# Omics



# Exponential growth of data in biology

Storing large databases and providing access

Data types:

- Nucleotide sequences
- Protein sequences
- Protein sequence motifs and patterns
- Three-dimensional structure of macromolecules
- Gene expression data
- Metabolic pathways
- Disease related variations

#### ==> Bioinformatics ==> BIG DATA



Figure 1 | **Biological database architecture.** Most biological databases use a three-tier architecture that consists of a database management system, a middleware layer and a web interface.

# Biological databases

Three basic requirements:

1. Ready access to the collected pool of information

2. An easy way to extract and manipulate data quickly

3. Enough computing power and storage capacity.

#### Updates and revisions

- Data entry and quality checks
- Researchers (groups) enter the data
- Curators add and update the data
- Marking and removing wrong data
- Types and extent of checking
- Consistency, redundancy, updates

#### Updates and revisions

Bioinformatics tools and databases can change with time!

- Save the data locally e.g. identifiers, sequences

## **Bioinformatics tools**

There is a great number of bioinformatics tools

- First check, if there is a tool already exists for your problem
- Don't worry if you don't know all the tools Good starting points: NAR Database and Web Servers Issues
- Convergence of resources

## Questions

- Which proteins belong to a given gene?
- What is their known function?
- What kind of diseases it is associated with?

- NCBI
- UNIPROT
- ENSEMBL

# Name and identifier

#### Identifier

Entry name in Uniprot (pl. ADH6\_HUMAN) Locus in GenBank (pl. HUMADH6A01)

#### **Accession number**

Unique Does not change

P28322 (Uniprot) AH001409 (Genbank) ENSG00000172955 (ENSEMBL)

Version number

#### Names

Rad24 in Saccharomyces cerevisiae (budding yeast) the DNA-damage checkpoint-pathway gene

Rad24 in Saccharomyces pombe (fission yeast) is involved in the checkpoint pathway,

BUT it is not ortholog of S. cerevisiae Rad24

The correct *S. pombe* orthilog is **rad17**, which is not the same as the **Rad17** gene in *S. cerevisiae*.

There are severl rad gene in C. elegans but these are not orthologs of the *S. cerevisiae* Rad17. The closest relative of this gene is mrt-2.

#### Sequence databases: NCBI

S NCRI	National Center for Biotechnology Information					
> NCBI	National Library of N	<u>ledicine</u>	Na	ational Ir	nstitutes o	of Health
PubMed All Dat	abases BLAST	OMIM	Books	TaxB	rowser	Structure
Search Gene	💌 for			Go		
SITE MAP	What does NCBI do?				Hot Spots	
Alphabetical List						
Resource Guide	Established in 1988 as a national resource			ce	Clusters of	
	for molecular biology information, NCBI				orthologous groups	
About NCBI	creates public databases, conducts research					
An introduction to	in computational biology, develops software			are	Coffee Break, Genes & Disease,	
NGBI	tools for analyzing genome data, and			3		
GenBank	enBank disseminates biomedical information - all for				NCBI Handbook	
Sequence	the better understanding of molecular				1004070074100 6411441044107	
submission support	processes affecting human health and				Electronic PCR	
and software	disease. More about NCBI				Entrez Home	
Literature	Protein Clusters					
databases	The new Protein Clu	uctore dotaba	na contain		Entrez	Tools
PubMed, OMIM,	ubMed, OMIM, ooks, and PubMed Reference Sequence (RefSeq) protein records that are grouped and annotated by					
Books, and PubMed					Gene expression	
Central	sequence and functi	ional similarit	y. Source	omnibus (GEO)		
	sequences come fro	om the comp	lete genon	nes	<b>•</b> • • • • • • • • • • • • • • • • • •	
Molecular	of prokaryotes, plas	f prokaryotes, plasmids, and organelles.			Human	genome
databases	Read more about Protein Clusters.			2	resources	

.

Sequences

### Sequence databases: Genbank

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences

GenBank is accessible through the NCBI Entrez retrieval system, which integrates data from the major **DNA and protein sequence databases** along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed.

Genbank flat file

Filtered data in the RefSeq database but not in GenBank.

#### UniProt (Knowledgebase)

- millions of protein sequences
- experimental information extracted from scientific literature
- provision of complete and up to date Reference proteomes
- integration of large-scale genomics and proteomics experiments
- interoperability between data and services in the biological, medical, translational and clinical domains



# Uniprot: SwissProt+TrEMBL

UniProt (Universal Protein): A central repository of **protein sequence** and function

#### Swiss-Prot

- The protein sequence database that contains core data enhanced by manually curated annotation.

Swiss-Prot annotation describes features such as function, posttranslational modifications, domains and sites, secondary and quaternary structure, diseases, and sequence variants.

#### TrEMBL

The computer-annotated section of UNIPROT

Contains translations of all coding regions and sequences extracted from literature.

The quality of the data is dependent upon the information provided by the submitter of the nucleotide data.

Enhanced through redundancy removal, and automatic annotation

## Swissprot/TrEMBL

#### **UniProtKB/Swiss-Prot**

- Non-redundant
- High level of integration
- High level of manual curation
- Contains 550116 entries

#### **UniProtKB/TrEMBL**

- Translations of CDS in EMBL/GenBank/DDBJ
- Automatic annotation
- Contains 54247468 entries

Pro	otein existence (PE):	entrie	es %
1:	Evidence at protein level	1436	16.6%
2:	Evidence at transcript level	57673	10.5%
3:	Inferred from homology	387609	70.5%
4:	Predicted	11444	2.1%
5 <b>:</b>	Uncertain	1954	0.4%

Protein existence (PE):		entries	00
1:	Evidence at protein level	121173	0.22%
2:	Evidence at transcript level	992520	1.80%
3:	Inferred from homology	11944683	21.61%
4:	Predicted	42212303	76.37%
5:	Uncertain	0	0.00%

#### What are proteomes?



#### Feature viewer



#### **Cross-references in UniProt**

Organism-specific databases AGD ArachnoServer CGD ConoServer CTD CYGD dictyBase	Sequence databases EMBL IPI PIR RefSeq UniGene	Enzyme database BioCyc BRENDA Pathway Reactome UniPathwa	and pathway	Family and domain databases Gene3D HAMAP InterPro PANTHER PIRSF Pfam PPINTS
EchoBASE EcoGene euHCVdb EuPathDB FlyBase GeneCards GeneDB_Spombe				ProDom PROSITE SMART SUPFAM TIGRFAMs
Generarm GeneLynx Gramene H-InvDB HGNC HPA LegioList Leproma				2D-gel databases 2DBase-Ecoli ANU-2DPAGE Aarhus/Ghent-2DPAGE COMPLUYEAST-2DPAGE Comea-2DPAGE DOSAC-COBS-2DPAGE ECO2DBASE
MaizeGDB	Uni	ProtKB/Swiss-P	Prot	OGP PHCI-2DPAGE
MIM		explicit links		PMMA-2DPAGE
neXtProt Orphanet PharmGKB PseudoCAP RGD SGD TAIR				REPRODUCTION-2DPAGE Siena-2DPAGE SWISS-2DPAGE UCD-2DPAGE World-2DPAGE
TubercuList				PTM databases
Xenbase ZFIN				GlycoSuiteDB PhosSite
Proteomics and PPI				Phosphosite
DIP IntAct Mint PeptideAtlas				Genome annotation databases Ensembl
PRIDE ProMEX		Phylogenomic	Miscellaneous	EnsemblBacteria
String	Protein family/group		ArrayExpress	EnsemblMetazoa
	databases	GeneTree	Bgee Bindin - DB	EnsemblPlants
3D structure databases	Allergome	HOGENOM	CleanEx	GenomeReviews
DisProt	GermOnline	HOVERGEN	dbSNP	GenelD
PDB	MEROPS	OMA	DrugBank	KEGG
PDBsum	PeroxiBase PotaseDB	OrthoDB	GO	TIGR
ProteinModelPortal	REBASE	PhylomeDB	NextBio	UCSC
SMR	TCDB	ProtClustDB	PMAP-CutDB	Vectorbase

## Function of proteins

What do we mean by function?

- could mean different things
- many levels of granularity
- functional categories are somewhat artificial
- ambiguities in naming

The same name can be used to describe different concepts, e.g.

- Glucose synthesis
- Glucose biosynthesis
- Glucose formation
- Glucose anabolism
- Gluconeogenesis

All refer to the process of making glucose Makes it difficult to compare the information

#### Solution: use Ontologies and Data Standards, Controlled vocabularies

# Ontologies

An ontology is a formal specification of terms and relationships between them – widely used in biology and boinformatics (e.g. taxonomy)

- The relationships are important and represented as graphs
- Ontology terms should have definitions
- Ontologies are machine-readable
- They are needed for ordering and comparing large data sets

# Gene Ontology (GO)

- http://www.geneontology.org
- Many annotation systems are organism-specific or different levels of granularity
- GO introduced standard vocabulary first used for mouse, fly and yeast, but now generic
- Three ontologies

# GO ontology

- Molecular function: tasks performed by gene product
  - -e.g. G-protein coupled receptor
- **Biological process**: broad biological goals accomplished by one or more gene products

-e.g. G-protein signaling pathway

- Cellular component: part(s) of a cell of which a gene product is a component; includes extracellular environment of cells
  - –e.g nucleus, membrane etc.

## GO hiearchy



## GO hierarchy



# How do gene products get GO terms?

- Electronic annotation:
- Through mappings to other biological entities and then automatic inference to proteins
- Manual annotation:
- Model organism databases
- Gene Ontology Annotation (GOA) project
- Evidence codes –attached to all GO annotations to show the source

## GO- slim

- GO slims are cut-down versions of the GO ontologies
- GO slims can give a summary of the result
- GO slims may be specific to species or to particular areas of the ontologies
- GO provides a generic GO slim
- Used for enrichment calculations

### Genome Browser: ENSEMBL

The Genome Reference Consortium (GRC) (1) released a new human genome assembly, GRCh38 (GCA\_000001405.15), in December 2013.

Ensembl processes large-scale genomic data for chordate and model organisms

Genes and transcripts are annotated by aligning protein and mRNA sequences to the genomic sequence

#### **Genome Variations**

#### Polymorphisms



#### **Cancer Genome Projects**



(Common) polymorphisms: Present in 1% in population Passenger and driver mutations

## Visualization of cancer data

#### COSMIC

http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=TP53

#### CbioPortal (TCGA)

http://www.cbioportal.org/

#### ICGC

https://dcc.icgc.org/genes/ENSG00000141510

#### Memphis (pediatric cancer)

https://pecan.stjude.org/#/proteinpaint/TP53

## **REST** interface



Easier way of providing interoperability between computer systems through web services.

 The data can be requested with simple HTTP requests (in the URL)

 returned in a variety of programatic and bioinformatical relevant formats such as JSON, XML, TXT and FASTA.

- reproducible, scalable, powerful

Uniprot, PDB, ELM ...