MetaHCR v1.0
Data Descriptions

Introduction

This document describes some of the data structures in the MetaHCR application. There are two classes of structures:

1. Database Entities
2. Analysis Data Sources

Database Entities are the relational tables and their relationships that are used to store all of the MetaHCR data. The full database schema can be found as an HTML document under the docs/schemas directory. Open the index.html file in a browser. The analysis tables require further explanation. A description of the analyses tables and their relationships is described below

Analysis data files are the files used (via the MetaHCR upload function) to create MetaHCR analyses. They are also described below. Also, consult the Users Guide for a description of the analysis uploading function.

Database Analysis Tables

The Django framework utilizes 'models' for describing data. Models are database-independent. In general there is one model per database table. It is possible to describe a particular model as being a subclass of another model. This is the case with the analysis models. There is a model named BiologicalAnalysis. This model has two subclasses: SingleGeneAnalysis and MetagenomeAnalysis. This means that two subclasses inherit the fields of the super class (Biological Analysis).

When the MetaHCR database is generated, three tables corresponding to the three models are created. At the same time, a field named biologicalanalysis_ptr_id is added to the two subclass tables. In usage, creating one of the subclass data rows also creates a BiologicalAnalysis row and this pointer field is given the id value of the BiologicalAnalysis row.

Analysis Data Sources

MetaHCR can read and process two types of Analysis data: Single Gene and Metagenome.

Single Gene

Single Gene analysis data are in a single file – a so-called 'L6' file, which refers to number of taxonomic levels. This file is also know as an OTU Summary file. The file is an 'csv' file that uses the tab character as a field delimiter. Each line corresponds to an organism. Its fields, in the order that they appear in a line, are:

- OTU # - the organism's taxonomy. It consist of the names of six taxonomic levels. Each level value is separated by a semi-colon '';'. The levels are (in order): kingdom, phylum, class, order, family, genus and species.

- 1 to n Analysis names – There is one column per sample. The sample name must correspond to the analysis name: the analysis_name field in the biological_analysis table. The value corresponding to these analyses represents the score for this organism in the analysis. It is a floating point number and can be in scientific notation or zero ('0').

Metagenome

Metagenome analysis data comes in the form of three files that are read in the following order:

1. Scaffold
2. Protein (Gene)
3. RNA

Unlike the single gene analysis files, there is one set of three files per metagenome analysis. When uploading these files to the MetaHCR database, you must specify the Sample on which this analysis was performed. Each file is a 'csv' file with the tab character as the field delimiter. Each file is described below.

Scaffold

The scaffold file has one line per organism. The fields that are used during upload are:

- Scaffold ID – a character string
- Scaffold Length (bp) – a floating point or empty
- Lineage Percentage - a floating point
- Gene Count – an integer
- Taxonomic fields: Lineage Domain, Lineage Phylum, Lineage Class, Lineage Order, Lineage Family, Lineage Genus and Lineage Species – character strings

Protein (Gene)

The Protein (Gene) file has one line per protein function:

- Scaffold – a character string corresponding to the Scaffold ID in the scaffold file.
- Gene ID, Gene Name, Taxon ID, Assembled?, Locus Type, Start Coord, End Coord, Gene Length, Strand, Scaffold Length, Scaffold GC, Scaffold Depth, # of Genes on Scaffold, COG ID, COG Function, Pfam ID, Pfam Function, TIGRfam ID, TIGRfam Function, EC Number, Enzyme Function, KO ID, KO Function

RNA

The RNA file has one line per Gene Product:

- Scaffold ID – a character string corresponding to the Scaffold ID in the scaffold file.
- Gene ID, Locus Type, Gene Product Name, Gene Symbol, Coordinates, Length, Scaffold Length, Scaffold GC Content, Scaffold Read Depth