

Laboratórios de Bioinformática

Sessão 1



1

Porquê Bases de dados 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



2

Redundância de dados

- É importante ter em conta, quando se realizam pesquisas nas bases de dados, que algumas destas sequências possam ser redundantes.
- A redundância das sequências deve-se, por exemplo, ao facto de o mesmo gene (ou genoma) ter sido sequenciado por diferentes laboratórios.
 - Durante o surto de uma doença causado por uma bactéria ou vírus (e.g. SARSCoV2), diferentes laboratórios sequenciam o genoma desta espécie, os quais podem apresentar diferenças devido à qualidade da sequenciação ou presença de mutações.
- Assim, diferentes bases de dados podem conter informação redundante para um dado gene.



3

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**
- 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
- Em 2008 **ENA** substitui EMBL-Bank integrando outros tipos de dados
 - e.g. next generation sequencing)



4

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>
- 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
 - KEGG <https://www.kegg.jp/>



5

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 (...)



6

Bases de dados primárias e secundárias

	Primary database	Secondary database
Synonyms	Archival database	Curated database; knowledgebase
Source of data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary databases
Examples	ENA , GenBank and DDBJ (nucleotide sequence) ArrayExpress and GEO (functional genomics data) Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)	InterPro (protein families, motifs and domains) UniProt Knowledgebase (sequence and functional information on proteins) Ensembl (variation, function, regulation and more layered onto whole genome sequences)

<https://www.ebi.ac.uk/training-beta/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/>



7

Classificação das Bases de dados para 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
 - **GenBank**
 - www.ncbi.nlm.nih.gov/GenBank
 - **DDBJ**
 - www.ddbj.nig.ac.jp

<http://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources>



8

Classificação das Bases de dados para sequências

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



9

Classificação das Bases de dados para sequências

- **Secundárias**
 - Bases de dados com dados analisados e eventualmente curados por especialistas
 - envolve 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



10

Classificação das Bases de dados para 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
- **UniProtKB** (<http://www.uniprot.org/>)
 - repositório de informação sobre proteínas
 - sequências e anotação
 - 3 componentes
 - UniParc
 - UniProtKB (contém a SwissProt e a TrEMBL)
 - UniRef

11

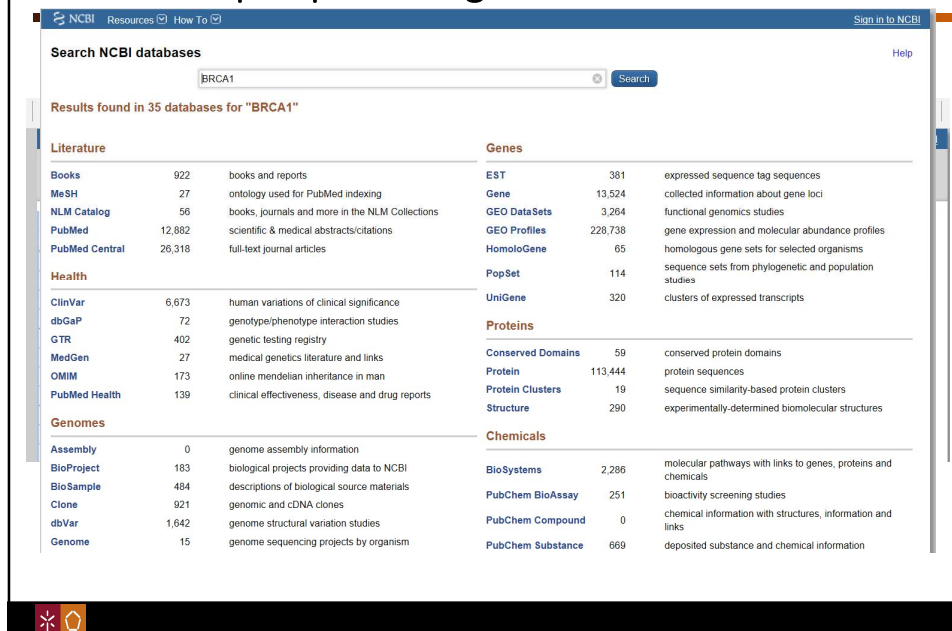
Interfaces: pesquisa integrada EBI

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

The screenshot displays the EBI Search website. At the top, there's a navigation bar with links for Services, Research, Training, and About us. Below this is a search bar with the text 'brca1' entered. The search results are displayed below the search bar, showing 21 results out of 53,363 in all results. The results are categorized by source, including Gene & protein summaries, Enzyme, and Protein families. The first result is for 'BRCA1 (PSCP, BRP53, PPP1R53, BRCC1, PNC4, BRCA1, FANCS, BROVCA1, TRIS, ENSG0000012048) Human (Homo sapiens)'. Other results include 'BRCA1 (PSCP, BRP53, PPP1R53, BRCC1, PNC4, BRCA1, FANCS, BROVCA1, TRIS, ENSG0000012048) Human (Homo sapiens)' and 'BRCA1 (PSCP, BRP53, PPP1R53, BRCC1, PNC4, BRCA1, FANCS, BROVCA1, TRIS, ENSG0000012048) Human (Homo sapiens)'. The bottom of the page features a footer with links to Structural Bioinformatics 2016, Ensembl Bite sized - September 2016, and ArrayExpress: why and how to submit your data.

12

Interfaces: pesquisa integrada NCBI



NCBI Resources How To Sign in to NCBI

Search NCBI databases

BRCA1 Search

Results found in 35 databases for "BRCA1"

Literature			Genes		
Books	922	books and reports	EST	381	expressed sequence tag sequences
MeSH	27	ontology used for PubMed indexing	Gene	13,524	collected information about gene loci
NLM Catalog	56	books, journals and more in the NLM Collections	GEO DataSets	3,264	functional genomics studies
PubMed	12,882	scientific & medical abstracts/citations	GEO Profiles	228,738	gene expression and molecular abundance profiles
PubMed Central	26,318	full-text journal articles	HomoloGene	65	homologous gene sets for selected organisms
Health			PopSet	114	sequence sets from phylogenetic and population studies
ClinVar	6,673	human variations of clinical significance	UniGene	320	clusters of expressed transcripts
dbGaP	72	genotype/phenotype interaction studies	Proteins		
GTR	402	genetic testing registry	Conserved Domains	59	conserved protein domains
MedGen	27	medical genetics literature and links	Protein	113,444	protein sequences
OMIM	173	online mendelian inheritance in man	Protein Clusters	19	sequence similarity-based protein clusters
PubMed Health	139	clinical effectiveness, disease and drug reports	Structure	290	experimentally-determined biomolecular structures
Genomes			Chemicals		
Assembly	0	genome assembly information	BioSystems	2,286	molecular pathways with links to genes, proteins and chemicals
BioProject	183	biological projects providing data to NCBI	PubChem BioAssay	251	bioactivity screening studies
BioSample	484	descriptions of biological source materials	PubChem Compound	0	chemical information with structures, information and links
Clone	921	genomic and cDNA clones	PubChem Substance	669	deposited substance and chemical information
dbVar	1,642	genome structural variation studies			
Genome	15	genome sequencing projects by organism			

13

NCBI

National Center for Biotechnology Information

14

Pesquisa por textos biomédicos: PubMed

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

The screenshot shows the PubMed website interface. At the top, there's a search bar with the text 'brca1 brca2' entered. Below the search bar, there's a notification about NCBJ testing https on public web servers. The main content area displays search results for 'brca1 brca2'. On the left, there's a sidebar with filters like 'Article types', 'Text availability', 'Publication dates', 'Species', and 'Other filters'. The main results list includes several articles, such as 'BRCA mutations and survival in breast cancer: an updated systematic review and meta-analysis' and 'Clinically Significant Unclassified Variants in BRCA1 and BRCA2 Genes Among Korean Breast Cancer Patients'. On the right, there's a 'Results by year' chart and a 'Related searches' section.

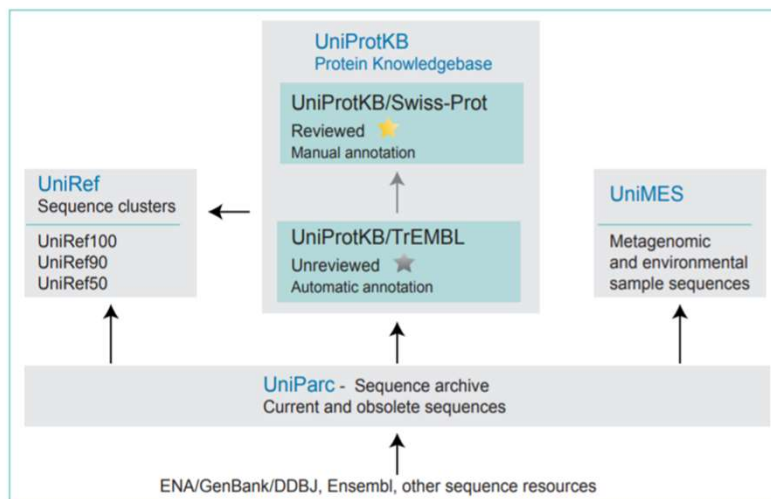
15

UniProt

Universal Protein Resource

16

UniProt



17

UniProt

■ 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 que aguardam curaço manual

18

UniProt

■ 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 bases de dados
 - SwissProt
 - TrEMBL
 - PIR
 - 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
 - BLAST



19

UniProt

■ 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



20

Explorando um registo UniProt

UniProtKB - P00964 (GLNA_NOSS1)

Display

Publications
Feature viewer
Feature table

Function

Protein | Glutamine synthetase
Gene | glNA
Organism | Nostoc sp. (Strain PCC 7120 / SAG 25.82 / UTEX 2576)
Status | Reviewed - Annotation score: **** - Experimental evidence at protein level

Function

Catalytic activity
ATP + L-glutamate + NH₃ = ADP + phosphate + L-glutamine.

Enzyme regulation
The activity of this enzyme is controlled by adenylation under conditions of abundant glutamine. The fully adenylylated enzyme complex is inactive (by similarity). [By similarity](#)

GO - Molecular function

- ATP binding [#Source: UniProtKB-KW](#)
- glutamate-ammonia ligase activity [#Source: UniProtKB-EC](#)

GO - Biological process

- glutamine biosynthetic process [#Source: InterPro](#)
- heterocyst differentiation [#Source: UniProtKB-KW](#)
- nitrogen fixation [#Source: UniProtKB-KW](#)

Complete GO annotation...

Keywords - Molecular function
ligase

Keywords - Biological process
nitrogen fixation

Keywords - Ligand
ATP-binding, Nucleotide-binding

Names & Taxonomy

21

Ensembl



22

Ensembl

- é um navegador de genomas de vertebrados, que suporta a investigação em genómica comparativa, evolução, variação de sequências e regulação da transcrição.
- O Ensembl anota genes, calcula alinhamentos múltiplos, prevê a função regulatória e coleta dados de doenças.
- As ferramentas de ensembl incluem BLAST, BLAT, BioMart e o Variant Effect Predictor (VEP).
- Cada espécie tem a sua própria página incluindo Homo sapiens
 - http://www.ensembl.org/Homo_sapiens/Info/Index

23

Ensembl

The screenshot displays the Ensembl genome browser interface for Homo sapiens. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, RefSeq, Downloads, Help & Docs, and Blog. A search bar is present in the top right corner. The main content area is divided into several sections:

- Search Human (Homo sapiens):** A search bar with a "Go" button and a link to "Search all categories".
- Genome assembly:** A section for GRCh38.p13 (GCA_000001485.28) with links to "More information and statistics", "Download DNA sequence (FASTA)", "Convert your data to GRCh38 coordinates", and "Display your data in Ensembl".
- Gene annotation:** A section with links to "What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNA", "More about the genebuild", "Download FASTA files for genes, cDNAs, ncRNA, proteins", "Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins", and "Update your old Ensembl Gtfs".
- Comparative genomics:** A section with links to "What can I find? Homologues, gene trees, and whole genome alignments across multiple species", "More about comparative analysis", and "Download alignments (GFF)".
- Regulation:** A section with links to "What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations", "More about the Ensembl regulatory build and microarray annotation", "Experimental data sources", and "Download all regulatory features (GFF)".
- Variation:** A section with links to "What can I find? Short sequence variants and longer structural variants, disease and other phenotypes", "More about variation in Ensembl", "Download all variants (GFF)", and "Variant Effect Predictor".

The footer includes links to "About Us", "Get help", "Our sister sites", and "Follow us".

24

Explorando o Ensembl

- Curso online no site do EBI:
 - <http://www.ebi.ac.uk/training/online/course/ensembl-quick-tour-0>
- Passos principais :
 - *What is Ensembl*
 - *What can you do with Ensembl*
 - *Searching and visualizing data from Ensembl*
 - *Getting data from Ensembl*
- Curso mais detalhado:
 - <http://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes>



25

Tutoriais



26

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



27

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

...



28

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*



29

Bioinformatics for the terrified

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

Bioinformatics
for the terrified



30

QUIZ



31