HMS-S-S Manual

Sebastian Tanabe

01.01.2022

Contents

1 What does HMS-S-S do?

Sulfur compounds are used in a variety of biological processes including respiration and photosynthesis. The biochemistry underlying the manifold transformations of inorganic sulfur compounds occurring in sulfur metabolizing prokaryotes is astonishingly complex and knowledge about it has increased over the last years.
HMS-S-S is a tool to search for proteins related to sulfur metabolism and sort the found genes by their location to reveal synteny. The output of the results can be filtered by co-location, or co-occurrence in the same genome. complex filtering is also possible using mysql. The tool is supported by a graphical interface. Furthermore, it is possible to expand the number of enzymes to be considered by including additional hidden markov models from pfam or other sources.
In its basic version, HMS-S-S contains 164 HMMs from sulfur metabolism and structurally related enzymes for better distinguishability. Thus, distinctions can be made between otherwise very similar protein sequences, which, however, have great significance for the probable metabolic potential.

2 Setting Up HMS-S-S

HMS-S-S runs on Linux, but requires some dependencies. Setup instructions are given below.

Essential

Perl

Bioperl

MySQL with table structure from schema.sql

HMMER3

Recommended

Prodigal

Ncbi taxonomy database

2.1 Set Up HMS-S-S

1. Download the latest release from github:
2. Extract the folder to a location of your choice
3. Run the perl script HMSSS.pl with perl

```
$ perl HMSSS.pl
```

2.2 Dependencies

2.2.1 Perl

The Perl interpreter is included in the default installation of Ubuntu, but can otherwise be downloaded via

```
$ sudo apt−get install perl
```

Packages that are required but possibly not installed by the default installation can be downloaded from CPAN:

- DBI;
- DBD::mysql;
- Parallel::ForkManager;

2.2.2 Bioperl

Bioperl is a toolbox of perl packages written specifically for computational molecular biology applications. These include but are not limited to reading sequences from fasta formats and processing BLAST or HMMER reports. To perform a hmmsearch and analysis HMSSS requires some dependencies in the path. Brief instructions on installation are given, but can also be obtained from the installation notes of the packages itself. Further information can be found at https://bioperl.org. The installation can be done with the following command:

```
$ sudo apt−get install bioperl
```

Optionally the packages can also be installed via CPAN. The following packages from bioperl are required:

- Bio::Seq
- Bio::SeqIO
- Bio::SearchIO
- Bio::SearchIO::FastHitEventBuilder
- Bio::Tools::Run::StandAloneBlastPlus
- Bio::DB::Taxonomy

2.2.3 HMMER

HMMER is a software package that uses profile hidden markov models (HMM) to detect homologous sequences. Compatible HMMs can be obtained from databases such as Pfam or Interpro, or created from a custom sequence alignment via the HMMER package itself. Further information, including a manual on HMMER installation, can be found at http://hmmer.org/. The installation can be done with the following command.

```
$ sudo apt−get install hmmer
```

2.2.4 Prodigal

Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) is a microbial (bacterial and archaeal) gene finding program, that learns the ORF prediction directly by the provided data. The installation can be done with the following command.

```
$ sudo apt-get install prodigal
```

2.2.5 MySQL/MariaDB

MySQL and MariaDB are completely interchangeable relational database management systems that even share the same syntax and commands. However, the source code of MariaDB is freely available. HMS-S-S uses a database based on MariaDB, but could just as well use a MySQL database. For installation and configuration under Ubuntu are some steps necessary, starting with:

```
$ sudo apt-get install mariadb-server
$ mysql --version          #Check for correct installation
mysql  Ver 15.1 Distrib 10.3.32-MariaDB, for debian-linux-gnu
(i686) using readline 5.2
$ sudo mysql -u root -p
Enter password:          #Enter your ubuntu password

#Choose a Username a password for your database account
> GRANT ALL ON *.* TO Username IDENTIFIED BY "password";
> GRANT FILE ON *.* TO Username;
> QUIT
```

To log in with root, the password of your computer is required. After that, you can create a local user with your own username and an independent password. With GRANT ALL the user is created. with GRANT FILE the write and read rights are assigned.
Now it is possible to log in with the local newly created user with the following command. Afterwards a database must be created, whereby the name can be freely selected again. This database will later be used to store the search results. With the last command the database schema is configured. this step must be executed from the same directory where schema.sql is stored, or specify the complete path to this file. The schema of the database is shown in Fig 5.

```
$ mysql -u Username -p
Enter password:

> CREATE DATABASE database_name;
> QUIT

#Enter your custom username, password and database_name
$ mysql -u Username -p database_name < schema.sql
```

2.2.6 BLAST+

NCBI BLAST+ enables local usage of the BLAST algorithm. It is not necessarily required by HMS-S-S but can be used as an addition to HMMER.

```
$ sudo apt-get install ncbi-blast+
```

2.2.7 Taxonomy database

For assemblies derived from NCBI the NCBI Taxonomy database can be used to add taxonomic information. The database taxdmp.zip can be downloaded from NCBI FTP server:

```
https://ftp.ncbi.nih.gov/pub/taxonomy/
```

The unpacked taxdump folder containing the .dmp files has to be moved to HMS-S-S source folder.

3 Running HMS-S-S

HMSSS has a graphical user interface that supports the program flow. The interface requires no installation and can be started directly from the directory with the command:

```
$ perl HMSSS.pl
```

All input FASTA should be located in a single folder. Nucleotide FASTA formats must have a .fa, .fna, .fa.gz or .fna.gz file suffix. For protein FASTA format a .faa.gz suffix is required, as well as a corresponding gff3 file with the same filename but a gff3.gz file suffix.

3.1 Login to the database

The connection menu allows login to the local database. This connection to a database with a table schema as specified in schema.sql is required for all functions and must therefore always be made. The name of the database, the user name and the password are the same as those used when connecting to the database. The menu frame allows switching between the functions of HMS-S-S (Fig. 1).
Username: MySql database username
Database: database with table structure from schema.sql
Password: Password for the MySQL login

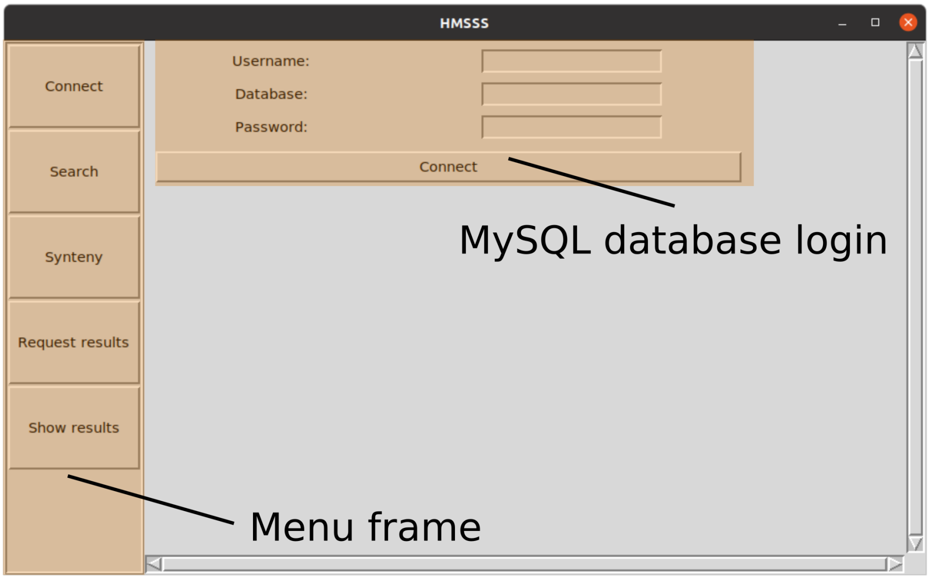


Figure 1: Connection login screen for the database. Left: Menu frame with search and analysis options.

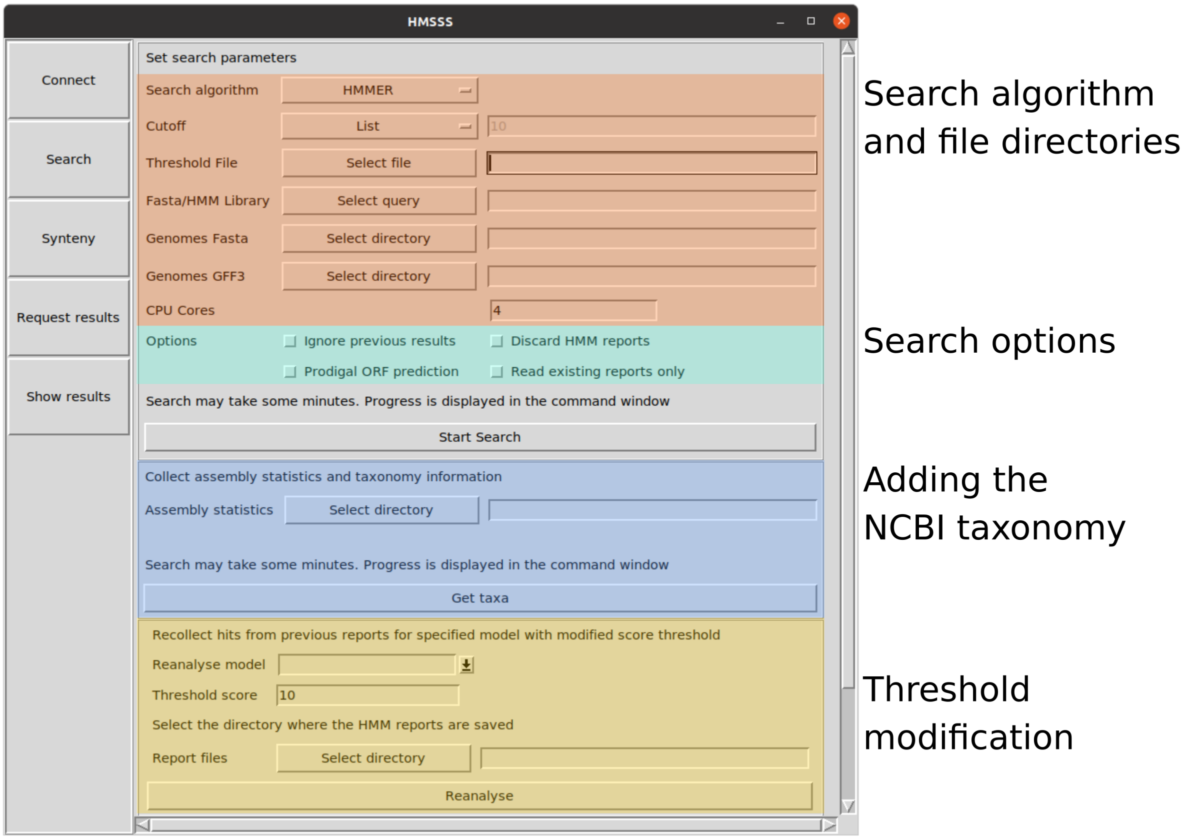
3.2 Searching

The search menu allows you to define search parameters, set target directories, insert taxonomy information and adjust threshold scores (Fig. 2). HMMER or BLAST can be used as the search algorithm. Either a single value or an external list with cutoff values can be defined as cutoff. If a single value is selected, all hits in the search that lie below this cutoff are not included in the local database. with a list, a specific cutoff is used depending on the respective query or HMM. The path to the external list is specified under threshold file option. A threshold list is formatted as a tab separated file with two columns. The first column contains the names of the HMM or querys the second contains the respective cutoff value. In the further target directories the HMM library for a hmmsearch or a fasta file with queries for a BLASTp search is specified. Also the folders containing the FASTA files to be searched and the general genmoic features in gff3 format have to be specified. If the files are nucleotide FASTA files without an associated gff3 file, the same location as that of the nucleotide fasta files should be specified.

Seach algorithm: HMMER or BLAST+
Cutoff: Single threshold value or threshold List
Threshold File: Location of the threshold List
Fasta/HMM Library: Location of a query FASTA file or HMM library File
Genomes Fasta: Directory with genome files
Genomes GFF3: Folder with GFF3 formatted files
CPU Cores: Number of cores used by HMMER

Input format The selected folders should include genomes or metagenomes, which must all be in FASTA format, comprised of contigs or scaffolds. Nucleotide FASTA formats must have a .fa, .fna, .fa.gz or .fna.gz file suffix. For protein FASTA format a .faa or faa.gz suffix is required, as well as a corresponding gff3 file with the same filename but a gff or gff.gz file suffix. The name of the file will be taken as genome identifier by the program.

Input: Directory with FASTA files, contig or scaffold
Suffix: .fa, .fna, .fa.gz, .fna.gz, .faa, .faa.gz, .gff, .gff.gz



Set search parameters

Search algorithm: HMMER
Cutoff: List, 10
Threshold File: Select file
Fasta/HMM Library: Select query
Genomes Fasta: Select directory
Genomes GFF3: Select directory
CPU Cores: 4

Options: □ Ignore previous results □ Discard HMM reports □ Prodigal ORF prediction □ Read existing reports only

Search may take some minutes. Progress is displayed in the command window

Collect assembly statistics and taxonomy information

Assembly statistics: Select directory

Search may take some minutes. Progress is displayed in the command window

Recollect hits from previous reports for specified model with modified score threshold

Reanalyse model:
Threshold score: 10

Select the directory where the HMM reports are saved

Report files: Select directory

Search algorithm and file directories

Search options

Adding the NCBI taxonomy

Threshold modification

Figure 2: Search option menu. Orange: Algorithm selection, cutoff and HMM library selection and pathway to folders. Cyan: Extended search options including the use Prodigal for nucleotide FASTA file ORF prediction and translation. Blue: NCBI assembly statistics pathway and taxonomy. Yellow: Reiteration of reports with altered cutoff threshold for single HMM.

Search options There are several options included. If nucleotide FASTA is provided, the "Prodigal ORF prediction" has to be checked in order to create protein FASTA files and corresponding gff3 files. Prodigal has to be installed for this option.

Once searched assemblies are stored in the local database at least with their name, as well as the hits. To avoid multiple searches in the same assembly and thus save resources, already existing assemblies in the database are skipped. However, if you want to search them again with hmmsearch you have to select the "Ignore previous results" option. This way these assemblies will not be skipped, searched again and new hits will be written into the database.

With the "Discard HMM reports" option the reports created by hmmsearch will be deleted after relevant results were transferred to the local database. This option saves memory but also prevents the subsequent use of these reports by "Read existing reports only". With the latter option no new hmmsearch will be performed but the old reports will be examined, skipping the actual hmmsearch. This can be useful if more than one cutoff value has been altered since the last hmmsearch.

Ignore previous results Use all FASTA files in the folder

Discard HMM reports Clean the hmmsearch reports after they were used

Prodigal ORF prediction Predict ORFs with Prodigal for nucleotide fasta

Read existing reports Take existing reports only, skip hmmsearch

Adding NCBI taxonomy For the inclusion of the NCBI taxonomy, the assembly statistic files for each assembly are required. These can be downloaded from NCBI. HMSSS then collects the NCBI taxonID from the assembly static files and matches the corresponding phylogenetic information to the taxonID from the NCBI Taxon database. The taxdump files have to be located in the source folder. All results are then written to the local database.

Assembly statistics: Directory with NCBI assembly statistics

Threshold modification Like the "read existing reports only" option this requires existing hmmsearch report files. Under reanalyse model a singel HMM model can be selected which is then added to the database with a modified threshold. Hits in the report files that meet the new cutoff are then additionally written to the database. This will overwrite entries if the reanalysed model has a higher score. Raising the cutoff does not lead to overwrites. If this is required please alter the threshold list file and use the "read existing reports" option. Please note that the paths to protein FASTA and gff3 files of these reports must also be specified in order to include general genomic features and the sequences in the database as well.

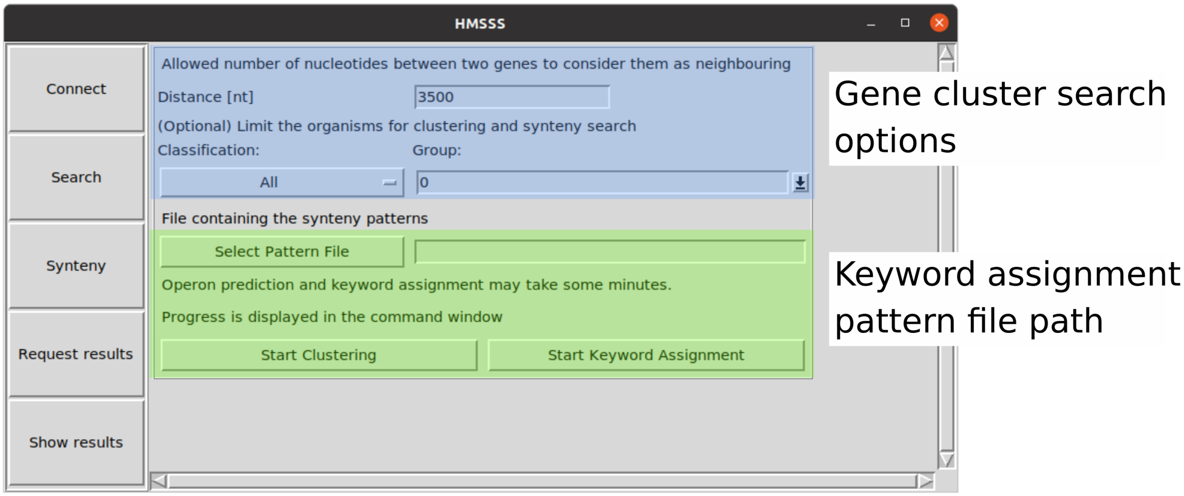
Reanalyse model: HMM to reanalyse with altered score

Threshold score: New cutoff score

Report files: Directory with hmmsearch reports

3.3 Synteny

Synteny describes the physical co-localization of genetic loci on the same chromosome. For HMSSS two genes are considered to be synthenic if the distance between them is less than a user defined nucleotide number. By default this distance is set to 3500 nucleotides. The synteny algorithm searches for two genes that meet these requirements and sorts them into a gene cluster. If another gene is within this distance, calculated from the outer boundaries of the cluster, the cluster is expanded by this gene. This continues until no further gene meets the requirement. Optionally, the search for gene clusters can be narrowed down to a specific phylogenetic group (Fig. 3).



Allowed number of nucleotides between two genes to consider them as neighbouring

Distance [nt] 3500

(Optional) Limit the organisms for clustering and synteny search

Classification: All

Group: 0

File containing the synteny patterns

Operon prediction and keyword assignment may take some minutes.

Progress is displayed in the command window

Figure 3: Cluster menu Blue: Minimum distance between two genes to define a gene cluster. Started by Start Clustering Button. Green: Directory with pattern file for keyword assignment to gene clusters. Started by Start Keyword Assignment.

The here so called synteny patterns are genes expected to be co-localized in a gene cluster. The pattern file to be selected is a tabular separated table with at least 4 columns. The first column contains an integer, which defines the number of genes from a pattern to be present to assign a keyword. The second column contains the keyword. Keywords are assigned to gene clusters if the cluster matches at least the minimal number of genes necessary, defined by the first column. All following columns define gene names, i.e. the HMM or query names, expected in a gene cluster. For example the synteny pattern:

4 Tus TusA TusB TusC TusD TusE

would match every combination of TusABCDE, where one of the gene may be missing would assign the keyword Tus to the cluster. In contrast synteny pattern:

5 Tus TusA TusB TusC TusD TusE

would only match a combination of TusABCDE, where all genes are present.

3.4 Output

The Hits from Hmmsearch are stored in a local relational database. This includes general genomic features like contig, start, end and strand information, the amino acid sequence, the hit score, information about the phylogeny of the assembly and assigned keywords. A ER diagram of the database is shown in figure 5. Results can be obtained from the database by sql commands. To help the user with this task, the interace has an interface for creating a mysql query. Optionally, this query can also be displayed and modified, or it can be completely defined directly by the user. The interface provides checkboxes to select columns to be shown in the retrieved results (Fig. 4 orange Box).

the interface also provides the possibility to limit the number of assemblies from which results are displayed. this can be done either by phylogeny or by keyword. If the phylogeny is limited, only entries belonging to a selected taxonomic level are shown. If a keyword is selected, only assemblies that have at least one gene cluster with the selected keyword are included (Fig. 4 cyan Box). Which sequences from the database are to be displayed can also be specified. the selection can be made by keywords or by protein type. Combinations are also possible. If a keyword is selected, all entries are displayed that are located in a gen cluster with the corresponding keyword. If a protein type is selected, the output will only contain entries that match the selected protein type. A possible combination of both would therefore only output entries that are located in a specified gene cluster and have a specific protein type e.g all genes in a Tus gene cluster or all TusA genes in a Tus gene cluster (Fig. 4 blue Box).

More complex requests concerning co-localisation and co-occurence or absence in the same genome, even of serveral genes are possible. For these complex request the interface provides a textfield for entering a custom query. Querys from this textfield are directly handed over to the database and therefore have to be written in correct syntax (Fig. 4 yellow Box).

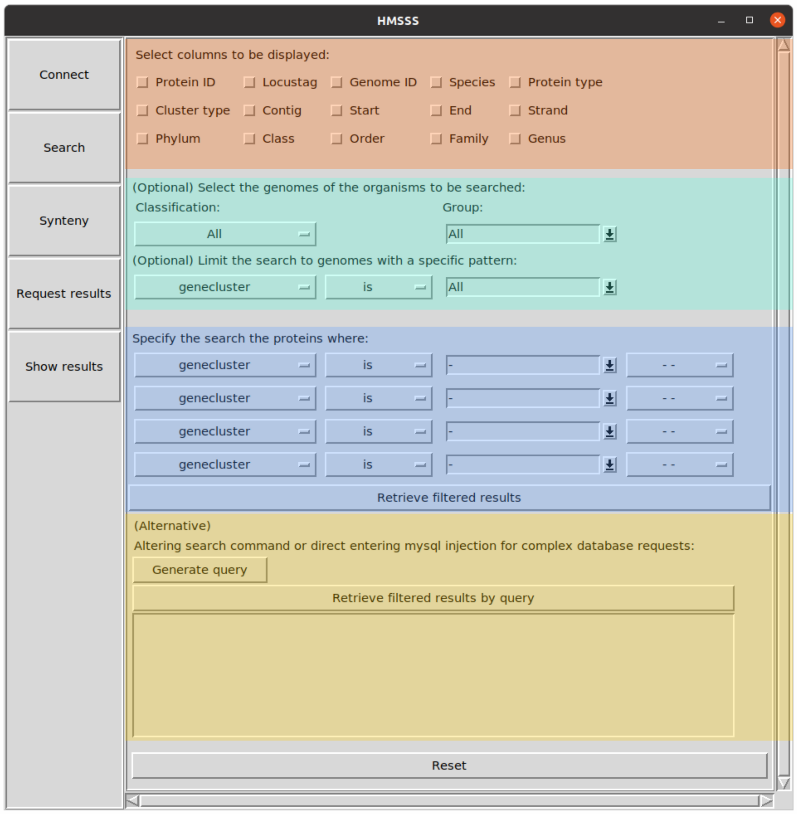
The generate query button creates the query but the selected options an displays it in a textfield, but does not execute it directly. The user can now modify the query and execute it with the "Retrieve filtered results by query" button. The textfield also allows to insert completely custom mysql querys.

Orange: Selection of information to be displayed from each entry.

Cyan: Limit assemblies from which results are retrieved.

Blue: Filtering displayed results by keyword or protein type or both.

Yellow: Custom query insertion, or interface supported generated query modification.


HMSSS
Connect
Search
Synteny
Request results
Show results
Select columns to be displayed:
☐ Protein ID ☐ Locustag ☐ Genome ID ☐ Species ☐ Protein type
☐ Cluster type ☐ Contig ☐ Start ☐ End ☐ Strand
☐ Phylum ☐ Class ☐ Order ☐ Family ☐ Genus
(Optional) Select the genomes of the organisms to be searched:
Classification:
Group:
All
All
(Optional) Limit the search to genomes with a specific pattern:
genecluster
is
All
Specify the search the proteins where:
genecluster
is
- -
Retrieve filtered results
(Alternative)
Altering search command or direct entering mysql injection for complex database requests:
Generate query
Retrieve filtered results by query
Reset

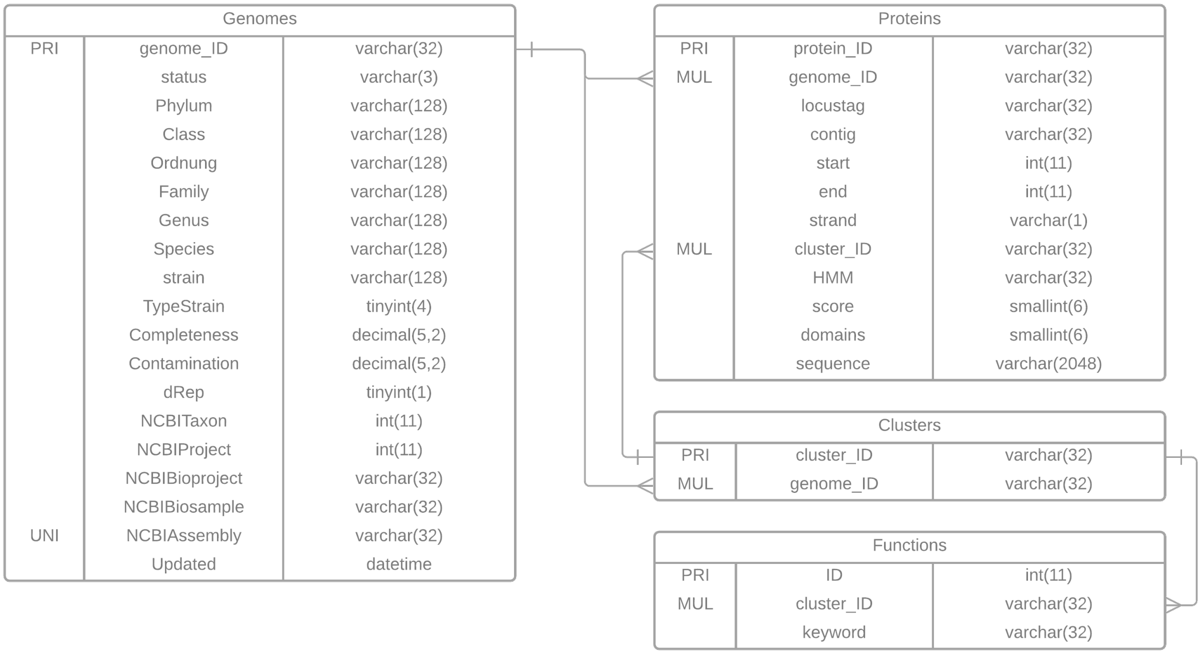
Select information presented in the results

Limit displayed results by phylogeny or existing keyword

Specify retrieved sequences

Custom query insertion

Figure 4: Request results Orange: Selection of information. Cyan: Limit assemblies from which results are retrieved. Blue: Filtering displayed results by keyword or protein type. Yellow: Custom query insertion, or interface supported generated query modification.



Genomes

PRI	genome_ID	varchar(32)
	status	varchar(3)
	Phylum	varchar(128)
	Class	varchar(128)
	Ordnung	varchar(128)
	Family	varchar(128)
	Genus	varchar(128)
	Species	varchar(128)
	strain	varchar(128)
	TypeStrain	tinyint(4)
	Completeness	decimal(5,2)
	Contamination	decimal(5,2)
	dRep	tinyint(1)
	NCBITaxon	int(11)
	NCBIProject	int(11)
	NCBIBioproject	varchar(32)
	NCBIBiosample	varchar(32)
UNI	NCBIAssembly	varchar(32)
	Updated	datetime

Proteins

PRI	protein_ID	varchar(32)
MUL	genome_ID	varchar(32)
	locustag	varchar(32)
	contig	varchar(32)
	start	int(11)
	end	int(11)
	strand	varchar(1)
MUL	cluster_ID	varchar(32)
	HMM	varchar(32)
	score	smallint(6)
	domains	smallint(6)
	sequence	varchar(2048)

Clusters

PRI	cluster_ID	varchar(32)
MUL	genome_ID	varchar(32)

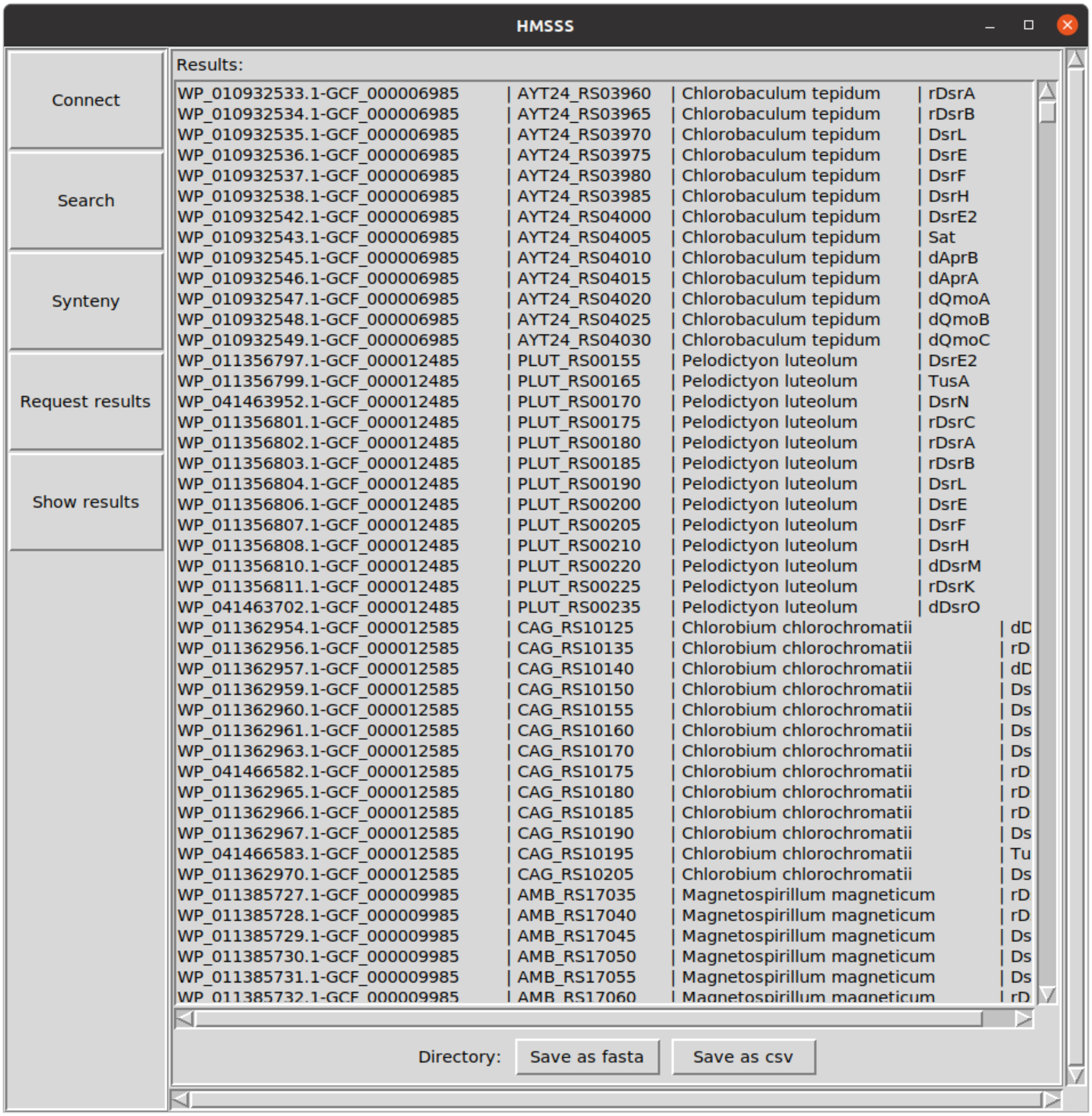
Functions

PRI	ID	int(11)
MUL	cluster_ID	varchar(32)
	keyword	varchar(32)

Figure 5: Database schema Entity relationship diagram with constrains of the HMS-S-S underlying database. Can be imported to any mysql database via schema.sql

Retrieved results are shown in an additional frame (Fig. 6). They can either be extracted and saved in a .csv file, containing all selected columns, or in protein FASTA file. Sequence headers of the FASTA file are derived from the selected columns.

CSV file: Selected columns for each entry
FASTA file: Selected columns for each entry as headers + protein sequence



```
WP_010932533.1-GCF_000006985  | AYT24_RS03960  | Chlorobaculum tepidum        | rDsrA
WP_010932534.1-GCF_000006985  | AYT24_RS03965  | Chlorobaculum tepidum        | rDsrB
WP_010932535.1-GCF_000006985  | AYT24_RS03970  | Chlorobaculum tepidum        | DsrL
WP_010932536.1-GCF_000006985  | AYT24_RS03975  | Chlorobaculum tepidum        | DsrE
WP_010932537.1-GCF_000006985  | AYT24_RS03980  | Chlorobaculum tepidum        | DsrF
WP_010932538.1-GCF_000006985  | AYT24_RS03985  | Chlorobaculum tepidum        | DsrH
WP_010932542.1-GCF_000006985  | AYT24_RS04000  | Chlorobaculum tepidum        | DsrE2
WP_010932543.1-GCF_000006985  | AYT24_RS04005  | Chlorobaculum tepidum        | Sat
WP_010932545.1-GCF_000006985  | AYT24_RS04010  | Chlorobaculum tepidum        | dAprB
WP_010932546.1-GCF_000006985  | AYT24_RS04015  | Chlorobaculum tepidum        | dAprA
WP_010932547.1-GCF_000006985  | AYT24_RS04020  | Chlorobaculum tepidum        | dQmoA
WP_010932548.1-GCF_000006985  | AYT24_RS04025  | Chlorobaculum tepidum        | dQmoB
WP_010932549.1-GCF_000006985  | AYT24_RS04030  | Chlorobaculum tepidum        | dQmoC
WP_011356797.1-GCF_000012485  | PLUT_RS00155   | Pelodictyon luteolum         | DsrE2
WP_011356799.1-GCF_000012485  | PLUT_RS00165   | Pelodictyon luteolum         | TusA
WP_041463952.1-GCF_000012485  | PLUT_RS00170   | Pelodictyon luteolum         | DsrN
WP_011356801.1-GCF_000012485  | PLUT_RS00175   | Pelodictyon luteolum         | rDsrC
WP_011356802.1-GCF_000012485  | PLUT_RS00180   | Pelodictyon luteolum         | rDsrA
WP_011356803.1-GCF_000012485  | PLUT_RS00185   | Pelodictyon luteolum         | rDsrB
WP_011356804.1-GCF_000012485  | PLUT_RS00190   | Pelodictyon luteolum         | DsrL
WP_011356806.1-GCF_000012485  | PLUT_RS00200   | Pelodictyon luteolum         | DsrE
WP_011356807.1-GCF_000012485  | PLUT_RS00205   | Pelodictyon luteolum         | DsrF
WP_011356808.1-GCF_000012485  | PLUT_RS00210   | Pelodictyon luteolum         | DsrH
WP_011356810.1-GCF_000012485  | PLUT_RS00220   | Pelodictyon luteolum         | dDsrM
WP_011356811.1-GCF_000012485  | PLUT_RS00225   | Pelodictyon luteolum         | rDsrK
WP_041463702.1-GCF_000012485  | PLUT_RS00235   | Pelodictyon luteolum         | dDsrO
WP_011362954.1-GCF_000012585  | CAG_RS10125    | Chlorobium chlorochromatii             | dD
WP_011362956.1-GCF_000012585  | CAG_RS10135    | Chlorobium chlorochromatii             | rD
WP_011362957.1-GCF_000012585  | CAG_RS10140    | Chlorobium chlorochromatii             | dD
WP_011362959.1-GCF_000012585  | CAG_RS10150    | Chlorobium chlorochromatii             | Ds
WP_011362960.1-GCF_000012585  | CAG_RS10155    | Chlorobium chlorochromatii             | Ds
WP_011362961.1-GCF_000012585  | CAG_RS10160    | Chlorobium chlorochromatii             | Ds
WP_011362963.1-GCF_000012585  | CAG_RS10170    | Chlorobium chlorochromatii             | Ds
WP_041466582.1-GCF_000012585  | CAG_RS10175    | Chlorobium chlorochromatii             | rD
WP_011362965.1-GCF_000012585  | CAG_RS10180    | Chlorobium chlorochromatii             | rD
WP_011362966.1-GCF_000012585  | CAG_RS10185    | Chlorobium chlorochromatii             | rD
WP_011362967.1-GCF_000012585  | CAG_RS10190    | Chlorobium chlorochromatii             | Ds
WP_041466583.1-GCF_000012585  | CAG_RS10195    | Chlorobium chlorochromatii             | Tu
WP_011362970.1-GCF_000012585  | CAG_RS10205    | Chlorobium chlorochromatii             | Ds
WP_011385727.1-GCF_000009985  | AMB_RS17035    | Magnetospirillum magneticum            | rD
WP_011385728.1-GCF_000009985  | AMB_RS17040    | Magnetospirillum magneticum            | rD
WP_011385729.1-GCF_000009985  | AMB_RS17045    | Magnetospirillum magneticum            | Ds
WP_011385730.1-GCF_000009985  | AMB_RS17050    | Magnetospirillum magneticum            | Ds
WP_011385731.1-GCF_000009985  | AMB_RS17055    | Magnetospirillum magneticum            | Ds
WP_011385732.1-GCF_000009985  | AMB_RS17060    | Magnetospirillum magneticum            | rD
```

Figure 6: Displayed results

3.5 Database backup

Creating a backup from a database can be done using mysqldump from the terminal command line.

```
$ mysqldump --databases database_name > backup.sql
```

Reloading SQL-format backups can be done in a similar way:

```
$ mysql -u Username -p database_name < backup.sql
```