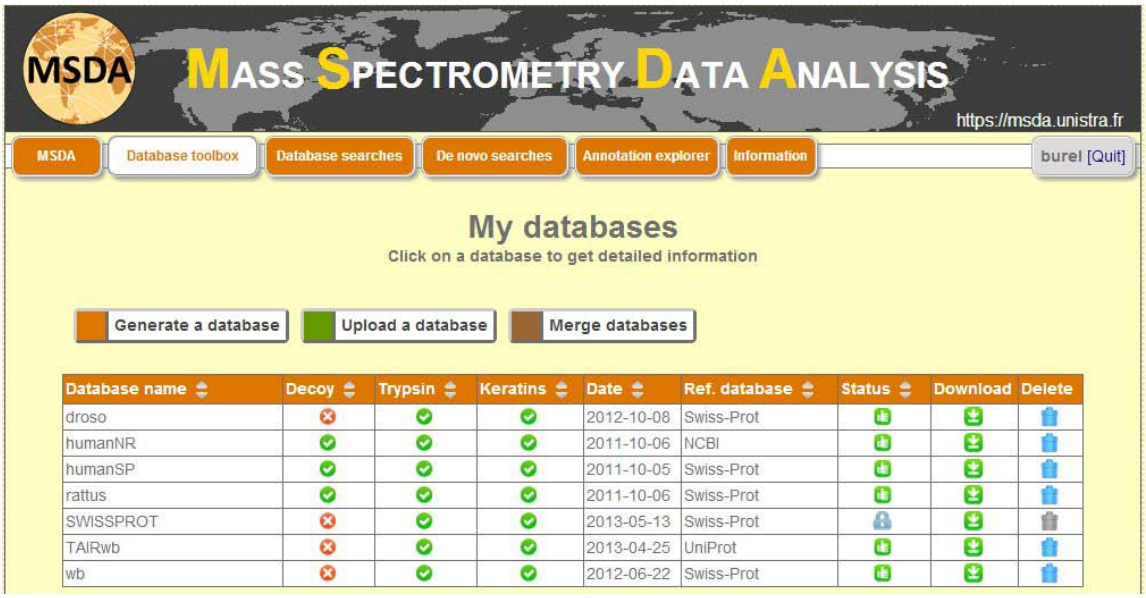
https://msda.unistra.fr

MSDA tutorial: How to create your fasta database

There are many ways to create a fasta database using the Database Toolbox in MSDA:

- Generate a database by extracting selected taxonomies
- Generate a database with a list of accession numbers
- Generate a database with custom amino acid sequences
- Generate a database containing common contaminants and decoy entries
- Upload a fasta database you have on your computer
- Generate a database by merging two or more databases



Generate a fasta database

The generation of a fasta database consists in extracting subsets of proteins from a reference database. The reference databases proposed in MSDA are:

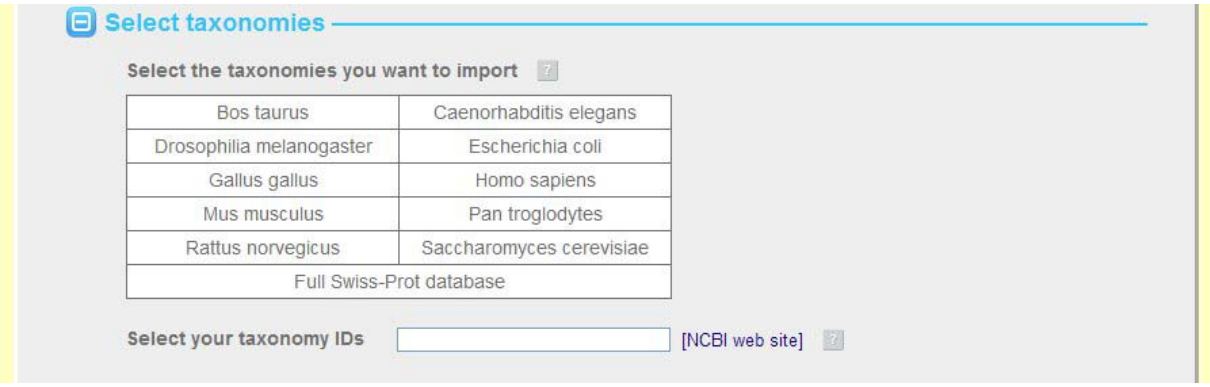
- UniProtKB-SwissProt
- UniProtKB (SwissProt + TrEMBL)
- NCBInr

The first thing to do to generate a database is to select your reference database (SwissProt, UniProtKB or NCBInr).

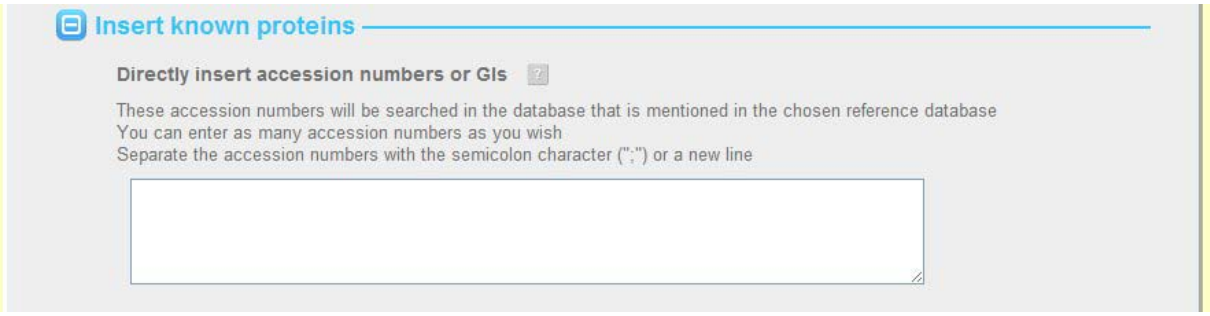
https://msda.unistra.fr

You can select one or several taxonomy IDs in order to extract all proteins contained in the selected taxonomy levels to generate your database. A list of common taxonomies is proposed in MSDA, but you can specify any taxonomy you want in the text area as long as you use their identifier (that can be found on the NCBI web site).

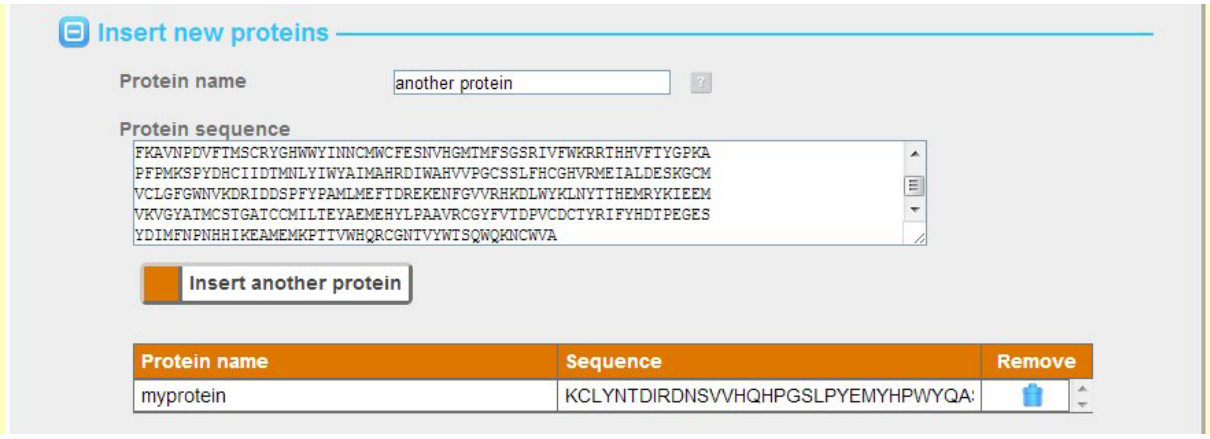
For UniProtKB-SwissProt it is possible to extract the root taxonomy (Taxonomy ID=1) but note that it is not allowed for UniProtKB and NCBInr, as those databases exceed 20GB, size that is even doubled when adding decoy entries.



You can also add or generate a database containing a list of known accession numbers, without any taxonomy restriction. The corresponding proteins will be extracted from the chosen reference database.

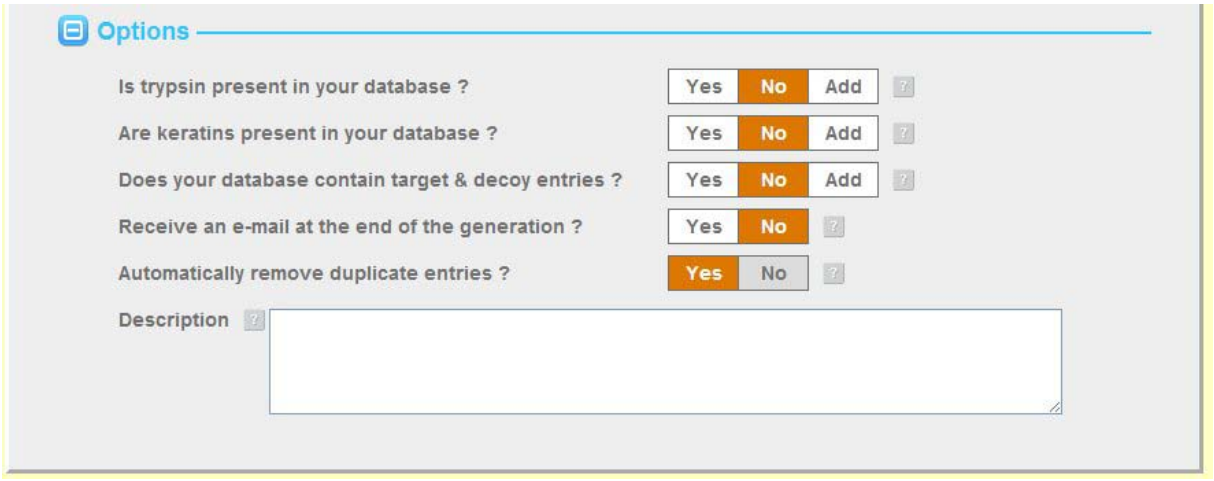


You can also add custom amino acid sequences, that will be added to your database.



https://msda.unistra.fr

You can combine all upper tools to generate your database in a single step. Once done, you still have a few options to set.



Common contaminants (trypsin and keratins) can be added to your database if you answer “Add” to the question “Are contaminants (trypsin or keratins) present in your database?”.

If you answer “Yes”, your database will be tagged (in its name and annotation) as containing these contaminants, but they won’t be added. If you answer “No”, these contaminants won’t be added and your database will not get the “Contaminant” tag in its name and annotation.

Either way, duplicate entries will be removed from your database as they cause problems for database search algorithms like Mascot or OMSSA. You won’t have the same accession number twice in your database, but you may have identical protein sequences as long as they have different accession numbers.

Decoy entries can be generated using the same system as for trypsin and keratins. Decoy proteins are generated at the end of the process, after the addition of contaminants and the removal of duplicates.

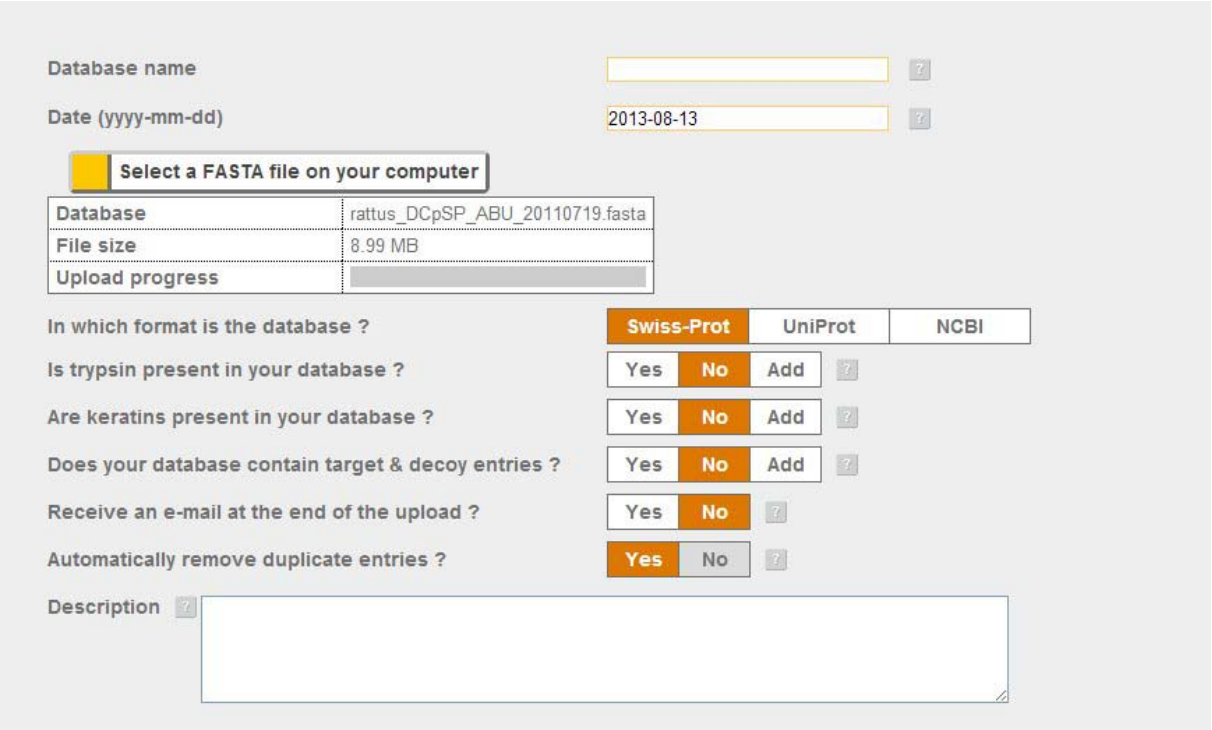
You can ask to be warned by E-mail when your database is generated. It can be useful for huge databases and in any case, you can continue your work on MSDA while your database is being compiled.

https://msda.unistra.fr

Upload your own fasta database

Uploading a fasta database is a straightforward operation. You just need to select your fasta file on your computer, name it and select your options.

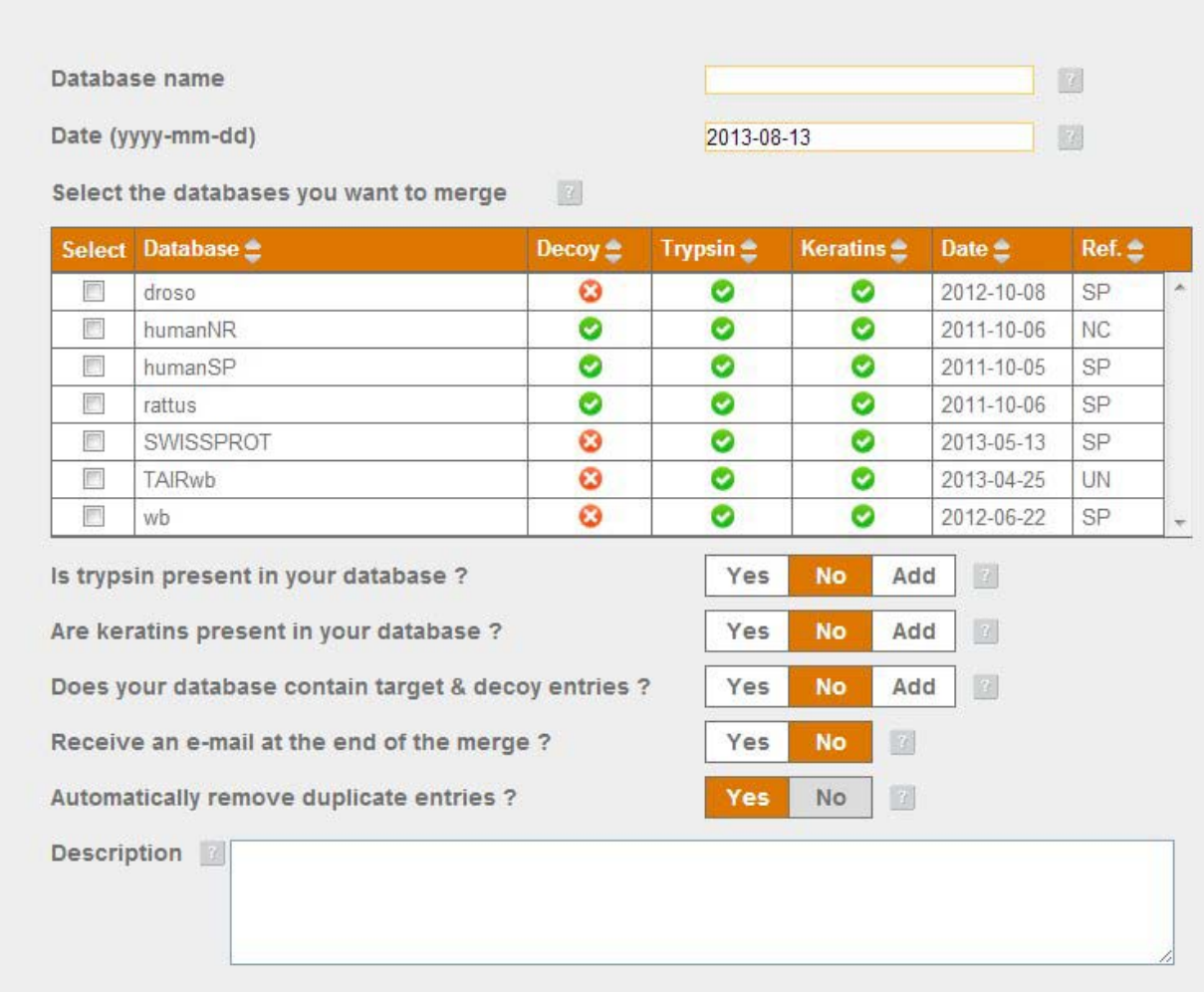
Note that you can select a zipped fasta file.



Merge fasta databases

The database toolbox allows you to merge two or more fasta databases fulfilling two conditions:

1. They must have the same reference database. We do not allow merging SwissProt and NCBInr extracted databases as this would lead to further parsing issues in any case.
2. They must have the same decoy tag: we do not allow merging databases containing decoy entries with databases containing no decoy entries.



Database naming rules

The name of the generated fasta file gives information about its content. For instance, the file "humanSP_DCpSP_burel_20130705.fasta" is composed of:

- The name given by the user (humanSP)
- A tag for decoy entries : "D" if the database contains decoy entries
- A tag for contaminants : "C" if the database contains trypsin and/or keratins

- A tag for the reference database : “SP” for SwissProt, “UN” for UniprotKB, “NC” for NCBInr
- The login of the user (burel)
- The date of creation, formatted as follows yyyymmdd (20130705 correspond to the 5th of July 2013)