Genome Annotation Tools

This section presents information on tools used for genome annotation, sequence analysis, and sites for data retrieval.

Appearance on this page does not imply endorsement by TAIR.

Gene Structural Annotation Tools

Repeat Masker

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked.

Codon Usage Database (Kazusa)

Codon usage tables for many organisms, including Arabidopsis thaliana, from the Kazusa Institute.

GENEMARK

Family of gene prediction programs provided by the Bioinformatics Group at the Georgia Institute of Technology.

GenScan

MIT's webserver for GenScan. GenScan is used to predict the location and intron/exon boundaries in a genomic sequence. Select Arabidopsis as the organism of choice for finding Arabidopsis genes in a genomic sequence.

NetGene2.2

Predictions of Arabidopsis splice sites from DTU.

NetStart1.0

Prediction software for Arabidopsis translation starts from DTU.

Software Downloads

Generic Model Organism Database (GMOD) GitHub

Everything you need to set up a MOD and annotate a genome- all open source software.

BioConductor

Open source software downloads and open development environment for bioinformatics software.

Comprehensive Sequence Analysis Resources

EBI Services

List of bioinformatic tools and resources

GenePalette

A cross-platform and cross-species desktop application for genome sequence visualization and navigation.

Comparative Resources

Inparanoid

A collection of pairwise comparisons between 640 eukaryotic whole genomes including Arabidopsis thaliana, useful for the identification of orthologs and differentiation between inparalogs and outparalogs.

Phytozome

Contains comparisons of many plant species.

PLAZA

Access point for plant comparative genomics centralizing genomic data produced by different genome sequencing initiatives. PLAZA integrates plant sequence data and comparative genomics methods and provides an online platform allowing to perform evolutionary analyses and data mining within the green plant lineage (Viridiplantae).

CoGe

A comparative genomics platform designed to allow easy access to genomic data from any organism and provide analysis tools for finding and comparing homologous sequences from multiple genomic regions.

Positional history of A. thaliana genes

Archived data set showing the chromosomal positional histories of Arabidopsis genes. This dataset accompanied the paper Woodhouse MR, Tang H, Freeling M (2011) Different gene families in Arabidopsis thaliana transposed in different epochs and at different frequencies throughout the rosids. The Plant Cell 23(12): 4241-4253. http://dx.doi.org/10.1105/tpc.111.093567

Plant Promoter and Regulatory Element Resources

AGRIS

Currently contains two databases, AtcisDB (Arabidopsis thaliana cis-regulatory database) and AtTFDB (Arabidopsis thaliana transcription factor database).

AthaMap

A genome-wide map of putative transcription factor binding sites in *Arabidopsis thaliana*. Because of the PI's retirement, this database may be switched of any time after July 1st 2023.

ATCOECIS

This resource can be used to query co-expression data, GO and cis-regulatory elements annotations, submit user-defined gene sets for motif analysis for Arabidopsis and provides an access point to unravel the regulatory code underlying transcriptional control in Arabidopsis. (non-https server)

PlantCare

Database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences.

PlantTFDB: Plant Transcription Factor Database

An integrative plant transcription factor database that provides a web interface to access large (close to complete) sets of transcription factors of a large number of plant species.

ppdb (Plant Promoter DB)

Database that provides transcription start sites (TSS) and other structural information for *Arabidopsis thaliana*, *Oryza sativa*, and *Physcomitrella patens* and poplar promoters.

Transfac

Database on eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles. Commercial site.

Proteome Resources

Links to proteome analysis tools and repositories.

Database Searches

Nucleotide and Protein Databases

Entrez

NCBI's Entrez Databases -retrieve sequences and other data, including literature, from PubMed.

UniProt

UniProt reflects a merge of 3 databases, SwissProt,PIR and TrEMBL and replaces these databases. Search UniProtKB, a database of curated protein sequences (formerly Swiss-Prot).

PDB

Protein Data Bank, the repository for the processing and distribution of 3-D macromolecular structure data.

miRBase

micro-RNA database for micro-RNA sequences from more than 270 species, including A. thaliana.

RNA Central

The non-coding RNA sequence database, a comprehensive ncRNA sequence collection representing all ncRNA types from a broad range of organisms .

BLAST servers

TAIR BLAST

Search against all public Arabidopsis sequences, several subsets of them, or all higher plant sequences from GenBank. These datasets can be downloaded.

NCBI BLAST

BLAST server at NCBI

BLAST help

BLAST manual and user guide from NCBI