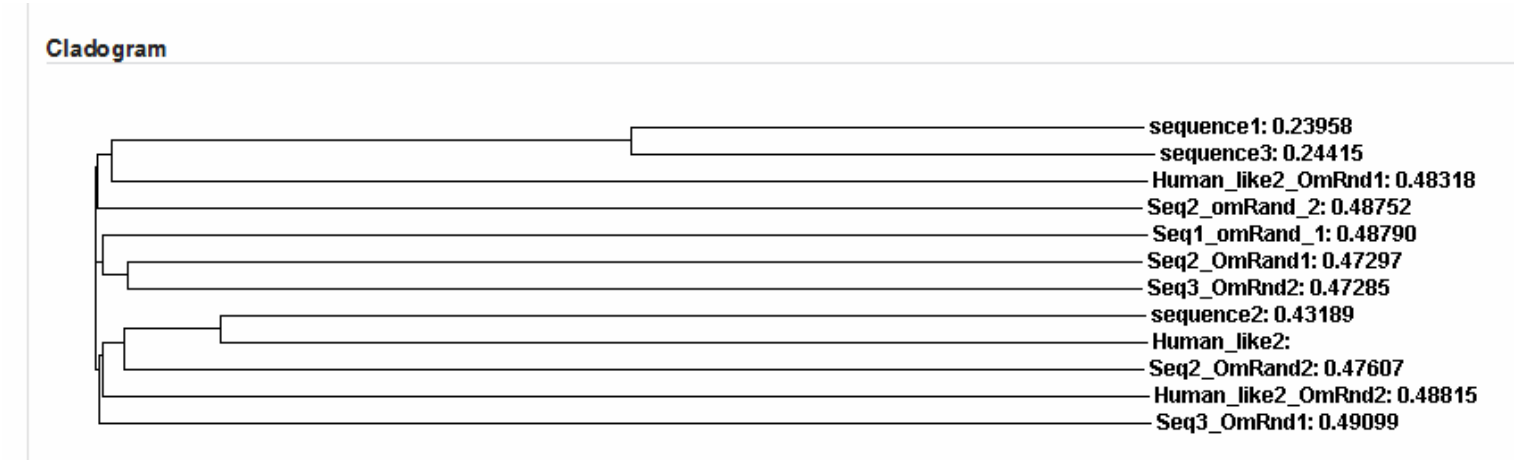
SeqA	Name	Len(nt)	SeqB	Name	Len(nt)	Score
1	sequence1	1383	2	sequence2	1052	3
1	sequence1	1383	3	sequence3	1414	51
1	sequence1	1383	4	Human_like2	8300	2
2	sequence2	1052	3	sequence3	1414	2
2	sequence2	1052	4	Human_like2	8300	13
3	sequence3	1414	4	Human_like2	8300	1

Vemos que obtiene un score más alto (13) (en verde) Puesto que Sequence2 y Human_like2 si están relacionadas. También comprobamos que el score de human_like2 con sequence3 y sequence1 es similar al que se obtenía con Sequence2. Confirmamos que no están relacionadas.

Para comprobar la sensibilidad preparo 2 secuencias aleatorias de cada una de las secuencias incluida la Human_like2, con el software OMKROGEN http://www.manet.uiuc.edu/nobai/nobai.php http://nar.oxfordjournals.org/cgi/screenpdf/36/suppl_2/W85.pdf

SeqA	Name	Len(nt)	SeqB	Name	Len(nt)	Score
1	sequence1	1383	7	sequence3	1414	51
4	sequence2	1052	10	Human_like2	8300	13
6	Seq2_OmRand2	1052	10	Human_like2	8300	7
5	Seq2_OmRand1	1052	9	Seq3_OmRnd2	1414	5
1	sequence1	1383	11	Human_like2_OmRnd1	8300	4
4	sequence2	1052	11	Human_like2_OmRnd1	8300	4
1	sequence1	1383	4	sequence2	1052	3

```
2     Seq1_omRand_1         1383      9     Seq3_OmRnd2           1414      3
5     Seq2_OmRand1          1052      10    Human_like2           8300      3
5     Seq2_OmRand1          1052      12    Human_like2_OmRnd2    8300      3
6     Seq2_OmRand2          1052      12    Human_like2_OmRnd2    8300      3
9     Seq3_OmRnd2           1414      11    Human_like2_OmRnd1    8300      3
1     sequence1             1383      3     Seq2_omRand_2         1383      2
1     sequence1             1383      6     Seq2_OmRand2          1052      2
1     sequence1             1383      9     Seq3_OmRnd2           1414      2
1     sequence1             1383      10    Human_like2           8300      2
1     sequence1             1383      12    Human_like2_OmRnd2    8300      2
2     Seq1_omRand_1         1383      3     Seq2_omRand_2         1383      2
2     Seq1_omRand_1         1383      4     sequence2             1052      2
2     Seq1_omRand_1         1383      6     Seq2_OmRand2          1052      2
3     Seq2_omRand_2         1383      8     Seq3_OmRnd1           1414      2
3     Seq2_omRand_2         1383      9     Seq3_OmRnd2           1414      2
3     Seq2_omRand_2         1383      10    Human_like2           8300      2
3     Seq2_omRand_2         1383      11    Human_like2_OmRnd1    8300      2
3     Seq2_omRand_2         1383      12    Human_like2_OmRnd2    8300      2
4     sequence2             1052      6     Seq2_OmRand2          1052      2
4     sequence2             1052      7     sequence3             1414      2
4     sequence2             1052      12    Human_like2_OmRnd2    8300      2
5     Seq2_OmRand1          1052      7     sequence3             1414      2
5     Seq2_OmRand1          1052      11    Human_like2_OmRnd1    8300      2
6     Seq2_OmRand2          1052      11    Human_like2_OmRnd1    8300      2
7     sequence3             1414      11    Human_like2_OmRnd1    8300      2
7     sequence3             1414      12    Human_like2_OmRnd2    8300      2
8     Seq3_OmRnd1           1414      10    Human_like2           8300      2
8     Seq3_OmRnd1           1414      12    Human_like2_OmRnd2    8300      2
1     sequence1             1383      2     Seq1_omRand_1         1383      1
1     sequence1             1383      5     Seq2_OmRand1          1052      1
1     sequence1             1383      8     Seq3_OmRnd1           1414      1
2     Seq1_omRand_1         1383      5     Seq2_OmRand1          1052      1
2     Seq1_omRand_1         1383      7     sequence3             1414      1
2     Seq1_omRand_1         1383      8     Seq3_OmRnd1           1414      1
2     Seq1_omRand_1         1383      10    Human_like2           8300      1
2     Seq1_omRand_1         1383      11    Human_like2_OmRnd1    8300      1
2     Seq1_omRand_1         1383      12    Human_like2_OmRnd2    8300      1
3     Seq2_omRand_2         1383      4     sequence2             1052      1
3     Seq2_omRand_2         1383      5     Seq2_OmRand1          1052      1
3     Seq2_omRand_2         1383      6     Seq2_OmRand2          1052      1
3     Seq2_omRand_2         1383      7     sequence3             1414      1
4     sequence2             1052      5     Seq2_OmRand1          1052      1
4     sequence2             1052      8     Seq3_OmRnd1           1414      1
4     sequence2             1052      9     Seq3_OmRnd2           1414      1
5     Seq2_OmRand1          1052      6     Seq2_OmRand2          1052      1
5     Seq2_OmRand1          1052      8     Seq3_OmRnd1           1414      1
6     Seq2_OmRand2          1052      7     sequence3             1414      1
6     Seq2_OmRand2          1052      8     Seq3_OmRnd1           1414      1
6     Seq2_OmRand2          1052      9     Seq3_OmRnd2           1414      1
7     sequence3             1414      8     Seq3_OmRnd1           1414      1
7     sequence3             1414      9     Seq3_OmRnd2           1414      1
7     sequence3             1414      10    Human_like2           8300      1
8     Seq3_OmRnd1           1414      9     Seq3_OmRnd2           1414      1
8     Seq3_OmRnd1           1414      11    Human_like2_OmRnd1    8300      1
9     Seq3_OmRnd2           1414      10    Human_like2           8300      1
9     Seq3_OmRnd2           1414      12    Human_like2_OmRnd2    8300      1
10    Human_like2           8300      11    Human_like2_OmRnd1    8300      0
10    Human_like2           8300      12    Human_like2_OmRnd2    8300      0
11    Human_like2_OmRnd1    8300      12    Human_like2_OmRnd2    8300      0
================================================================================
```



Como conclusión final podemos decir que todo secuencia (en esta comparación) con score por debajo o igual a 7, tiene muchas probabilidades de no estas relacionada.