List of Common Bioinformatics Programs

Detecting Sequence Similarity

Program	Purpose
NCBI BLAST	Detect regions of local similarity between a query sequence and sequences in a database
FlyBase BLAST	Perform BLAST searches against the *D. melanogaster* genome assembly, annotated genes, and annotated proteins
FASTA	Suite of programs for performing global and local similarity searches
BLAT	Quickly generate alignments of query sequences against a genome assembly; BLAT is faster but less sensitive than BLAST
Clustal Omega	Generate alignment for multiple protein or nucleotide sequences
Stretcher	Use the Needleman and Wunsch algorithm to identify the optimal global alignment between two sequences
Matcher	Identify local regions of similarities between two input sequences using the Waterman-Eggert local alignment algorithm
HMMER	Use profile hidden Markov Models to detect sequence similarity between the query protein sequence and sequences in protein databases
SmartBLAST	Quickly identify matches to protein sequences in the landmark database

Web Databases

Program	Purpose
FlyBase	Access the most recent set of *D. melanogaster* gene annotations
ENCODE Portal	Access the *D. melanogaster* datasets produced by the ENCODE3 project. View these datasets on the UCSC Genome Browser.
UCSC Genome Browser	Access the genome assemblies and annotations generated by UCSC (e.g., human, chimpanzee)
UCSC Genome Archive	Access the genome browsers and gene annotations for more than 3,200 genome assemblies curated by NCBI and the Vertebrate Genomes Project (VGP). The UCSC Genome Archive (GenArk) includes UCSC Assembly Hubs for more than 500 invertebrate genomes from GenBank and the RefSeq database.
Ensembl Metazoa	Access the genome assemblies and annotations for different insects. Use the Ensembl interface to easily retrieve individual exon sequences.

Gene Predictors

Program	Purpose
Genscan	An *ab initio* gene predictor (optimized for mammalian genomes)
Augustus	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	Find interspersed repeats and low complexity DNA in the query sequence
Dfam Sequence Search	Find matches to known repeats in the Dfam database

Motif Finding

Program	Purpose
JASPAR	This database contains the profiles of experimentally-confirmed transcription factor binding sites for many eukaryotes. For example, the insecta section of JASPAR CORE includes the transcription factor binding sites for *D. melanogaster*.
matrix-scan	Search for matches to a motif in a nucleotide sequence using the Regulatory Sequence Analysis Tools (RSAT) web server
MEME Suite	Suite of tools for *de novo* motif discovery and analyses

Sequence Analysis Tools

Program	Purpose
EMBOSS	Large collection of bioinformatics tools for manipulating and analyzing sequences (e.g., translation, extract subsequence)
Galaxy	Public instance of Galaxy — a platform for analyzing next generation sequencing data (e.g., ChIP-Seq, RNA-Seq) and sharing analysis results
G-OnRamp	Galaxy tools and workflows which can be used to create UCSC Assembly Hubs and JBrowse/Apollo for genome annotation
Apollo	A web-based collaborative genomic annotation editor
JBrowse2	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	This database combines the protein signatures from 13 member databases (e.g., NCBI CDD, Pfam, PROSITE, SMART) to facilitate the identification of conserved domains within a protein sequence, and the classification of proteins into families. (See the release notes for the InterPro database statistics.) The Sequence Search Box on the InterPro website allows users to compare a protein sequence against the protein signatures in the InterPro database with InterProScan.
CD-Search	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
HMMER web server	This web server provides access to four programs in HMMER (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	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 (e.g., perform structure prediction via comparative modelling with MODELLER).