



Laboratórios de Bioinformática

Sessão 1



Porquê BDs biológicas ?



- **Tornar os dados acessíveis aos investigadores**

- Integração de dados de fontes diversas
- Fornecer acesso a conjuntos de dados demasiado grandes para serem explicitamente publicados (e.g. genomas, dados experimentais, ...)

- **Disponibilizar dados em formatos para processamento automático**

- Disponibilizar dados em grande escala em formatos não ideais para leitura humana mas sim para processamento por programas



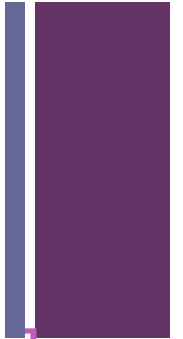
Bases de dados de sequências - nota histórica



- Primeira base de dados nos anos 1960/70: **PIR** (Dayhoff) – sequências de proteínas
- 1ª BD de DNA – **EMBL**, 1982; logo seguida pelo GenBank do NCBI
- Em 1986 surge a 1ª versão da **Swiss-Prot** (base de dados curada)
- 1988 – **EBI** (Europa), **NCBI** (EUA) e **DDBJ** (Japão) criam o INSDC – International Nucleotide Sequence Database Collaboration
 - O INSDC permite partilha das sequências depositadas nas 3 BDs
- Em 2003 **EBI**, **PIR** e **Swiss-Prot** juntam-se na UniProt (dados curados e não curados)
- Em 2008 **ENA** substitui EMBL-Bank integrando outros tipos de dados (e.g. next generation sequencing)



Bases de dados biológicas



■ Sequências de DNA, RNA

- GenBank (NCBI) <http://www.ncbi.nlm.nih.gov/Genbank>
- EMBLBank (EBI) <http://www.ebi.ac.uk/embl/> (ENA)
- DDBJ (Japan) [http:// www.ddbj.nig.ac.jp](http://www.ddbj.nig.ac.jp)

■ Sequências de proteínas

- UniProt <http://www.ebi.uniprot.org>

■ Estruturas de proteínas

- PDB <http:// www.rcsb.org/pdb>

■ Domínios de proteínas

- CDD <http://www.ncbi.nlm.nih.gov/cdd>

■ Metabolismo – reações, vias metabólicas (e.g. KEGG)



Bases de dados biológicas

- **Genomas** de diversas espécies (Ensembl)
- **Dados expressão genética** (NCBI GEO, ArrayExpress)
- **Bibliografia** (PubMed)
- **Taxonomia** (NCBI Taxonomy, Tree of Life)
- **Ontologias** (terminologia) – Gene Ontology, MESH
- **Mutações / doenças genéticas** (e.g. SNPs, OMIM)
- **Metabolitos** e dados de **metabolómica**: ChEBI, PubChem, HMDB, Metabolites, Metabolomics Workbench

(...)



+ Interfaces: pesquisa integrada EBI

<http://www.ebi.ac.uk/>

ebi.ac.uk

EMBL-EBI

ServicesResearchTrainingAbout us

EMBL-EBI

ServicesResearchTrainingAbout us

EBI Search

brca1

Search

Examples: VAV_HUMAN, tpi1, Sulston ...

Build Query

Help & Documentation

About EBI Search

Share

Feedback

Search results for *brca1*

Showing 21 results out of 53,363 in All results

Filter your results

Source

- All results (53,363)
- Genomes & metagenomes (5,171)
- Nucleotide sequences (13,012)
- Protein sequences (14,100)
- Macromolecular structures (79)
- Small molecules (45)
- Gene expression (4,701)
- Molecular interactions (400)
- Reactions, pathways & diseases (745)
- Protein families (59)
- Protein expression data (11)
- Enzymes (124)
- Literature (11,156)
- Samples & ontologies (3,742)
- EBI web (18)

Gene & protein summaries (includes expression, structures, literature...) (3 results found)

[BRCA1, DNA repair associated](#)
BRCA1 (PSCP, RNF53, PPP1R53, BRCC1, PNCA4, BRCAI, FANCS, BROVCA1, IRIS, ENSG00000012048)
Human (Homo sapiens)

[Breast cancer 1, early onset](#)
Brca1 (ENSMUSG00000017146)
House Mouse (Mus musculus)

[Protein BREAST CANCER SUSCEPTIBILITY 1 homolog](#)
BRCA1 (ATBRCA1, ARABIDOPSIS THALIANA BREAST CANCER SUSCEPTIBILITY1, **BRCA1**, breast cancer susceptibility1, AT4G21070)
Thale Cress (Arabidopsis thaliana)

Enzymes (124 results found)

[FANCI_MOUSE](#)
Fanconi anemia group J protein homolog
Fanconi anemia group J protein homolog

Related data

Views

Source: Enzyme Portal
ID: FANCI_MOUSE

Structural Bioinformatics 2016

This course will explore bioinformatics data resources and

Ensembl Bite sized – September 2016

Work with the Ensembl team to get to grips with the Ensembl browser

ArrayExpress: why and how to submit your data

Join Melissa Burke, a curator with ArrayExpress, for a webinar on why




Data submission

Research

Training

News


+ Interfaces: pesquisa integrada NCBI

 [Resources](#)  [How To](#) 

Sign in to NCBI

Search NCBI databases

Help



Results found in 35 databases for "BRCA1"

Literature

Books	922	books and reports
MeSH	27	ontology used for PubMed indexing
NLM Catalog	56	books, journals and more in the NLM Collections
PubMed	12,882	scientific & medical abstracts/citations
PubMed Central	26,318	full-text journal articles

Health

ClinVar	6,673	human variations of clinical significance
dbGaP	72	genotype/phenotype interaction studies
GTR	402	genetic testing registry
MedGen	27	medical genetics literature and links
OMIM	173	online mendelian inheritance in man
PubMed Health	139	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	183	biological projects providing data to NCBI
BioSample	484	descriptions of biological source materials
Clone	921	genomic and cDNA clones
dbVar	1,642	genome structural variation studies
Genome	15	genome sequencing projects by organism

Genes

EST	381	expressed sequence tag sequences
Gene	13,524	collected information about gene loci
GEO DataSets	3,264	functional genomics studies
GEO Profiles	228,738	gene expression and molecular abundance profiles
HomoloGene	65	homologous gene sets for selected organisms
PopSet	114	sequence sets from phylogenetic and population studies
UniGene	320	clusters of expressed transcripts

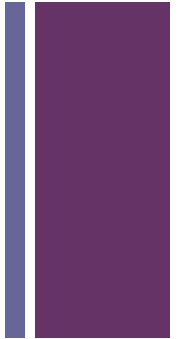
Proteins

Conserved Domains	59	conserved protein domains
Protein	113,444	protein sequences
Protein Clusters	19	sequence similarity-based protein clusters
Structure	290	experimentally-determined biomolecular structures

Chemicals

BioSystems	2,286	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	251	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	669	deposited substance and chemical information

+ Classificação - BDs de sequências



■ Primárias

- Contêm dados de sequenciação da responsabilidade dos seus autores; dados não são tratados nem curados
- Existe redundância
- Exemplos:
 - ENA (European Nucleotide Archive)
www.ebi.ac.uk/ena – (inclui EMBL-bank)
 - GenBank: www.ncbi.nlm.nih.gov/GenBank
 - DDBJ: www.ddbj.nig.ac.jp



Potenciais problemas das BDs primárias



- Se uma “*feature*” contém informação errónea (e.g. sobre tradução) esta irá ser propagada as outras BDs que extraem a sua informação das BDs primárias
- Se a informação sobre a proteína não está no sítio correto no registo, os programas de extração de informação irão falhar.



Classificação das BDs de sequências



■ Secundárias

- BDs com dados curados por especialistas; implicam trabalho de validação dos dados
- Exemplos:
 - NCBI Gene - <http://www.ncbi.nlm.nih.gov/gene>
 - base de dados curada com informação centrada nos genes
 - NCBI Protein - <http://www.ncbi.nlm.nih.gov/protein>
 - tradução das sequências do GenBank, RefSeq e TPA. Inclui registos da SwissProt, PIR, PRF e PDB



Classificação das BDs de sequências



- **NCBI RefSeq** - <http://www.ncbi.nlm.nih.gov/refseq>
 - (curada partir do GenBank; evita redundância, i.e. sequências repetidas; inclui DNA, RNA e proteínas; liga explicitamente sequências de DNA e proteínas)
- **UniProt** - <http://www.uniprot.org/>
 - Repositório de informação sobre proteínas: sequências e anotação
 - Resulta da união das BD's Swiss Prot, TrEMBL e PIR-PSD
 - 3 componentes: UniParc, UniProtKB (contém a SwissProt e a TrEMBL), UniRef



NCBI

National Center for Biotechnology Information



Redundância de dados

- É importante ter em conta, quando se realizam pesquisas nas bases de dados NCBI, que algumas destas sequências possam ser redundantes.
- A redundância das sequências deve-se ao facto de o mesmo gene (ou genoma) ter sido sequenciado por diferentes laboratórios que mais tarde submetem estas sequências.
- Exemplos de situações que isto pode acontecer: num surto de uma doença causado por uma bactéria ou vírus, diferentes laboratórios podem sequenciar o genoma desta espécie para seu estudo. Mais tarde submetem estas mesmos genomas que podem apresentar diferenças devido à qualidade da sequenciação ou presença de mutações.
- Assim, diferentes bases de dados do NCBI podem conter informação redundante para um dado gene, com sequências de qualidade variável, contendo dados curados ou não curados.



Redundância de dados



Exemplos de bases de dados **não curadas**:

- **GenBank/GenPept** - contém sequências não revistas submetidas (eventualmente com pouca qualidade) por laboratórios individuais e por projetos de sequenciação em larga escala.

Exemplos de bases de dados **curadas**:

- **Refseq** - As sequências presentes no NCBI Genbank que passam por uma curação manual passam para o Refseq. Esta base de dados é abrangente, integrada, não-redundante e com uma anotação completa, incluindo a sequência genômica, transcritos e proteínas.
- **SwissProt (UniProt)** - sequências de proteínas curadas e revistas manualmente.



Pesquisa por textos biomédicos:

PubMed <http://www.ncbi.nlm.nih.gov/pubmed>

brca1 brca2 - PubMed - X +

ncbi.nlm.nih.gov/pubmed/?term=brca1+brca2

NCBI Resources How To Sign in to NCBI

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

PubMed Search

Create RSS Create alert Advanced Help

NCBI will be testing https on public web servers from 8:00 AM to 12:00 PM EDT (12:00-16:00 UTC) on Monday, September 26. You may experience problems with NCBI web sites during that time. Please plan accordingly. [Read more.](#)

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

PubMed
Commons
Reader comments
Trending articles

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

[Clear all](#)

[Show additional filters](#)

Format: Summary - Sort by: Most Recent -

Send to - Filters: [Manage Filters](#)

Search results

Items: 1 to 20 of 6525

<< First < Prev Page 1 of 327 Next > Last >>

☐ [BRCA mutations and survival in breast cancer: an updated systematic review and meta-analysis.](#)

1. Zhu Y, Wu J, Zhang C, Sun S, Zhang J, Liu W, Huang J, Zhang Z. *Oncotarget*. 2016 Sep 21. doi: 10.18632/oncotarget.12158. [Epub ahead of print]
PMID: 27659521
[Similar articles](#)

☐ [Clinically Significant Unclassified Variants in BRCA1 and BRCA2 Genes Among Korean Breast Cancer Patients.](#)

2. Yoon KA, Park B, Lee BI, Yang MJ, Kong SY, Lee ES. *Cancer Res Treat*. 2016 Sep 13. doi: 10.4143/crt.2016.292. [Epub ahead of print]
PMID: 27658390
[Similar articles](#)

☐ [Multiplex Gene Expression Profiling of 16 Target Genes in Neoplastic and Non-Neoplastic Canine Mammary Tissues Using Branched-DNA Assay.](#)

3. Lüder Ripoli F, Conradine Hammer S, Mohr A, Willenbrock S, Hewicker-Trautwein M, Brenig B, Murua Escobar H, Nolte I. *Int J Mol Sci*. 2016 Sep 21;17(9). pii: E1589.
PMID: 27657059
[Similar articles](#)

☐ [Uncovering synthetic lethal interactions for therapeutic targets and predictive markers in lung adenocarcinoma.](#)

4. Chang JG, Chen CC, Wu YY, Che TF, Huang YS, Yeh KT, Shieh GS, Yang PC. *Oncotarget*. 2016 Sep 15. doi: 10.18632/oncotarget.12046. [Epub ahead of print]
PMID: 27655641
[Similar articles](#)

☐ [Risk reduction and survival benefit of prophylactic surgery in BRCA mutation carriers, a systematic review.](#)

5. Ludwig KK, Neuner J, Butler A, Geurts JL, Kong AL. *Am J Surg*. 2016 Jul 18. pii: S0002-9610(16)30348-8. doi: 10.1016/j.amjsurg.2016.06.010. [Epub ahead of print]
PMID: 27649974
[Similar articles](#)

Results by year

Download CSV

Related searches

breast cancer brca1 brca2
brca1 brca2 mutations
meta-analysis of brca1 and brca2 penetrance
brca1 brca2 review
brca1 brca2 gene

PMC Images search for brca1 brca2

See more (1100)...

Titles with your search terms

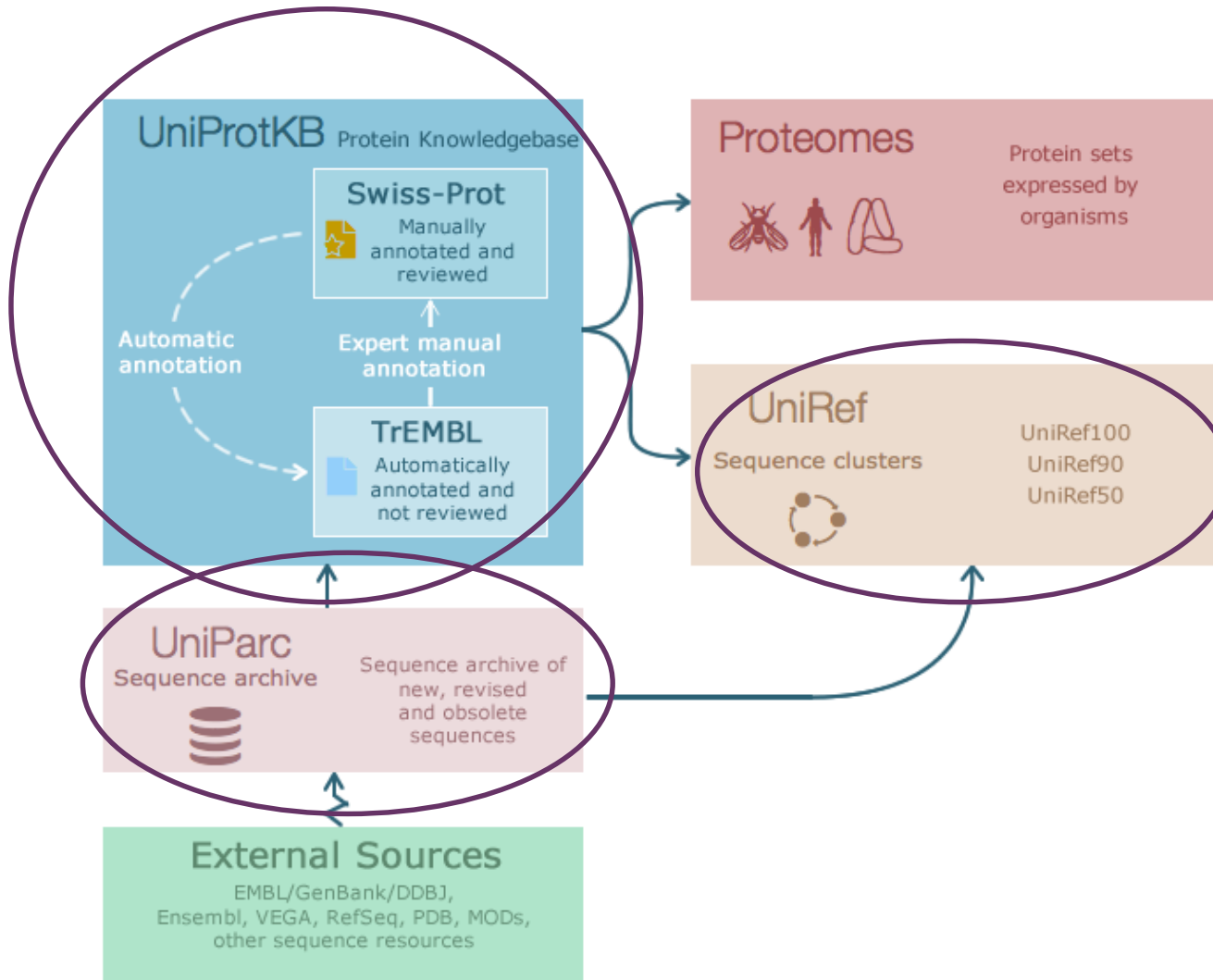
Bilateral Oophorectomy and Breast Cancer Risk in BRCA1 and BRCA2 [*J Natl Cancer Inst*. 2017]
Effect of decision aid for breast cancer



UniProt

Universal Protein Resource

+ UniProt





UniProt: componentes



■ UniParc (UniProt archive)

- Base de dados primária; inclui traduções automáticas dos registos do EMBL-Bank/ ENA e GenBank;
 - inclui submissões de sequências de proteínas das bd's SwissProt, TrEMBL, PIR e outras fontes
- A maior fonte de sequências de proteínas não redundantes atualmente existente (cada sequência só é guardada uma vez)
- Usada por muitas ferramentas de procura por homologia (e.g. BLAST)



UniProt: componentes



■ UniProtKB (UniProt Knowledgebase)

Repositório central de informação funcional sobre proteínas contendo anotação rica, precisa e consistente

Junta todas as sequências referidas ao mesmo gene e reúne toda a informação conhecida sobre a proteína

■ Swiss Prot

- Registos já curados

■ TrEMBL

- Anotações automáticas aguardando curaço manual



UniProt: componentes



■ UniRef (UniProt Reference Clusters)

- Clusters de sequências da UniProKB e de alguns registos da UniParc
- Elimina redundância das sequências disponíveis na UniProt e agrupa as sequências em 3 níveis:
 - UniRef100: sequências idênticas com 11 ou mais resíduos
 - UniRef90: grupos de sequências da UniRef100 com 90% de identidade e 80% de overlap
 - UniRef50: grupos de sequências da UniRef90 com 50% de identidade e 80% de overlap



Explorando um registo UniProt

The screenshot shows the UniProtKB entry for P00964 (GLNA_NOSS1). The browser address bar shows the URL `uniprot.org/uniprot/P00964?sort=score`. The UniProt logo is in the top left, and a search bar is in the top right. The main header displays the entry ID and name: **UniProtKB - P00964 (GLNA_NOSS1)**. Below this, there are tabs for **Display**, **Entry**, **Publications**, **Feature viewer**, and **Feature table**. The **Display** tab is selected, showing a list of categories on the left: **Function**, **Names & Taxonomy**, **Subcellular location**, **Pathology & Biotech**, **PTM / Processing**, **Expression**, **Interaction**, **Structure**, **Family & Domains**, **Sequence**, **Cross-references**, **Entry information**, **Miscellaneous**, and **Similar proteins**. The **Function** section is expanded, showing the following information:

- Protein**: Glutamine synthetase
- Gene**: glnA
- Organism**: *Nostoc sp. (strain PCC 7120 / SAG 25.82 / UTEX 2576)*
- Status**: Reviewed - Annotation score: 100% - Experimental evidence at protein level¹

The **Function** section is further detailed with the following information:

- Catalytic activity**¹: ATP + L-glutamate + NH₃ = ADP + phosphate + L-glutamine.
- Enzyme regulation**¹: The activity of this enzyme is controlled by adenylation under conditions of abundant glutamine. The fully adenylation enzyme complex is inactive (By similarity). [By similarity](#)
- GO - Molecular function**¹
 - ATP binding [Source: UniProtKB-KW](#)
 - glutamate-ammonia ligase activity [Source: UniProtKB-EC](#)
- GO - Biological process**¹
 - glutamine biosynthetic process [Source: InterPro](#)
 - heterocyst differentiation [Source: UniProtKB-KW](#)
 - nitrogen fixation [Source: UniProtKB-KW](#)

Complete GO annotation...
Keywords - Molecular function¹
Ligase
Keywords - Biological process¹
Nitrogen fixation
Keywords - Ligand¹
ATP-binding, Nucleotide-binding
Names & Taxonomy¹



Formatos de sequências

fasta

genbank



Formatos de sequências



- Por razões históricas, as BDs de sequências permitem a visualização (ou exportação) dos seus registos em *flat files* (texto) com uma dada estrutura
- Vários formatos distintos são usados pelas BDs e pelas ferramentas existentes
- Formatos usados pelo NCBI, EBI e DDBJ são muito semelhantes



Formato FASTA

.fna

```
>gi|1322283|gb|U54469.1|DMU54469 Drosophila melanogaster (...)  
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTT (...)  
GCTGCCTTTGGCCACCAAATCCCAAACCTTAATTAAAGAATTAAATAATT (...)  
TAACCTACGCAGCTTGAGTGCGTAACCGATATCTAGTATACATTTTCGATA (...)
```

.faa

```
>gi|1322285|gb|AAC03525.1| eukaryotic initiation factor 4E-I  
[Drosophila melanogaster]  
MQSDFHRMKNFANPKSMFKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQET  
GEPAGNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEI  
TSFDTVEDFWSLYNHKPPSEIKLGSDYSLFKKNIRPMWEDAAN (...)
```

- Linha de definições/ comentários iniciada com >
- Esta linha pode incluir vários identificadores de BDs e identificação da sequência
- Seguem-se várias linhas com a sequência

+ **Formato GenBank**

- Começa com um header iniciado pela palavra chave LOCUS e que tem um conjunto de outros campos identificados por um título em maiúsculas
- No meio tem a tabela de “features” com a anotação, i.e. a informação biológica relevante
- No final tem a sequência iniciada pela palavra ORIGIN; no início de cada linha tem o n° da posição
- Cada registo termina com // (terminador)

+ Dissecando o formato GenBank

LOCUS:

LOCUS DMU54469 2881 bp DNA linear INV 22-FEB-1998

Locus name

Tamanho da sequência

tipo

divisão

Data submissão

Divisões do NCBI:

BCT – bactérias

INV – invertebrados

MAM – outros mamíferos

PRI – primatas

PLN – plantas

...

DEFINITION:

DEFINITION Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene, alternative splice products, complete cds.

Na linha de definição temos um sumário do conteúdo biológico do registo

+ Dissecando o formato GenBank ...

ACCESSION:

ACCESSION U54469

Código do registo (chave primária)

VERSION:

U54469.1 GI:1322283

Acession.version

GI:geninfo identifier; cada versão de um registo tem um GI diferente

KEYWORDS:

KEYWORDS .

Este campo serve para colocar palavras chave sobre o registo. O seu uso é desencorajado por muitos.

SOURCE:

Organismo e informação taxonómica

SOURCE *Drosophila melanogaster* (fruit fly)

ORGANISM [*Drosophila melanogaster*](#)

Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta;
Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
Muscomorpha; Ephydroidea; Drosophilidae; *Drosophila*.

+ Dissecando o formato GenBank ...



REFERENCE: Referências bibliográficas relacionadas com o registo

REFERENCE 1 (bases 1 to 2881)
AUTHORS Lavoie,C.A., Lachance,P.E., Sonenberg,N. and Lasko,P.
TITLE Alternatively spliced transcripts from the Drosophila eIF4E gene
produce two different Cap-binding proteins
JOURNAL J. Biol. Chem. 271 (27), 16393-16398 (1996)
PUBMED [8663200](#)
(...)

FEATURES:

Esta secção do registo contém as anotações biológicas sendo iniciada com a palavra chave FEATURES. Está organizada em pares chave/ localização:

Na 1ª coluna temos as chaves (tipo de anotação) que pode tomar valores como: source, gene, mRNA, CDS, etc.

Na 2ª colunas temos a informação respeitante a esta chave.

+ Dissecando o formato GenBank : features

```
source          1..2881
                /organism="Drosophila melanogaster"
                /mol_type="genomic DNA"
                /db_xref="taxon:7227"
                /chromosome="3"
                /map="67A8-B2"
```

Bases às
quais se
refere a
anotação

Referência à BD
de taxonomia do
NCBI

Cromossoma e
localização no
cromossoma

Coding sequences

exons

Refs ao registo
da proteína

Seq. proteína

```
CDS            join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
                /gene="eIF4E"
                /note="Method: conceptual translation with partial peptide
                sequencing"
                /codon_start=1
                /product="eukaryotic initiation factor 4E-II"
                /protein_id="AAC03524.1"
                /db_xref="GI:1322284"
                /translation="MVVLETEKTSAPSTEQGRPEPPTSAAAPAEA(...)"
```

+ Explorando o ENA

- Cursos online no site do EBI:
 - <http://www.ebi.ac.uk/training/online/course/european-nucleotide-archive-quick-tour>
 - <http://www.ebi.ac.uk/training/online/course/european-nucleotide-archive-using-primary-nucleoti>
- Passos para explorar o conteúdo da ENA:
 - *What is ENA*
 - *When to use ENA*
 - *How to search and browse ENA*
 - *Exploring an EMBL-Bank entry*
 - *How to export sequence and download data*
 - *Guided examples*
 - *Exercises*

+ NCBI how to's

<http://www.ncbi.nlm.nih.gov/guide/all/#howtos>

How to: Find a curated version of a sequence record (NCBI Reference Sequence)

How to: Obtain genomic sequence for/near a gene, marker, transcript or protein

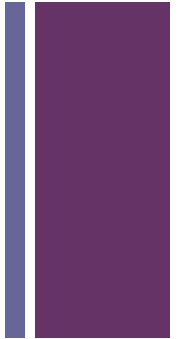
How to: Retrieve all sequences for an organism or taxon

How to: Download the complete genome for an organism

How to: Find transcript sequences for a gene

...

+ Explorando a UniProt



- Curso online no site EBI:

- <http://www.ebi.ac.uk/training/online/course/uniprot-quick-tourversion-0>

- Passos para explorar o conteúdo da UniProt:

- *What is UniProt*
 - *UniProt databases*
 - *Searching data from UniProt*



Bioinformatics for the terrified



- Hands-on

- <http://www.ebi.ac.uk/training/online/course/bioinformatics-terrified>