List of Common Bioinformatics Programs

# Detecting Sequence Similarity

|  |  |
| --- | --- |
| Program | Purpose |
| [NCBI BLAST](https://blast.ncbi.nlm.nih.gov/Blast.cgi) | Detect regions of local similarity between a query sequence and sequences in a database |
| [FlyBase BLAST](https://flybase.org/blast/) | Perform BLAST searches against the *D. melanogaster* genome assembly, annotated genes, and annotated proteins |
| [FASTA](https://www.ebi.ac.uk/jdispatcher/sss/fasta) | Suite of programs for performing global and local similarity searches |
| [BLAT](https://genome.ucsc.edu/cgi-bin/hgBlat) | Quickly generate alignments of query sequences against a genome assembly; BLAT is faster but less sensitive than BLAST |
| [Clustal Omega](https://www.ebi.ac.uk/jdispatcher/msa/clustalo) | Generate alignment for multiple protein or nucleotide sequences |
| [*Stretcher*](https://www.ebi.ac.uk/jdispatcher/psa/emboss_stretcher) | Use the Needleman and Wunsch algorithm to identify the optimal global alignment between two sequences |
| [*Matcher*](https://www.ebi.ac.uk/jdispatcher/psa/emboss_matcher) | Identify local regions of similarities between two input sequences using the Waterman-Eggert local alignment algorithm |
| [HMMER](https://www.ebi.ac.uk/jdispatcher/pfa/hmmer3_phmmer) | Use profile hidden Markov Models to detect sequence similarity between the query protein sequence and sequences in protein databases |
| [SmartBLAST](https://blast.ncbi.nlm.nih.gov/blast/smartblast/) | Quickly identify matches to protein sequences in the [landmark database](https://blast.ncbi.nlm.nih.gov/smartblast/smartBlast.cgi?CMD=Web&PAGE_TYPE=BlastDocs#searchSets) |

# Web Databases

|  |  |
| --- | --- |
| Program | Purpose |
| [FlyBase](https://flybase.org/) | Access the most recent set of *D. melanogaster* gene annotations |
| [ENCODE Portal](https://www.encodeproject.org/) | Access the *D. melanogaster* datasets produced by the ENCODE3 project. View these datasets on the UCSC Genome Browser. |
| [UCSC Genome Browser](https://genome.ucsc.edu/) | Access the genome assemblies and annotations generated by UCSC (e.g., human, chimpanzee) |
| [UCSC Genome Archive](https://hgdownload.soe.ucsc.edu/hubs/) | Access the genome browsers and gene annotations for more than 3,700 genome assemblies curated by NCBI and the Vertebrate Genomes Project (VGP). The UCSC Genome Archive (GenArk) includes UCSC Assembly Hubs for more than 600 [invertebrate genomes](https://hgdownload.soe.ucsc.edu/hubs/invertebrate/asmStats.html) from GenBank and the RefSeq database. |
| [Ensembl Metazoa](https://metazoa.ensembl.org/index.html) | Access the genome assemblies and annotations for different insects. Use the Ensembl interface to easily retrieve individual exon sequences. |

# Gene Predictors

|  |  |
| --- | --- |
| Program | Purpose |
| [Genscan](http://hollywood.mit.edu/GENSCAN.html) | An *ab initio* gene predictor (optimized for mammalian genomes) |
| [Augustus](https://bioinf.uni-greifswald.de/webaugustus/) | An *ab initio* and evidence-based gene predictor that supports many organisms. Uses extrinsic evidence such as sequence alignments and RNA-Seq data to improve gene predictions. |

# Repeat Finders

|  |  |
| --- | --- |
| Program | Purpose |
| [RepeatMasker](https://www.repeatmasker.org/) | Find interspersed repeats and low complexity DNA in the query sequence |
| [Dfam Sequence Search](https://www.dfam.org/search/sequence) | Find matches to known repeats in the [Dfam](https://www.dfam.org/home) database |

# Motif Finding

|  |  |
| --- | --- |
| Program | Purpose |
| [JASPAR](https://jaspar.elixir.no/) | This database contains the profiles of experimentally-confirmed transcription factor binding sites for many eukaryotes. For example, the [insecta section of JASPAR CORE](https://jaspar.elixir.no/search?q=&collection=CORE&tax_group=insects) includes the transcription factor binding sites for *D. melanogaster*. |
| [matrix-scan](http://rsat.sb-roscoff.fr/matrix-scan_form.cgi) | Search for matches to a motif in a nucleotide sequence using the Regulatory Sequence Analysis Tools ([RSAT](http://rsat.sb-roscoff.fr/index.php)) web server |
| [MEME Suite](https://meme-suite.org/meme/) | Suite of tools for *de novo* motif discovery and analyses |

# Sequence Analysis Tools

|  |  |
| --- | --- |
| Program | Purpose |
| [EMBOSS](http://emboss.toulouse.inra.fr/cgi-bin/emboss/) | Large collection of bioinformatics tools for manipulating and analyzing sequences (e.g., translation, extract subsequence) |
| [Galaxy](https://usegalaxy.org/) | Public instance of Galaxy — a platform for analyzing next generation sequencing data (e.g., ChIP-Seq, RNA-Seq) and sharing analysis results |
| [G-OnRamp](https://g-onramp.org/) | Galaxy tools and workflows which can be used to create UCSC Assembly Hubs and JBrowse/Apollo for genome annotation |
| [Apollo](https://genomearchitect.readthedocs.io/en/latest/) | A web-based collaborative genomic annotation editor |
| [JBrowse2](https://jbrowse.org/jb2/) | A web-based genome browser which supports linear, circular, dot plot, and synteny views of large genomic datasets for comparative genomics and variant analyses. |

# Protein Domains

|  |  |
| --- | --- |
| Program | Purpose |
| [InterPro](https://www.ebi.ac.uk/interpro/) | This database combines the protein signatures from [13 member databases](https://interpro-documentation.readthedocs.io/en/latest/interpro.html) (e.g., [NCBI CDD](https://www.ncbi.nlm.nih.gov/cdd), [Pfam](https://www.ebi.ac.uk/interpro/entry/pfam/#table), [PROSITE](https://prosite.expasy.org/), [SMART](http://smart.embl-heidelberg.de/)) to facilitate the identification of conserved domains within a protein sequence, and the classification of proteins into families. (See the [release notes](https://www.ebi.ac.uk/interpro/release_notes/) for the InterPro database statistics.) The [Sequence Search Box](https://www.ebi.ac.uk/interpro/search/sequence/) on the InterPro website allows users to compare a protein sequence against the protein signatures in the InterPro database with [InterProScan](https://interpro-documentation.readthedocs.io/en/latest/interproscan.html). |
| [CD-Search](https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) | This program compares a nucleotide or protein sequence against the Position-Specific Scoring Matrices (PSSMs) for the conserved domains in the NCBI Conserved Domains Database (CDD) with [RPS-BLAST](https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#RPSBWhat) |
| [HMMER web server](https://www.ebi.ac.uk/Tools/hmmer/) | This web server provides access to four programs in [HMMER](http://hmmer.org/) (phmmer, hmmscan, hmmsearch, and jackhmmer) that use profile hidden Markov models to identify remote homologs. For example, phmmer is designed to compare a protein sequence against the UniProtKB/Swiss-Prot database, and hmmscan can be used to compare a protein sequence against conserved domain databases such as Pfam. |
| [HHpred](https://toolkit.tuebingen.mpg.de/tools/hhpred) | This program uses pairwise comparisons of profile hidden Markov models (HMMs) to identify remote homologs and conserved protein domains. Pairwise comparisons of profile HMMs typically show higher sensitivity than the comparisons of two sequences (e.g., BLAST) and the comparison of a sequence against a profile HMM (e.g., HMMER). The alignment output produced by HHpred can be used with other tools in the [MPI Bioinformatics Toolkit](https://toolkit.tuebingen.mpg.de/) (e.g., perform structure prediction via comparative modelling with [MODELLER](https://toolkit.tuebingen.mpg.de/tools/modeller)). |