**List of useful bioinformatics tools**

**Elaborated by: Patrick da Silva**

 **DATABASES**

In general, databases are huge collections of curated information where you can search specific features of a query. There are two main multi-functional databases divided into several sub-sections, also providing a great assort of useful online tools.

First, the NCBI (National Center for Biotechnology Information) is part of the United States NLM (National Library of Medicine), a branch of the NIH (National Institutes of Health). It is approved and funded by the government of the United States. The NCBI houses one of the most important bibliographic databases for biomedical literature, the PubMed.

The EBI (European Bioinformatics Institute) is an intergovernmental organization inside the EMBL (European Molecular Biology Laboratory), which handles the bioinformatics research and services.

**1. Nucleotide Sequence Databases**

The three main databases inside this section follow the rules of the INSDC (International Nucleotide Sequence Database Collaboration). New and updated data on nucleotide sequences contributed by research teams to each of the three databases are synchronized on a daily basis. They gather annotated information of any kind of DNA and RNA sequences: genes, transcripts, plasmids, chromosomes, genomes…

**- DDBJ (DNA Data Bank of Japan):** <https://www.ddbj.nig.ac.jp/index-e.html>

**- GenBank:** <https://www.ncbi.nlm.nih.gov/genbank/>

**- ENA (European Nucleotide Archive):** <https://www.ebi.ac.uk/ena/browser/home>

**2. Protein Databases**

**- InterPro:** Database for functional analysis of proteins by classifying them into families and predicting domains and important sites.

Access: <https://www.ebi.ac.uk/interpro/>

**- Uniprot:** Resource of protein sequence, functional information, protein-protein interaction, family & domains, subcellular location.

Access: <https://www.uniprot.org/>

**- RCSB Protein Data Bank (RCSB PDB):** Most complete database for protein structural information from X-crystallography, Magnetic Resonance Imaging (MRI), Cryo-EM experiments, and also 3D structure prediction.

Access: <https://www.rcsb.org/>

**- AlphaFold:** Protein structures predicted by DeepMind’s algorithm together with EMBL-EBI.

Access: <https://alphafold.ebi.ac.uk/>

**- STRING:** Database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases. Interactions in STRING are derived from five main sources: Genomic Context Predictions, High-throughput Lab Experiments, (Conserved) Co-Expression, Automated Textmining, and Previous Knowledge in Databases. The STRING database currently covers 24584628 proteins from 5090 organisms.

Usage: search the interactions of a single or multiple proteins by name/identifier or sequence.

Access: <https://string-db.org/>

**3. Genes Databases**

**- Genecards:** Searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. The knowledgebase automatically integrates gene-centric data from ~150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information.

Access: <https://www.genecards.org/>

**4. Omics databases**

**- SRA (Sequene Read Archive):** largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information.

Access: <https://www.ncbi.nlm.nih.gov/sra>

**- GEO (Gene Expression Omnibus):** Public functional genomics data repository with array- and sequence-based data curated by NCBI.

Access: <https://www.ncbi.nlm.nih.gov/geo/>

**- Expression Atlas:** Gene expression datasets across species and biological conditions curated by European Bioinformatics Institute (EBI).

Access: <https://www.ebi.ac.uk/gxa/home>

**- Single Cell Expression Atlas:** Single Cell gene expression datasets curated by European Bioinformatics Institute (EBI).

Access: <https://www.ebi.ac.uk/gxa/sc/home>

**- EGA (European Genome-Phenome Archive):** a service for permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects or in the context of research-focused healthcare systems.

Access: <https://ega-archive.org/>

**- ChIP-Atlas:** An integrative, comprehensive database to explore public Epigenetic dataset, including ChIP-Seq, DNase-Seq, ATAC-Seq, and Bisulfite-Seq data.

Access: <https://chip-atlas.org/>

**- Signaling Pathways Project (SPP):** multi-omics knowledgemine based upon public, manually curated transcriptomic and cistromic (ChIP-Seq) datasets involving genetic and small molecule manipulations of cellular receptors, enzymes and transcription factors.

Access: <http://www.signalingpathways.org/index.jsf>

**- ProteomeXchange Consortium:** was established to provide globally coordinated standard data submission and dissemination pipelines involving the main proteomics repositories, and to encourage open data policies in the field.The current members of the Consortium are: [PRIDE](http://www.ebi.ac.uk/pride/archive/) (EMBL-EBI, Cambridge, UK), [PeptideAtlas](http://www.peptideatlas.org/) (ISB, Seattle, WA, USA) (both of them are the founding members), [MassIVE](http://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) (UCSD, San Diego, CA, USA), [jPOST](http://jpost.org/) (various institutions, Japan), [iProx](http://www.iprox.org/) (National Center for Protein Sciences, Beijing, China) and [Panorama Public](https://panoramaweb.org/project/home/begin.view) (University of Washington, Seattle, WA, USA).

Access: <http://www.proteomexchange.org/>

**5. Pathways and Ontologies**

**- QuickGO:** fast web-based browser of the Gene Ontology and Gene Ontology annotation data, providing GOs related to genes, terms or term IDs.

Access: <https://www.ebi.ac.uk/QuickGO/>

**- AmiGO2:** project to create the next official web-based set of tools for searching and browsing the Gene Ontology database, which consists of a controlled vocabulary of terms covering biological concepts, and a large number of genes or gene products whose attributes have been annotated using GO terms.

Access: <http://amigo.geneontology.org/amigo/landing>

**- KEGG (Kyoto Encyclopedia of Genes and Genomes):** Database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.

Access: <https://www.genome.jp/kegg/>

**- BioCyc:** collection of 20,023 Pathway/Genome Databases (PGDBs) for model eukaryotes and for thousands of microbes, plus software tools for exploring them. Some tools are under paid subscription.

Access: <https://biocyc.org/>

**- Reactome:** open-source relational database of signaling and metabolic molecules and their relations organized into biological pathways and processes. The core unit of the Reactome data model is the reaction. Entities (nucleic acids, proteins, complexes, vaccines, anti-cancer therapeutics and small molecules) participating in reactions form a network of biological interactions and are grouped into pathways.

Access: <https://reactome.org/>

**6. Other useful databases**

**- The Human Protein Atlas:** Swedish-based program initiated in 2003 with the aim to map all the human proteins in cells, tissues, and organs using an integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology.

Access: <https://www.proteinatlas.org/>

**- ENCODE (Encyclopedia of DNA Elements):** The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Access: <https://www.encodeproject.org/>

**- Allen Brain Map:** The Allen Institute for Brain Science uses a unique approach to generate data, tools and knowledge for researchers to explore the biological complexity of the mammalian brain. This portal provides access to high quality data and web-based applications created for the benefit of the global research community. The Allen Brain Atlases capture patterns of gene expression across the brain in various species. Allen Cell Types Database resources include exploring gene expression patterns across cell types, brain regions, and species; comparing cell types across transcriptomic, electrophysiological, and morphological profiles; defining cell types and taxonomies; and matching data to defined cell types.

Access: <https://portal.brain-map.org/>

**- EggNOG:** a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.

Access: <http://eggnog5.embl.de/#/app/home>

**ONLINE TOOLS**

**1. Alignment Tools**

Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Computational approaches to sequence alignment generally fall into two categories: global and local alignments.

 Usually, we use alignment tools providing them the target sequence(s) (nucleotides or aminoacids).

**1.1. Global alignments**

Global alignment tools create an end-to-end alignment of the sequences to be aligned. They can be pairwise or multiple sequence alignments (MSA).

**Pairwise:**

- **EMBOSS Needle:** creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

Access: <https://www.ebi.ac.uk/Tools/psa/emboss_needle/>

**- EMBOSS Stretcher:** uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

Access: <https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/>

**MSA:**

**- Clustal Omega:** MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Access: <https://www.ebi.ac.uk/Tools/msa/clustalo/>

**- EMBOSS Cons:** creates a consensus sequence from a protein or nucleotide multiple alignment.

Access: <https://www.ebi.ac.uk/Tools/msa/emboss_cons>

**- WebPRANK:** phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions.

Access: <https://www.ebi.ac.uk/goldman-srv/webprank/>

**1.2. Local alignments**

Local alignments identify regions of similarity within sequences that could be widely divergent overall.

**- BLAST (Basic Local Alignment Search Tool):** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Access: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

**- EMBOSS Water:** uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

Access: <https://www.ebi.ac.uk/Tools/psa/emboss_water>

**2. Ontology Tools**

An ontology is a formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of classes (or terms or concepts) with relations that operate between them. The ontology tools usually analyze a list of target genes, classifying them into ontology terms, and then perform analysis to determine the fold enrichment of each term and statistical significance inside the list.

**- Gene Ontology/PANTHER:** describes our knowledge of the biological domain with respect to three aspects: Molecular Function (describe activities that occur at the molecular level, such as “catalysis” or “transport”), Cellular Component (locations relative to cellular structures in which a gene product performs a function, either cellular compartments or stable macromolecular complexes of which they are parts) and Biological Process (larger processes, or ‘biological programs’ accomplished by multiple molecular activities).

Usage: provide a list of gene IDs, select the GO aspect and species.

Access: <http://geneontology.org/>

**- WebGestalt (WEB-based Gene SeT AnaLysis Toolkit):** functional gene enrichment analysis web tool, powered by the main databases and also supports three well-established and complementary methods for enrichment analysis, including Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA), and Network Topology-based Analysis (NTA).

Usage: Select the species, method of interest (ORA, GSEA and NTA), and a functional database. Provide a list of gene targets and select their type of ID.

Access: <http://www.webgestalt.org/>

**- KEGG Mapper:** Search tool searches various KEGG objects, including genes, KOs, EC numbers, metabolites and drugs, against KEGG pathway maps and other network entities.

Usage: provide a list of gene IDs (KEGG identifiers are highly recommended).

Access: <https://www.genome.jp/kegg/mapper/search.html>

**- Enrichr:** integrative web-based and mobile software application that includes new gene-set libraries, an alternative approach to rank enriched terms, and various interactive visualization approaches to display enrichment results using the JavaScript library, Data Driven Documents (D3).

Usage: provide a list of gene IDs. On the results page, the analysis is divided into different categories of enrichment. You can open a particular analysis by tapping on the name of the gene-set library, presenting a multitude of visualizations.

Access: <https://maayanlab.cloud/Enrichr/>

**- Metascape:** web-based portal designed to provide a comprehensive gene list annotation and analysis resource for experimental biologists. Metascape combines functional enrichment, interactome analysis, gene annotation, and membership search to leverage over 40 independent knowledgebases within one integrated portal.

Usage: provide a list of gene IDs.

Access: <https://metascape.org/gp/index.html#/main/step1>

**3. Orthology and annotation tools**

Orthologs are genes which evolved from a common ancestral gene by speciation that usually have retained a similar function in different species. This information is extremely useful for annotating and predicting the function and features of new sequences.

**- OrthoVenn2:** web platform for comparison and annotation of orthologous gene clusters among multiple species. OrthoVenn has an efficient and interactive graphic tool which provide a Venn diagram view for comparing multiple species protein sequences.

Usage: upload a list of protein sequences (FASTA aminoacids or .faa).

Access: <https://orthovenn2.bioinfotoolkits.net/home>

**- eggNOG-mapper:** tool for functional annotation of large sets of sequences based on fast orthology assignments using precomputed eggNOG v5.0clusters and phylogenies.

Usage: upload FASTA sequence(s) (.fa .faa .fna file formats) of a protein list, chromosome, genome, metagenome.

Access: <http://eggnog-mapper.embl.de/>

**4. Primer design**

Primer design tools offer automated ways to obtain primer sequence for all kinds of usage (conventional PCR, qPCR, cloning…) and primer quality control.

**- Primer3:** Pick primers from a DNA sequence.

Usage: Paste source sequence (5'->3', string of ACGTNacgtn -- other letters treated as N -- numbers and blanks ignored) in FASTA format. Set general conditions: primer size, Tm, GC%.

Access: <https://primer3.ut.ee/>

**- PrimerBLAST:** A tool for finding specific primers.

Usage: same as Primer3.

Access: <https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome>

**- IDT’s OligoAnalyzer:** Quality control for primers, assessing dimer and hairpin formation.

Usage: simply paste the primer sequence. It needs to create an IDT account, but is free of charge.

Access: <https://www.idtdna.com/calc/analyzer>

**5. NEB molecular biology helpful online tools**

New England Biolabs (NEB) provides a great assort of online tools to help molecular biologists in tasks such as PCR, ligation, digestion, cloning…

**- NEBcloner:** Select a traditional cloning workflow step (restriction digest, end modification, ligation, transformation, site directed mutagenesis) to determine recommended products and protocols.

Access: <https://nebcloner.neb.com/#!/>

**- Tm calculator:** estimate an appropriate annealing temperature when using NEB PCR products.

Access: <https://tmcalculator.neb.com/#!/main>

**- NEBioCalculator:** This tool will calculate the mass of insert required at several molar insert:vector ratios in the range needed for typical ligation reactions.

Access: <https://nebiocalculator.neb.com/#!/ligation>

**- NEBcutter:** Enter a DNA sequence to identify cut sites.

Access: <https://nc3.neb.com/NEBcutter/>

**- NEBuilder Assembly Tool:** design primers for a Gibson Assembly reaction, based on the entered fragment sequences and the polymerase being used for amplification.

Access: <https://nebuilder.neb.com/#!/>

**6. Structural biology tools**

Structural biology bioinformatics tools deals with problems related to macromolecule structures, as 3D protein structures prediction, protein-ligand and protein-protein interaction (docking).

**- RaptorX:** tertiary and contact prediction for protein sequences without close homologs in the Protein Data Bank (PDB). RaptorX predicts protein secondary and tertiary structures, contact and distance map, solvent accessibility, disordered regions, functional annotation and binding sites.

Usage: For structure prediction provide the aminoacid sequence; for property and complex contact prediction provide the aminoacid sequence or a structure file (.pdb); for structure alignment provide a structure file (.pdb).

Access: <http://raptorx.uchicago.edu/>

**- HHPred:** method for sequence database searching and structure prediction based on the pairwise comparison of profile hidden Markov models (HMMs).

Usage: provide a protein sequence (fasta) or multiple sequence alignment file in A3M/CLUSTAL/FASTA/STOCKHOLM format.

Access: <https://toolkit.tuebingen.mpg.de/tools/hhpred>

**- PatchDock:** Molecular docking algorithm based on shape complementarity principles.

Usage: provide to the program the receptor and ligand PDB ID or a file in the .pdb format.

Access: <https://bioinfo3d.cs.tau.ac.il/PatchDock/php.php>

**- PrankWeb:** it builds upon P2Rank a machine learning-based method for prediction of ligand binding sites from protein structure.

Usage: provide to the program the protein target PDB ID or a file in the .pdb format.

Access: <https://prankweb.cz/>

**- ConSurf:** server for the identification of functional regions in proteins.

Usage: provide to the program the protein target PDB ID; a file in the .pdb format, or FASTA sequence.

Access: <https://consurf.tau.ac.il/consurf_index.php>

**- ClusPro:** protein-protein docking in which a Fast Fourier Transform (FFT)-based rigid docking program provides 1000 low energy results to a clustering program to attempt to find the native site under the assumption that it will have a wide free-energy attractor with the largest number of results.

Usage: provide to the program the receptor and ligand PDB ID or a file in the .pdb format.

Access: <https://cluspro.org/home.php>

**7. Other useful tools**

**- WebLogo:** it is a web-based application designed to make the generation of sequence logos easy and painless. A sequence logo is a graphical representation of an amino acid or nucleic acid multiple sequence alignment. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

Usage: enter a multi-FASTA file or a list of FASTA sequences.

Access: <https://weblogo.threeplusone.com/create.cgi>

**- The MEME Suite:** Allows you to discover novel motifs in collections of unaligned nucleotide or protein sequences, and to perform a wide variety of other motif-based analyses. It provides motif discovery algorithms using both probabilistic (MEME) and discrete models (STREME), which have complementary strengths. It also allows discovery of motifs with arbitrary insertions and deletions (GLAM2). The MEME Suite provides three tools for motif enrichment analysis--measuring the enrichment of known motifs in sets of sequences. The motif enrichment may be anywhere in the sequences (SEA, AME), or concentrated in the central regions of the sequences (CentriMo). The Suite also provides an algorithm for measuring the similarity between motifs (Tomtom). These three types of analysis are combined in a pipeline in two MEME Suite tools--XSTREME and MEME-ChIP--that perform comprehensive motif analysis in general sequences (XSTREME) or in ChIP-seq peaks (MEME-ChIP). In addition to motif discovery, the MEME Suite provides tools for scanning sequences for matches to motifs (FIMO, MAST and GLAM2Scan), scanning for clusters of motifs (MCAST), finding preferred spacings between motifs (SpaMo), predicting the biological roles of motifs (GOMo), and predicting the regulatory targets of transcription factors.

Usage: the initial steps involve entering a collection of unaligned nucleotide or protein sequences.

Access: <https://meme-suite.org/meme/>

**- Ominer:** Discover relationships between signaling pathway nodes, gene targets and tissues. Utilizes cistromics data (ChIP-seq) to infer transcription factor regulation. Curated by SPP.

Usage: Select a single or list of target genes, select Cistromics as the Omics category, select species and physiological system of interest and submit.

Access: <http://www.signalingpathways.org/ominer/query.jsf>

**INTEGRATED BIOINFORMATICS ANALYSIS PLATFORM**

**- Galaxy:** scientific workflow, data integration, and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming or systems administration experience.

Access: <https://usegalaxy.org/>

**- ASAP (Automated Single-Cell Analysis Pipeline):** Collaborative web portal aimed at democratizing single-cell omics data analyses for researchers. The entire single-cell analysis pipeline is available in ASAP, allowing users to choose from a panel of tools and guiding them through tutorials.

Usage: Upload count / normalized gene expression matrix [10x(.h5 file), loom, text, zip, tar.gz, MEX/MTX formats supported]

Access: <https://asap.epfl.ch/projects/new>

**LEARNING PLATFORMS**

**For learning programming languages and coding**

**- Codecademy:**

Access: <https://www.codecademy.com/>

**- Sololearn:**

Access: <https://www.sololearn.com/home>

**- Datacamp:**

Access: <https://www.datacamp.com/>

**For learning bioinformatics skills**

**- Rosalind:**

Access: <https://rosalind.info/problems/locations/>