**USER GUIDE FOR DENOVO GENOMICS PIPELINE**

**Contents**

1. Introduction
2. Installation
3. Genome Information file
4. Configuration file
5. SQL Database Structure
6. Running Denovo Pipeline
7. Output Files
8. Exploring output using GUI

**Introduction**

Denovo Genomic Analysis Pipeline (DeNoGAP) is a software package for comparative analysis of multiple completed or draft genomes. The pipeline incorporates number of tools and databases for gene prediction, homolog prediction, ortholog prediction, functional annotation, phylogenetic profiling and core genome prediction.

**Installation**

The package for DeNoGAP is available at the github site XX.

**Directory structure of DeNoGAP package:**

**bin**: contain main-pipeline execution script, installation script, and other analysis scripts.

**config:** contain configuration files for defining parameters for different analysis phases.

**lib:**  contains DeNoGAP-specific Perl modules required for the analysis.

**data**: directory to store input data files.

**output:** directory to store output data / results.

**doc:** contains manual for setting up and using DeNoGAP.

**SYSTEM REQUIREMENT**

**Operating system: Linux**

**RAM: at least 2GB**

The package provides a script (install.pl) to install necessary programs and Perl modules for performing analysis with DeNoGAP on Linux platform. Currently DeNoGAP has been only tested on Ubuntu Linux system. Although not tested, we assume that DeNoGAP can also run under other Unix based OS after installation of necessary programs and prerequisites. Windows OS user can try using ‘cygwin’ to run DeNoGAP. We may release system-specific version in near future.

List of required Perl modules / programs:

**Perl modules:**

FindBin, Env, Exporter, Getopt::Long, File::Basename, File::Copy, Tie::File Parallel::ForkManager, List::MoreUtils, List::Util, File::Path, Hash::Merge, DBI, CGI, English, File::Spec::Functions, FileHandle, IO::Scalar, IO::String, Mail::Send, Sys::Hostname, URI::Escape, XML::Parser, XML::Quote

**BioPerl modules:**

Bio::Perl, Bio::SeqIO, Bio::Seq, Bio::SearchIO, Bio::Tools::Phylo::Phylip::ProtDist, Bio::AlignIO

**Programs:**

Muscle v3.8.31 or above (http://www.drive5.com/muscle)

Kalign2 (http://msa.sbc.su.se/downloads/kalign)

MCL (http://micans.org/mcl)

Hmmer version 3 or above (http://selab.janelia.org/software/hmmer3)

Phylip v3.6 or above (http://evolution.gs.washington.edu/phylip)

Glimmer (<http://ccb.jhu.edu/software/glimmer>)

Prodigal (<http://prodigal.googlecode.com>)

FragScan (<http://omics.informatics.indiana.edu/mg/get.php?software=FragGeneScan1.16.tar.gz>)

GeneMark (<http://opal.biology.gatech.edu>)

InterProScan5 (<https://code.google.com/p/interproscan>)

EMBOSS (<http://emboss.sourceforge.net>)

SQLite (<https://sqlite.org>)

Apache (<http://www.apache.org>)

Users can either manually install each of the programs or run “install.pl” script under DeNoGAP for installation.

DeNoGAP do not install Apache. Users need to manually install apache server and configure it for running GUI for DeNoGAP database exploration

To install required programs, execute following commands in terminal:

**cd DeNoGAP**

**cd bin**

**perl install.pl <install\_directory>**

Note: By default installation of most of the external programs will take place under “exe” directory of DeNoGAP package. However, some programs are installed under root directory by default, which may need appropriate permissions from user for installation.

**Input Data Files**

DeNoGAP requires three input files to perform any analysis.

(1) The tab-delimited file containing metadata information about the genomes used for the analysis.

(2) The configuration file containing defined parameters for the analysis.

(3) SQLite Database file to store output.

The format and description for each input file is given below.

**(1) Genome Information File**

The first line of the genome information file should start with “#” followed by tab-delimited column names.

* **Mandatory Column names**
* **genome\_name** : Full genome name.
* **species**: Full name of the species**.**
* **abbreviation**: Short abbreviation for the genome. This will be used to identify and name all sequence files and output files. (Abbreviation should not have any dots or special characters except “\_”).
* **genome\_type:** Indicate if genome is a reference or query. (Acceptable values: reference / query).
* **outgroup:** Indicate if genome is an outgroup or not. (Acceptable values: Yes / No).
* **Optional Columns**
* Users can create any number of new columns to add any additional information to the genome information table.
* The name of additional columns should be in lower case without any special characters except “\_”.

The example table is given in the data directory of DeNoGAP package.

**(2) Configuration Files**

DeNoGAP uses separate configuration file for each analysis phase. All parameters, file paths, and directory paths required for performing the analysis should be defined in respective configuration file. Parameters are divided into different sections named within [] brackets. The description of each configuration file and parameters included in it is given below:

* **PARSE\_GENBANK.config**

This configuration file defines parameters for extracting sequences and genomic information from the GenBank Files.

* **PARSE\_GENBANK**: Initiates parsing of genebank files (Default value: YES).
* **GENBANK\_DIR\_PATH**: Define directory path for genebank files.
* **PROJECT\_DIR\_NAME**: Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **GENOME\_DIR\_NAME:** Sub-directory name to store fasta formatted genome sequence files.
* **CDS\_DIR\_NAME:** Sub-directory name to store fasta formatted coding sequence files.
* **PROTEIN\_DIR\_NAME:** Sub-directory name to store fasta formatted protein sequence files.
* **FEATURE\_DIR\_NAME:** Sub-directory name to store tab-delimited genomic feature files.
* **PREDICT\_GENE.config**

This configuration file defines parameters to predict genes from the genome sequences using four gene prediction programs.

* **PREDICT\_GENE:** Initiate gene prediction analysis. (Default value: YES).
* **GENOME\_DIR\_PATH:** Define directory path for fasta-formatted genome sequence files.
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **GILMMER\_RESULT\_DIR\_NAME:** Sub-directory name to store glimmer output files.
* **GENEMARK\_RESULT\_DIR\_NAME:** Sub-directory name to store GeneMark output files.
* **PRODIGAL\_RESULT\_DIR\_NAME:** Sub-directory name to store prodigal output files.
* **FRAGSCAN\_RESULT\_DIR\_NAME:** Sub-directory name to store fragscan output files.
* **CDS\_DIR\_NAME:** Sub-directory to store predicted coding sequence files.
* **PROTEIN\_DIR\_NAME:** Sub-directory to store translated protein sequence files.
* **FEATURE\_DIR\_NAME:** Sub-directory to store tab-delimited genomic feature files.
* **TRANSLATION\_CODE:** Genebank codon table for translating coding sequences into proteins.
* **OVERLAP\_BASE:** Number of overlapping bases allowed between adjacent genes.
* **PARALLEL\_CPU\_CORE:** Number of CPU core to be used for parallel processing.
* **GLIMMER3:** Define options for running glimmer3 program. Check available options from glimmer manual. All options should be defined within “ ”.
* **LONG\_ORF:** Define options for running long-orfs program. Check available options from glimmer manual. All options should be defined within “ ”.
* **MULTI\_EXTRACT:** Define options for running multi-extract program. Check available options from glimmer manual. All options should be defined within “ ”.
* **BUILD\_ICM:** Define options for running build-icm program. Check available options from glimmer manual. All options should be defined within “ ”.
* **GMSN:** Define options for running GeneMark program. Check available options from GeneMark manual. DeNoGAP automatically takes value for “-- name” and “--species” options from the genome table. All other options should be defined here within “ ”.
* **PRODIGAL:** Define options for running Prodigal program. Check available options from prodigal help. DeNoGAP automatically takes value for -i , -t, -o, -a, -d , -s from the genome table. All other options should be defined here within “ ”.
* **FRAGSCAN:** Define options for running FragGeneScan program. Check available options from FragGeneScan help. DeNoGAP automatically takes value for “-genome” and “-out” options from the genome table. All other options should be defined here within “ ”.
* **GENE\_VERIFICATION.config**

This configuration file defines parameters to verify and annotate predicted protein sequences by comparing sequence match within Uniprot database.

* **VERIFY\_SEQUENCE:** Initiate verification of predicted protein sequences using Uniprot database. (Default Value: YES).
* **BLAST\_ALIGNMENT\_FILE:** Pairwise blast alignment result between protein sequences and UniPort database.
* **FEATURE\_DIR:** Define full name of the directory including complete path containing genomic feature files.
* **CDS\_DIR:** Define full name of the directory including complete path containing coding sequence files.
* **PROTEIN\_DIR:** Define full name of the directory including complete path containing protein sequence files.
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **EVALUE\_THRESHOLD:** Define minimum e-value cut-off for significant hits.
* **ALIGNMENT\_IDENTITY:** Define minimum sequence identity for significant hits.
* **QUERY\_COVERAGE:** Define minimum query coverage for significant match.
* **MIN\_PROTEIN\_LENGTH:** Define minimum protein sequence length cutoff to discard insignificant sequences.
* **LOAD\_DATA.config**

This configuration file defines parameters to load sequences and genomic feature information in to the SQLite database.

* **LOAD\_DATA:**  Initiate module for loading sequences and genomic data. (Default: YES).
* **FEATURE\_DIR:** Define full name of the directory including complete path containing genomic feature files.
* **CDS\_DIR:** Define full name of the directory including complete path containing coding sequence files.
* **PROTEIN\_DIR:** Define full name of the directory including complete path containing protein sequence files.
* **ADJUST\_HEADER:** Default value: YES. Adjust sequence identifier and format it as “genome\_abbreviation|sequence\_identifier”.
* **COMPARE\_REFERENCE.config**

This configuration file defines parameter for pairwise sequence comparison between reference genomes using Phmmer program.

* **COMPARE\_REFERENCE:** (Default value: YES). Initiates pairwise sequence comparison between reference genomes defined by user in genome table.
* **MODEL\_DB:** Define name for the database file to be created for Hidden Markov models of the protein families. (Default value: HMM\_MODEL\_DB).
* **SEQ\_DB:** Define name for the database file to be created for Singleton protein family sequences. (Default value: HMM\_SEQ\_DB).
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **HMMER\_OPT:** Define options for running phmmer program. Check available options for phmmer from hmmer package manual. DeNoGAP automatically takes value for “-o” and “-domtblout” options from the genome table. All other options should be defined here within “ ”.
* **MAX\_NUM\_DOMAIN:** Define maximum number of hmmer domains allowed between matched sequences. (Default value: 5).
* **ACCURACY\_THRESHOLD:** Define hmmer accuracy probability cutoff for significant match. The value range is between [0 - 1]. (Default value: 0.8)
* **IDENTITY:** Define percentage identity cutoff for significant match. (Default value: 70).
* **SIMILARITY:** Define percentage similarity cutoff for significant match. (Default value: 60).
* **QUERY\_COVERAGE:** Define percentage cutoff for query sequence covered in a significant match. (Default value: 70).
* **HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a significant match. (Default value: 70).
* **MIN\_CHIMERA\_IDENTITY:** Define percentage identity cutoff for predicting chimera-like match. (Default value: 70)
* **MIN\_CHIMERA\_SIMILARITY:** Define percentage similarity cutoff for predicting chimera-like match. (Default value: 60)
* **MIN\_CHIMERA\_QUERY\_COVERAGE:**  Define percentage cutoff for query sequence covered in a chimera match. (Default value: 25).
* **MIN\_CHIMERA\_HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a chimera match. (Default value: 25).
* **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).
* **CLUSTER\_INDEX:** Define index value for naming the hmm family cluster file. (Default value: 1).
* **PREDICT\_HMM\_FAMILY.config**

This configuration files define parameters for iterative prediction of protein families in additional genomes.

* **PREDICT\_HMM:** Initiate iterative comparison of protein sequences from new genomes. (Default value: YES).
* **MODEL\_DB:** Define name for the database file to be created for Hidden Markov models of the protein families. (Default value: HMM\_MODEL\_DB).
* **SEQ\_DB:** Define name for the database file to be created for Singleton protein family sequences. (Default value: HMM\_SEQ\_DB).
* **HMM\_CLUSTER\_FILE:** Define complete path and file name of the seed family cluster file.
* **MODEL\_DB\_FILE:** Define complete path and file name of the seed HMM model database file.
* **SINGLETON\_DB\_FILE:** Define complete path and file name of the seed Singleton sequence database file.
* **HMM\_FAMILY\_OUTFILE:** Define complete path and file name for the final HMM model database output file.
* **SUPER\_HOMOLOG\_OUTFILE:** Define complete path and file name for the super-homolog cluster output file.
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **HMMER\_OPT:** Define options for running hmmscan and phmmer program. Check available options for hmmscan and phmmer from hmmer package manual. DeNoGAP automatically takes value for “-o” and “-domtblout” options from the genome table. All other options should be defined here within “ ”.
* **MAX\_NUM\_DOMAIN:** Define maximum number of hmmer domains allowed between matched sequences. (Default value: 5).
* **ACCURACY\_THRESHOLD:** Define hmmer accuracy probability cutoff for significant match. The value range is between [0 - 1]. (Default value: 0.8)
* **IDENTITY:** Define percentage identity cutoff for significant match. (Default value: 70).
* **SIMILARITY:** Define percentage similarity cutoff for significant match. (Default value: 60).
* **QUERY\_COVERAGE:** Define percentage cutoff for query sequence covered in a significant match. (Default value: 70).
* **HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a significant match. (Default value: 70).
* **MIN\_CHIMERA\_IDENTITY:** Define percentage identity cutoff for predicting chimera-like match. (Default value: 70)
* **MIN\_CHIMERA\_SIMILARITY:** Define percentage similarity cutoff for predicting chimera-like match. (Default value: 60)
* **MIN\_CHIMERA\_QUERY\_COVERAGE:**  Define percentage cutoff for query sequence covered in a chimera match. (Default value: 25).
* **MIN\_CHIMERA\_HMM\_COVERAGE:** Define percentage cutoff for hmm model sequence covered in a chimera match. (Default value: 25).
* **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).
* **CLUSTER\_INDEX:** Define index value for naming the hmm family cluster file. (Default value: 2).
* **PREDICT\_SUPER\_HOMOLOG.config**

This configuration file defines parameters for predicting super-homolog family and identifying links between partial gene families and full-length gene families.

* **PREDICT\_SUPER\_HOMOLOG:** Initiate prediction of super-homolog families (Default value: YES).
* **HMM\_CLUSTER\_FILE:** Define complete path and file name of the hmm-family cluster file.
* **IDENTITY\_THRESHOLD:** Define percentage identity cutoff for significant partial match. (Default value: 70).
* **SIMILARITY\_THRESHOLD:** Define percentage similarity cutoff for significant partial match. (Default value: 60).
* **QUERY\_COVERAGE\_THRESHOLD:** Define percentage cutoff for query sequence covered in a significant partial match. (Default value: 70).
* **SUBJECT\_COVERAGE\_THRESHOLD:** Define percentage cutoff for subject sequence covered in a significant partial match. (Default value: 70).
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **PREDICT\_ORTHOLOG.config**

This configuration file defines parameters for predicting ortholog and inparalog protein pairs and cluster ortholog families.

* **PREDICT\_ORTHOLOG**: Initiate prediction of ortholog and in paralog pairs. (Default value: YES).
* **CLUSTER\_ORTHOLOG**: Initiate clustering of ortholog and inparalog pairs. (Default value: YES).
* **HMM\_CLUSTER\_FILE:** Define complete path and file name of the hmm-family cluster file.
* **HOMOLOG\_CLUSTER\_FILE:** Define complete path and file name of super-homolog cluster file.
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **ORTHOLOG\_DIVERGENCE\_THRESHOLD:** Define distance cut-off for predicting ortholog pairs in case out-group is absent. (Default value: 0.8).
* **INPARALOG\_DIVERGENCE\_THRESHOLD:** Define distance cut-off for predicting inparalog pairs in case out-group is absent. (Default value: 0.5).
* **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).
* **PHYLOGENETIC\_PROFILE.config**

This configuration file generates a phylogenetic profile matrix to represent presence or absence of protein families across genomes.

* **PHYLOGENETIC\_PROFILE:** Initiate analysis for making binary Phylogenetic profile. (Default value: YES).
* **CLUSTER\_FILE:** Define complete path and name of the protein family cluster file.
* **GROUP\_PROFILE:** Define complete path and name of the output profile file.
* **GROUP\_PROTEIN\_TAB:** Define complete path and name of the tab-delimited list of protein sequence identifiers sorted by protein families.
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **CORE\_GENOME.config**

This configuration file defines parameters for predicting core protein families and create concatenated core-genome alignment.

* **CORE\_GENOME:** Initiate analysis for predicting core-genome. (Default value: YES).
* **CLUSTER\_FILE:** Define complete path and name of the protein family cluster file.
* **CORE\_ALIGNMENT\_FILE:** Define complete path and name of the concatenated core alignment file.
* **CORE\_THRESHOLD:** Define minimum percentage of required genome for predicting core-genome.
* **SEQUENCE\_TYPE:** Define sequence type (Options: nucleotide / protein). (Default value: protein).
* **INCLUDE\_OUTGROUP:** Include out-group sequences in core genome alignment. (Default value: NO).
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **ANNOTATION.config**

This configuration files defines parameters for predicting functional annotation for protein sequences using InterProScan.

* **PREDICT\_ANNOTATION:** Initiates annotation of protein sequences. (Default value: YES).
* **INTERPRO\_SCAN\_PATH:** Define source directory path for interproscan databases and files.
* **INTERPRO\_SCAN\_OPTS:** Define options for running interproscan analysis. Check available options from InterProscan help. DeNoGAP automatically takes value for “-i”, “-f” and “-o” options from the genome table. All other options should be defined here within “ ”.
* **PROJECT\_DIR\_NAME:** Define name of the project directory. If not present, a new directory will be created with project name under main output directory. All result files and sub-directories will be created under project directory.
* **PARALLEL\_CPU\_CORE:** Define number of CPU cores to be used for the analysis. (Default value: 1).

**(3) SQLite Database Schema**

DeNoGAP uses SQLite database to store analyzed information. DeNoGAP creates 19 database tables to store results from various analysis phases. The description of each table and its column is given below:

|  |  |  |
| --- | --- | --- |
| **Table: OrganismInfo** | | |
| **Column** | **Data type** | **Description** |
| genome\_name | TEXT | Full name of the genome |
| species | TEXT | Full name of the species |
| genome\_abbreviation | TEXT | short name for the genome |
| genome\_type | TEXT | Is Reference or Query |
| outgroup | TEXT | Is outgroup or not |

|  |  |  |
| --- | --- | --- |
| **Table: GeneFeature** | | |
| **Column** | **Data type** | **Description** |
| feature\_id | TEXT | Unique sequence identifier for the gene |
| feature\_type | TEXT | By default: CDS |
| genome\_id | TEXT | Unique sequence identifier for the genome sequence |
| genome\_type | TEXT | Chromosome / Plasmid / Contig |
| genome\_name | TEXT | Full name of the genome |
| genome\_length | INT | Length of the genome sequence |
| feature\_start | INT | Start co-ordinate of the gene sequence |
| feature\_end | INT | End co-ordinate of the gene sequence |
| nuc\_length | INT | Length of the coding sequence |
| aa\_length | INT | Length of the protein sequence |
| strand | INT | Genome strand on which gene is located ( + or -) |
| index\_on\_genome | INT | Order on the genome sequence |
| description | TEXT | Product description |

|  |  |  |
| --- | --- | --- |
| **Table: ProteinSequence** | | |
| **Column** | **Data type** | **Description** |
| pseq\_index\_id | INT | Auto-incremented primary key index |
| protein\_id | TEXT | Unique sequence identifier for protein |
| genome\_abbreviation | TEXT | Short name for the genome |
| seq\_type | TEXT | Protein |
| seq\_length | INT | Length of amino acid sequence |
| aa\_sequence | TEXT | Protein sequence |

|  |  |  |
| --- | --- | --- |
| **Table: NucleotideSequence** | | |
| **Column** | **Data type** | **Description** |
| nseq\_index\_id | INT | Auto-incremented primary key index |
| nucleotide\_id | TEXT | Unique sequence identifier for CDS |
| genome\_abbreviation | TEXT | Short name for the genome |
| seq\_type | TEXT | Protein |
| seq\_length | INT | Length of coding sequence sequence |
| nuc\_sequence | TEXT | CDS sequence |

|  |  |  |
| --- | --- | --- |
| **Table: Similarity** | | |
| **Column** | **Data type** | **Description** |
| query\_id | TEXT | Sequence identifier for the query protein |
| subject\_id | TEXT | Identifier for the target sequence or target hmm group |
| query\_length | INT | Length of query sequence |
| subject\_length | INT | Length of target sequence or target hmm model |
| num\_total\_domain | INT | Total domains predicted in the query sequence |
| num\_significant\_domain | INT | Number of domains with significant match |
| query\_start | INT | Start position of query sequence |
| query\_end | INT | End position of query sequence |
| subject\_start | INT | Start position of the target |
| subject\_end | INT | End position of the target |
| evalue | REAL | Significance value |
| bit\_score | INT | Bit score of the alignment |
| percent\_identity | REAL | Percentage identity between sequences |
| percent\_similarity | REAL | Percentage similarity between sequences |
| query\_coverage | REAL | Percentage query sequence coverage |
| subject\_coverage | REAL | Percentage target sequence coverage |
| pair\_relation | TEXT | Match type (best / truncated / chimera / insignificant) |

|  |  |  |
| --- | --- | --- |
| **Table: LinkFamily** | | |
| **Column** | **Data type** | **Description** |
| family\_idA | TEXT | Group id of hmm family |
| family\_idB | TEXT | Group id of hmm family |
| significance | REAL | Significance value between pair |

|  |  |  |
| --- | --- | --- |
| **Table: GenetoSuperFamily** | | |
| **Column** | **Data type** | **Description** |
| gene\_superfamily\_index\_id | INT | Auto increment primary key index |
| gene\_id | TEXT | Unique sequence identifier |
| genome\_name | TEXT | Genome abbreviation |
| hmm\_family\_id | TEXT | Hmm family identifier |
| super\_family\_id | TEXT | Super-homolog family identifier |

|  |  |  |
| --- | --- | --- |
| **Table: MultipleAlignment** | | |
| **Column** | **Data type** | **Description** |
| seq\_id | TEXT | Unique sequence identifier |
| genome\_abbreviation | TEXT | Genome abbreviation |
| seq\_type | TEXT | Genome abbreviation |
| alignment\_id | TEXT | Super-homolog family identifier |
| alignment\_length | INT | Length of super-homolog alignment |
| alignment\_sequence | TEXT | Aligned sequence |

|  |  |  |
| --- | --- | --- |
| **Table: DistancePair** | | |
| **Column** | **Data type** | **Description** |
| taxonA | TEXT | Genome abbreviation for SeqA |
| idA | TEXT | Unique sequence identifier SeqA |
| taxonB | TEXT | Genome abbreviation for SeqB |
| idB | TEXT | Unique sequence identifier SeqB |
| divergence | REAL | Pairwise distance between SeqA and SeqB |
| homolog\_cluster\_id | TEXT | Super-homolog family identifier |

|  |  |  |
| --- | --- | --- |
| **Table: OrthologPair** | | |
| **Column** | **Data type** | **Description** |
| taxonA | TEXT | Genome abbreviation for SeqA |
| idA | TEXT | Unique sequence identifier SeqA |
| taxonB | TEXT | Genome abbreviation for SeqB |
| idB | TEXT | Unique sequence identifier SeqB |
| divergence | REAL | Pairwise distance between SeqA and SeqB |
| homolog\_cluster\_id | TEXT | Super-homolog family identifier |

|  |  |  |
| --- | --- | --- |
| **Table: InParalogPair** | | |
| **Column** | **Data type** | **Description** |
| taxonA | TEXT | Genome abbreviation for SeqA |
| idA | TEXT | Unique sequence identifier SeqA |
| taxonB | TEXT | Genome abbreviation for SeqB |
| idB | TEXT | Unique sequence identifier SeqB |
| divergence | REAL | Pairwise distance between SeqA and SeqB |
| min\_ortholog\_divergence | REAL | Minimum pairwise distance between SeqA and ortholog from any other genome. |
| homolog\_cluster\_id | TEXT | Super-homolog family identifier |

|  |  |  |
| --- | --- | --- |
| **Table: DomainAnnotation** | | |
| **Column** | **Data type** | **Description** |
| protein\_id | TEXT | Unique sequence identifier |
| genome\_name | TEXT | Genome abbreviation |
| seq\_len | INT | Length of query sequence |
| domain\_id | TEXT | Unique identifier for the predicted domain |
| domain\_name | TEXT | Name of the predicted domain |
| domain\_start | INT | Start position of the domain |
| domain\_end | INT | End position of the domain |
| significance\_value | REAL | Significance of the domain match |
| description | TEXT | Domain description |

|  |  |  |
| --- | --- | --- |
| **Table: InterProAnnotation** | | |
| **Column** | **Data type** | **Description** |
| protein\_id | TEXT | Unique sequence identifier |
| genome\_name | TEXT | Genome abbreviation |
| interpro\_id | TEXT | Unique identifier for the interpro domain |
| interpro\_name | TEXT | Name of the predicted interpro domain |

|  |  |  |
| --- | --- | --- |
| **Table: GOAnnotation** | | |
| **Column** | **Data type** | **Description** |
| protein\_id | TEXT | Unique sequence identifier |
| genome\_name | TEXT | Genome abbreviation |
| go\_id | TEXT | Unique identifier for the gene ontology term |
| go\_category | TEXT | Classification category for go term |
| go\_description | TEXT | Description of the go term |

|  |  |  |
| --- | --- | --- |
| **Table: PathwayAnnotation** | | |
| **Column** | **Data type** | **Description** |
| protein\_id | TEXT | Unique sequence identifier |
| genome\_name | TEXT | Genome abbreviation |
| pathway\_id | TEXT | Unique identifier for the predicted pathway |
| pathway\_name | TEXT | Name of the predicted pathway |

|  |  |  |
| --- | --- | --- |
| **Table: PhylogeneticProfile** | | |
| **Column** | **Data type** | **Description** |
| id | TEXT | Ortholog family id |
| genome\_name | TEXT | Genome abbreviation |

|  |  |  |
| --- | --- | --- |
| **Table: MapGeneIdtoGeneFamily** | | |
| **Column** | **Data type** | **Description** |
| familymap\_index\_id | INT | Auto-incremented primary key index |
| genefamily\_id | TEXT | Unique identifier for predicted ortholog family |
| gene\_id | TEXT | Unique identifier for the sequence |
| specie\_abbreviation | TEXT | Genome abbreviation |

**RUNNING DeNoGAP**

In order to run analysis using DeNoGAP execute commands shown below in the command line terminal.

cd DeNoGap\_v1.0

cd bin

perl Denogap\_v1.0.pl **–genome\_info** <genome information file> **-db\_dir** <sqlite database dir path> **-db\_name** <name of the sqlite database> **-config** <path and name of configuration file> **-output\_dir** <path to the output directory>

**Running DeNoGAP with test dataset**

The package contains 2 dataset that are under **data/test1** and **data/test2** directories.

The data/test1 contains input files to initiate new analysis with new set of genomes.

The data/test2 contains input files to add new genomes to the existing analysis database.

* **Test parse genbank files**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/PARSE\_GENBANK.config –output\_dir ../output/TEST\_RUN

* **Test gene prediction**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/PREDICT\_GENE.config –output\_dir ../output/TEST\_RUN

* **Test gene verification**

- Download and extract UniProt database.

(<ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz>)

* Merge sequences predicted by multiple programs and single program for all genomes in a single sequence file using cat command.

* Perform protein Blast and store output in default format.

blastp –query <All\_predicted\_sequence\_file> -db <Uniprot\_database\_file> -evalue 1e-05 –out <blast\_output\_file> -max\_target\_seq 1

* Perform gene verification analysis using DeNoGAP. Change path and name of Blast alignment file in configuration file to match user-defined blast output file.

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/GENE\_VERIFICATION.config –output\_dir ../output/TEST\_RUN

* **Test Load Data**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/LOAD\_DATA.config –output\_dir ../output/TEST\_RUN

* **Test Reference comparision**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/COMPARE\_REFRENCE.config –output\_dir ../output/TEST\_RUN

* **Test Iterative HMM family prediction**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/PREDICT\_HMM\_FAMILY.config –output\_dir ../output/TEST\_RUN

* **Test Super-homolog family prediction**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/PREDICT\_SUPER\_HOMOLOG.config –output\_dir ../output/TEST\_RUN

* **Test Ortholog Prediction and Ortholog Clustering**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/PREDICT\_ORTHOLOG.config –output\_dir ../output/TEST\_RUN

* **Test Phylogenetic profile**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/PHYLOGENETIC\_PROFILE.config –output\_dir ../output/TEST\_RUN

* **Test Core genome prediction**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/CORE\_GENOME.config -output\_dir ../output/TEST\_RUN

* **Test Function Annotation prediction**

perl DeNoGap\_v1.0.pl -genome\_info ../data/test1/test\_genome\_table\_1.txt –db\_dir ../output/TEST\_RUN –db\_name test.sqlite –config ../config/ANNOTATION.config -output\_dir ../output/TEST\_RUN

**OUTPUT**

This section given as overview of the output directory structure and output files created for each analysis phase.

**Database directory:** User while running DeNoGAP should define the name and path of database directory.

**Database name:** User while running DeNoGAP should define the name of the SQLite database.

**Analysis: Parse GenBank**

The output directory for parsed genbank files contains four sub-directories defined in the configuration file.

* **Genome sequence directory**: This directory contains fasta formatted genome sequence files (one file for each organism).
* **Coding sequence directory**: This directory contains fasta formatted coding gene sequences (one file for each organism).
* **Protein sequence directory**: This directory contains fasta formatted protein sequences (one file for each organism).
* **Genomic feature directory**: This directory contains tab-delimited genomic feature files (one file for each organism).

**Analysis: Gene Prediction**

The output directory for gene prediction contains seven sub-directories defined in the configuration file.

* **Glimmer directory:** This directory stores output from Glimmer software (one directory for each genome).
* **GeneMark directory:** This directory stores output from GeneMark software (one directory for each genome).
* **Prodigal directory:** This directory stores output from Prodigal software (one directory for each genome).
* **FragGeneScan directory:** This directory stores output from FragGeneScan software (one directory for each genome).
* **Coding sequence directory**: This directory contains fasta formatted coding gene sequences (one file for each organism). Coding sequences predicted by multiple programs and single program for each genome are stored in separate sub-directories respectively.
* **Protein sequence directory**: This directory contains fasta formatted translated protein sequences (one file for each organism). Protein sequences predicted by multiple programs and single program for each genome are stored in separate sub-directories respectively.
* **Genomic feature directory**: This directory contains tab-delimited genomic feature files (one file for each organism). Genomic features predicted by multiple programs and single program for each genome are stored in separate sub-directories respectively.

**Analysis: Gene prediction verification and annotation**

The output directory for gene prediction verification stores output files for verified Coding sequences, protein sequences and their genomic features. It also stores genbank files for each genome with verified sequences.

* **VERIFY\_CDS:** This sub-directory stores verified coding sequences for each genome (one file for each genome).
* **VERIFY\_PROTEIN:** This sub-directory stores verified protein sequences for each genome (one file for each genome).
* **VERIFY\_GENOME\_FEATURE:** This sub-directory stores genomic features for verified sequences (one file for each genome).
* **VERIFY\_GENBANK\_FILE:** This sub-directory stores genebank file for each genome (one file for each genome).

**Analysis: Homolog prediction**

The output directory for homolog prediction stores output files for reference genome comparison, iterative hmm-family prediction and ortholog prediction.

* **HOMOLOG\_SCAN**: This directory stores all the output files and sub-directories for homolog prediction analysis.
* **HMMER\_OUT**: This directory stores un-parsed hmmscan and phmmer output files in alignment format and tabular format under HMM\_FULL and HMM\_DOM folders respectively (one file for each organism).
* **BEST\_PAIR:** This sub-directory stores similarity information for highly similar sequences (one file for each organism).
* **CHIMERA\_PAIR:** This sub-directory stores similarity information for chimera-like protein sequences (one file for each organism).
* **ALL\_PAIR:** This sub-directory stores similarity information for all kind of pairwise matches including highly significant hit, significant partial hits, chimera-like hits and insignificant sequence match.
* **MCL:** This sub-directory stores output from MCL clustering.
* **HMM:** This sub-directory stores files for HMM models and singleton sequences for protein families MODEL and SINGLETON sub-directories respectively (one file for each protein family).
* **HMM\_DB:** This sub-directory stores HMM model database files and Singleton sequence database file.
* **SUPER\_HOMOLOG:** This sub-directory stores output files from super-homolog family prediction.
* **ORTHOLOG:** This directory consists of other sub-directories for storing output files from ortholog prediction analysis.
* **PAIR\_DISTANCE:** This sub-directory stores pairwise genetic distance between each pair of proteins in a super-homolog family (one file for each super-homolog family).
* **PAIR\_ORTHOLOG:** This sub-directory stores pairwise genetic distance between each pair of ortholog proteins (one file for each super-homolog family).
* **PAIR\_INPARALOG:** This sub-directory stores pairwise genetic distance between each pair of inparalog proteins (one file for each super-homolog family).
* **ALIGNMENT:** This sub-directory stores multiple alignment of each super-homolog family (one file for each super-homolog family).
* **ORTHO\_CLUSTER:** This sub-directory stores ortholog protein family information (one file for each super-homolog family).
* **RESULT:** This sub-directory stores clustering results for latest protein hmm models, super-homolog family and ortholog family. It also stores latest HMM model database and Singleton sequence database.

**Analysis: Phylogenetic profile**

The output directory to store results from phylogenetic profile analysis.

* **PROJECT\_DIR\_NAME:** The name for this output directory is taken from the PROJECT\_DIR\_NAME parameter in the configuration file. It stores presence and absence information of the protein families across compared genomes as a binary matrix and tabular list.

**Analysis: Core Genome Prediction**

The output directory for core genome prediction stores amino acid sequences and alignments for core gene families.

* **CORE\_SEQ:** This sub-directory stores amino acid sequences for the core gene families in fasta format (one file for each core family).
* **CORE\_ALN:** This sub-directory stores amino acid sequence alignment for the core gene families in fasta format (one file for each core family).
* **CORE\_ALIGNMENT\_FILE:** The name and location of concatenated core alignment file is taken from the CORE\_ALIGNMENT\_FILE parameter in the configuration file.

**Analysis: InterProScan Annotation**

The output directory for InterProScan annotation stores output files for domain prediction, pathway prediction, gene ontology, and signal peptide prediction.

* **INTERPRO\_SCAN\_OUTPUT:** This directory stores all the output files and sub-directories for interproscan annotation.
* **XML\_FILE:** This sub-directory stores un-parsed InterProScan output files in XML format (one file for each genome).
* **TAB\_FILE:** This sub-directory stores un-parsed InterProScan output files in Tab-delimited format (one file for each genome).
* **DOMAIN:** This sub-directory stores parsed domain information (one file for each genome).
* **INTERPRO:** This sub-directory stores parsed interpro domain information (one file for each genome).
* **GENE\_ONTOLOGY:** This sub-directory stores parsed gene ontology information (one file for each genome).
* **PATHWAY:** This sub-directory stores parsed metabolic pathway information (one file for each genome).
* **SIGNALP:** This sub-directory stores parsed signal peptide information (one file for each genome).
* **TMHMM:** This sub-directory stores parsed transmembrane information (one file for each genome).
* **PHOBIUS:** This sub-directory stores parsed results from phobius (one file for each genome).