About us

The EMBL-European Bioinformatics Institute (EMBL-EBI) maintains one of the world’s most comprehensive collections of freely available biological databases. Our free tools allow you to analyse this information and share it with the research community.

About this guide

This document provides an overview of EMBL-EBI’s core resources. You can find overviews of these resources in Train online.

www.ebi.ac.uk/training/online

Support

Detailed help documentation is available through each of the resources described here. You can also find answers to some of your questions through our homepage:

www.ebi.ac.uk/help

Training

EMBL-EBI offers training at all levels. Our hands-on training courses are held regularly on the Genome Campus in Hinxton, UK and at host institutions throughout the world. Train online guides you through many bioinformatics resources, in your own time and at your own pace.

www.ebi.ac.uk/training

Stay in touch

You can find EMBL-EBI news, service updates and training opportunities on our website, Twitter and Facebook.

https://www.ebi.ac.uk

@emblebi

/EMBLEBI

Genes and genomes

Next-generation sequencing has transformed the task of collecting DNA sequence data. We no longer sequence just single genomes but also sets of related genomes in structured studies. This has made possible the 1000 Genomes Project, an international initiative to build a comprehensive catalogue of all common human variation. We also capture genetic variation of medical relevance. Such data require careful attention to access control, ensuring that only authorised scientists can use them, and that they do so in a way that reflects the spirit of consent and confidentiality.

Ensembl

Ensembl produces and maintains both automatic and manually curated annotation on selected eukaryotic genomes. Automatic annotation is based on mRNA and protein information. Ensembl provides valuable insights into variation within and between species, and allows you to compare whole genomes to identify conserved elements. It is integrated with several other important molecular resources, for example UniProt, and can be accessed programmatically. Ensembl is developed as a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute.

Ensembl Genomes

Ensembl Genomes is a portal to genome-scale data from bacteria, protists, fungi, plants and invertebrate metazoa, through a unified set of interactive and programmatic interfaces. Domain-centric resources (focused on particular areas of scientific interest) are developed in collaboration with the appropriate scientific communities, and integrated in the broad taxonomic context through comparative analysis and standardised annotation.

European Nucleotide Archive

ENA, a member of the International Sequence Database Collaboration, contains all the nucleotide sequences in the public domain and consolidates data from EMBL-Bank, the European Trace Archive and the Sequence Read Archive. The EBI Search returns results from selected annotated sections of the ENA.

European Genome-phenome Archive

EGA allows you to explore datasets from numerous genotype experiments – including case-control, population and family studies – that are supplied by a range of data providers.

Metagenomics Portal

Our Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data, and provides insights into the functional and metabolic potential of a sample.

Search www.ebi.ac.uk

Our search service is your gateway to a world of information spanning DNA, genes, functional genomics, proteins, structures, small molecules, enzymes, interactions, pathways, scientific publications and patent sequences.
Gene expression

Genome-wide expression assays, originally using microarrays and more recently high-throughput sequencing, can answer specific biological questions and provide reference data sets. The associated large-scale expression datasets can be used to answer questions that might not be related to the study for which the data were originally generated. For example, a gene expression study that reveals differentially expressed genes characteristic of a particular type of cancer may also reveal candidate genes for therapeutics development.

EMBL-EBI’s functional genomics resources provide easy access to gene expression data and related information. Our visualisation tools collate information from multiple data sets and present it in an intuitive way.

ArrayExpress
www.ebi.ac.uk/arrayexpress

The MIAME-standard compliant ArrayExpress Archive stores functional genomics experiments performed using RNA-Seq/ChIP-Seq and array-based technologies.

Expression Atlas
www.ebi.ac.uk/gxa

The Expression Atlas allows you to search for gene expression changes measured in various cell types, organism parts, disease states and many other biological and experimental conditions. It represents a curated subset of the ArrayExpress Archive experiments, and can be downloaded for local use.

R-Workbench
www.ebi.ac.uk/Tools/rcloud

The R-Workbench provides access to R and Bioconductor in the EBI cloud, and rapid access to data stored in EMBL-EBI databases such as ArrayExpress. You can upload your own data and use EMBL-EBI’s infrastructure to perform your data analyses.

Leanne works on discovery and validation of candidate drug targets in a pharmaceutical company. She explains: ‘I do a lot of cloning and expression analysis. I need to find published cDNA sequences and their orthologues, so that I can find appropriate models in which to validate my targets. For discovery, I’d like a simple way to search across all public-domain biological information. A colleague mentioned a gene to me in passing recently and I only caught the name of it. I’d like to be able to find out more about it without having to forage around in lots of different publications.’

EBI resources for Leanne:

EBI Search for quick summary reports on specific genes.
BLAST to root out vector contamination.
The Expression Atlas for transcription profiles and functional genomics.
UniProt to learn more about protein function.
ChEMBL for druggability of target molecules.
Francesco, marine biologist

Francesco is Head of Department in a new collaborative centre for marine biology. His group is doing metagenomic studies of extreme marine environments but his department has just secured funding to look for novel wound-healing compounds in marine algae. “We’re about to scale up our sequencing efforts severalfold. We need to put some pipelines in place that will enable us to make sense of all the data that we’re generating. The extremophile project is a collaborative effort, and we want to be able to share information with our research network.”

**EBI resources for Francesco:**

**European Nucleotide Archive** for DNA barcodes to identify organisms in his samples. His team could set up a pipeline to submit their data to ENA and make it publicly available.

**Ensembl and Ensembl Genomes** to characterise organisms by comparing them with a wealth of information on the genomes of many fully sequenced organisms.

**InterProScan** to help identify domains and motifs in the novel sequences he is generating that may imply specific functions.

**UniMes**, a repository specifically developed for metagenomic and environmental data.

**Metagenomics Portal** for powerful analysis of samples containing a variety of organisms.

**Web Services** to help create pipelines.

---

**Proteins and proteomics**

Our protein and nucleotide data are extensively linked, both at the level of underlying data and through the coordination of web resources. This helps us capture, organise and interpret sequence-related data, providing information freely in a variety of formats including user-friendly web sites. Wherever possible, we collaborate with other groups worldwide with similar aims.

**UniProt**

[www.uniprot.org](http://www.uniprot.org)

This comprehensive resource organises protein sequence information so that you can find splice variants, functional information and links to related resources from a single entry. UniProt is a joint effort with the Swiss Institute of Bioinformatics and Georgetown University. It comprises four components: the UniProt Knowledgebase (UniProtKB), Reference Clusters (UniRef), Archive (UniParc) and Metagenomic and Environmental Sequences (UniMES). UniProtKB is the hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.

**InterPro**

[www.ebi.ac.uk/interpro](http://www.ebi.ac.uk/interpro)

InterPro classifies proteins into families and predicts the presence of important domains and sites. InterProScan allows you to query your sequence against the database.

**PRIDE**

[www.ebi.ac.uk/pride](http://www.ebi.ac.uk/pride)

The Proteomics Identifications Database is a centralised, standards-compliant, public repository for proteomics data. It contains protein and peptide identifications and their associated supporting evidence. PRIDE also captures details of post-translational modifications.

---

**Macromolecular structures**

Three-dimensional structures give us mechanistic insight into how macromolecules work, and help explain how their functions are disrupted by mutation or interaction with small molecules. Our structural data resources are engineered to reflect the needs of a growing and diverse structural biology community.

**PDBe**

[www.ebi.ac.uk/pdbe](http://www.ebi.ac.uk/pdbe)

The Protein Data Bank in Europe (PDBe), part of the Worldwide Protein Data Bank (wwPDB), allows you to search all the macromolecular structures in the public domain. Its tools help you perform sophisticated analyses of the data, including sub-structure searches and structural comparisons. You can analyse detailed molecular interactions and correlate them with sequence or structure patterns.

**PDBsum**

[www.ebi.ac.uk/pdbsum](http://www.ebi.ac.uk/pdbsum)

This pictorial database provides an overview of macromolecular structures deposited in the Protein Data Bank archive.

**ProFunc**

[www.ebi.ac.uk/profunc](http://www.ebi.ac.uk/profunc)

ProFunc allows you to identify the likely biochemical function of a protein from its 3D structure.
John is a medicinal chemist in a large, multinational pharmaceutical company. “I'm constantly looking for new information, so searchable databases of existing chemical literature are very useful,” he says. “For hit and lead optimisation, new ideas that could get me into novel chemical space would be great. Information on my pharmacophore would be really useful too; things like: is it likely to be toxic? Or, can we predict if analogues will be biologically active?”

EBI resources for John:

ChEMBL for detailed information about over a million bioactive compounds.

Kinase SARfari, an integrated chemogenomics workbench linking kinase sequences with their structures, compounds that target them and screening data.

GPCR SARfari for G-protein-coupled receptors.

ChEBI for information about small molecular entities and links to the macromolecules with which they interact.

PDBe for 3D structures of macromolecules, including protein–ligand interactions.
Leon, postdoc

Leon is a postdoc working on quorum sensing in bacteria. He wants to understand what makes a normally harmless bacterium pathogenic in the lungs of people with cystic fibrosis. "I'm using a combination of transcriptomics, proteomics and metabolomics to understand these pathogenic changes better," he says. "I end up with big spreadsheets of protein or gene IDs and I'm trying to piece together which signalling pathways are involved in flipping to the pathogenic state. When I come across a gene or protein that's not familiar to me, I want to be able to find out about it quickly and put it in the context of other genes or proteins that are expressed at the same time."

EBI resources for Leon:

EBI Search for quick summary reports of genes or proteins.

Enzyme Portal to learn more about a particular enzyme that switches a pathway on or off.

InterPro for uncharacterised proteins or novel protein families and domains.

UniProt for the function of proteins in his lists of IDs.

Gene Ontology to find out which ones are involved in the same processes.

Reactome to map gene and protein lists to pathways.

Leon could also submit his transcriptomics and proteomics experiments to ArrayExpress and PRIDE to analyse his results in the context of other similar experiments.

Interactions, pathways & networks

Computational systems biology takes us beyond the identification of molecular ‘parts lists’ for living organisms and towards synthesising information to generate and test new hypotheses about how biological systems work.

IntAct

www.ebi.ac.uk/intact

The IntAct molecular interaction database and analysis system provides a central, standards-compliant repository of molecular interactions, including protein–protein, protein–small molecule and protein–nucleic acid interactions.

Reactome

www.reactome.org

Reactome is a database of human biological pathways built from connected reactions that encompass all biological events as well as classical biochemical events. Content is authored by expert biologists and linked to scientific literature that experimentally validates the steps in a pathway. The goal is to represent the current biological consensus of human pathways, as a free online reference and a downloadable core dataset. Reactome offers tools for searching and viewing pathways as interactive diagrams, and for analysing datasets by comparing them with other pathways and overlaying expression or interaction data.

BioModels

www.ebi.ac.uk/biomodels

BioModels is a database of published mathematical models describing biological processes. You can simulate the models online and use them to develop your own. Biologists can store, search and retrieve published and annotated mathematical models of biological interest.

Patent sequences

We offer free, unrestricted access to dedicated patent sequence databases that contain valuable information about patent family data, patent equivalents and ‘earliest priority’ dates. These databases also offer access to full-text patent documents.

Our non-redundant patent sequence databases help you search efficiently and return de-duplicated results, saving valuable time and effort.

Patent nucleotide and protein sequences are contributed by the European Patent Office (EPO), US Patent and Trademark Office, and the Japanese (JPO) and Korean (KIPO) patent offices, including World Intellectual Property Organization (WIPO)-approved patents from these offices.

Further agreements with patent offices from Brazil, Canada, China and Mexico have been achieved or are in progress.

www.ebi.ac.uk/patentdata/nr
Judith, plant geneticist

Judith works for a plant-breeding institute that is using genomics to identify new crop strains that are resistant to drought, salt and fungal diseases. “We’re doing linkage studies to find out which genes are involved in resistance to different types of stress,” she explains. “Some of the crops we work on haven’t been sequenced yet. We’ve got a lot of data, both on genomic variants and on differences in expression levels, that we need to map on to well-characterised plants to make sense of it. We also need to find out whether any of our sequences have associated patents so that we don’t waste time and effort.”

CiteXplore
www.ebi.ac.uk/citexplore

CiteXplore covers 25 million abstracts including PubMed, Agricola and patents. It combines literature search with text-mining tools, citation networks and cross references to other bioinformatics resources.

UK PubMed Central
http://ukpmc.ac.uk

UKPMC contains over 2 million full text life science research articles, of which around 250 000 are open access. Combined with the CiteXplore content and functions listed above, it provides integrated text-mining tools as well as grant-reporting services.

Ontologies

The wide variations in terminology between different research communities pose a challenge both for the researcher conducting a search and the computer executing it. Ontology services help identify the most useful information by identifying functionally equivalent terms.

Gene Ontology
www.geneontology.org

The Gene Ontology Consortium provides a controlled vocabulary to describe gene and gene product attributes in any organism.

GO Annotation
www.ebi.ac.uk/QuickGo

GOA provides Gene Ontology assignments for proteins for all species in the UniProt Knowledge Base.

ChEBI
www.ebi.ac.uk/ChEBI

The ChEBI chemistry ontology provides a controlled vocabulary to describe small molecules, including their biological and chemical roles.

Systems Biology Ontologies
www.ebi.ac.uk/sbo

The SBO project develops controlled vocabularies and ontologies for problems in systems biology. SBO is part of BioModels.net.

Experimental Factor Ontology
www.ebi.ac.uk/EFO

EFO is used to describe experimental variables in functional genomics experiments, and is cross-referenced to more than 20 species or domain-specific terminologies. It provides a way to integrate databases using different terminologies and is deployed in searching for the ArrayExpress Archive and Expression Atlas databases.

Ontology Lookup Service
www.ebi.ac.uk/ontology-lookup

OLS is a centralised query interface for ontology and controlled vocabulary lookup. It features interactive graphics that clarify the relationships between terms.
Ola is an MD/PhD. When she’s not practising as an oncologist, she’s working on her research project. “I’m looking for proteomics-based biomarkers for the early detection of bladder cancer, specifically peptides from transitional-cell carcinomas that are shed into the urine,” she explains. “I do mass spec of samples from patients coming in for biopsies. I need to identify the proteins whose production is up- or down-regulated. I’ve started looking at post-translational modifications on these proteins, and have found a phosphoprotein that seems to be upregulated in some patients. I’d like to know more about the cellular processes leading to this change. This might help us find novel therapies, too.”

EBI resources for Ola:

PRIDE for proteomics experiments that have identified proteins from carcinoma cells. Ola could also submit her own research results to PRIDE.

UniProt to find out about the phosphorylation sites on her new biomarker and what type of kinase might catalyse these changes.

IntAct to find evidence of other proteins that interact with her protein of interest.

---

**Tools**

We provide a comprehensive range of bioinformatics tools. Here we highlight a selection of those used most frequently, but you can find a more extended list on our website ([www.ebi.ac.uk/Tools](http://www.ebi.ac.uk/Tools)).

For programmatic access, take a look at our Web Services (SOAP or REST APIs): [www.ebi.ac.uk/Tools/webservices](http://www.ebi.ac.uk/Tools/webservices)

### Nucleotide sequence search

You can use these tools to search nucleotide sequence databases, including ENA and resources for coding sequences, immunoglobulins and high-throughput sequences:
- BLAST Nucleotide
- ENA Sequence Search (Exonerate)

### Multiple sequence alignment

For alignment of three or more sequences to identify regions of conservation, which may indicate functional constraints and infer evolutionary relationships:
- Clustal Omega
- ClustalW2
- WebPRANK
- T-Coffee
- MUSCLE

### Protein functional analysis

For identification of potential protein families, domains and functional sites allowing the the function of a protein sequence to be predicted:
- InterProScan
- Pratt
- Phobius
- RADAR

### Molecular structure analysis

These tools help you determine and visualise macromolecular structures:
- PDBeFold
- QuaternaryStructure

---

### Protein sequence search

You can use these tools to search protein sequence databases including UniProtKB, sequences derived from macromolecular structures, immunoglobulins and sequences from patents:
- BLAST Protein
- PSI-Search

### Pairwise sequence alignment

For alignment of two sequences to identify regions where the sequence is conserved:
- Needle
- LAAlign

### Functional genomics tools

To explore data from gene expression experiments:
- Expression Atlas
- EFO Tools
- EBI R-Workbench

### Text mining

For analysis of text documents to identify important terms and relate them to database entries and ontologies:
- EBIMed
- Whatizit

### Data retrieval and ID mapping

- dbfetch
- ENA / SRA
- ENA / EMBL-SVA
- UniProt / UniSave
- SRS
- PICR
The European Molecular Biology Laboratory

Hinxton, UK: European Bioinformatics Institute

Grenoble, France: Structural biology research and services

Monterotondo, Italy: Mouse biology programme

Hamburg, Germany: Structural biology research and services

Heidelberg, Germany: Main laboratory