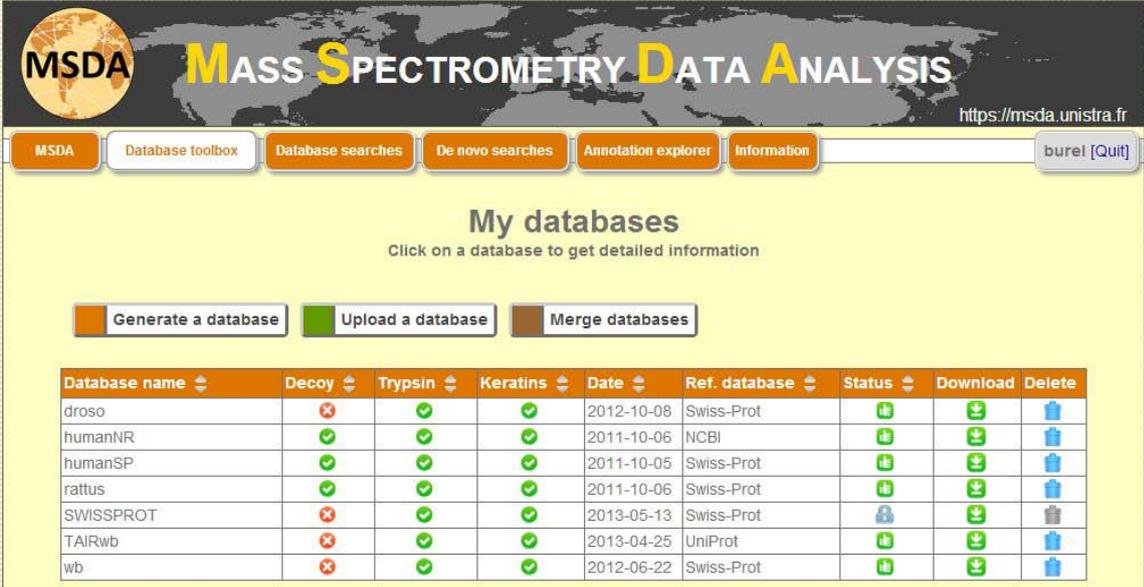


<https://msda.unistra.fr>

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



My databases
Click on a database to get detailed information

Database name	Decoy	Trypsin	Keratins	Date	Ref. database	Status	Download	Delete
droso	✗	✓	✓	2012-10-08	Swiss-Prot	🔒	⬇️	🗑️
humanNR	✓	✓	✓	2011-10-06	NCBI	🔒	⬇️	🗑️
humanSP	✓	✓	✓	2011-10-05	Swiss-Prot	🔒	⬇️	🗑️
rattus	✓	✓	✓	2011-10-06	Swiss-Prot	🔒	⬇️	🗑️
SWISSPROT	✗	✓	✓	2013-05-13	Swiss-Prot	🔒	⬇️	🗑️
TAIRwb	✗	✓	✓	2013-04-25	UniProt	🔒	⬇️	🗑️
wb	✗	✓	✓	2012-06-22	Swiss-Prot	🔒	⬇️	🗑️

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)
- NCBI

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



Generate a database
Generate a database defined by taxonomies, sequences or accession numbers

Database name

Date (yyyy-mm-dd)

From which database do you want to extract your sequences ?

Swiss-Prot
 UniProt
 NCBI

<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 NCBI nr, as those databases exceed 20GB, size that is even doubled when adding decoy entries.

Select taxonomies

Select the taxonomies you want to import ?

Bos taurus	Caenorhabditis elegans
Drosophila melanogaster	Escherichia coli
Gallus gallus	Homo sapiens
Mus musculus	Pan troglodytes
Rattus norvegicus	Saccharomyces cerevisiae
Full Swiss-Prot database	

Select your taxonomy IDs [\[NCBI web site\]](#) ?

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.

Insert known proteins

Directly insert accession numbers or GIs ?

These accession numbers will be searched in the database that is mentioned in the chosen reference database
You can enter as many accession numbers as you wish
Separate the accession numbers with the semicolon character (";") or a new line

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

Insert new proteins

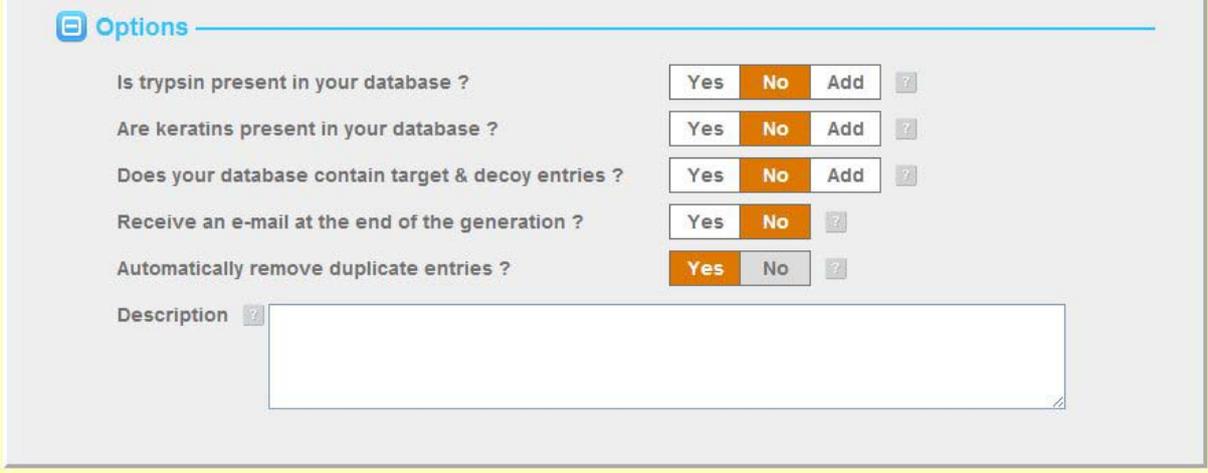
Protein name ?

Protein sequence

```
FKAVNPDVFTMSCRYGHWWYINNMCWCFESNVHGMIMFSGSRIVFWKRRTHHVFYGPKA
FFPMKSPYDHCIIIDTMNLYIWIYAIAHARDIWAHVVPGCCSLFHCGHVRMEIALDESKGCM
VCLGFGWNVKDRIDDSFFYPAMLMEFTDREKENFGVVRHKDLWYKLNYYTHEMRYKIEEM
VKVGYATMCSTIGATCCMILTEYAEMEHVLPAAVRCGYFVTDPVCDCTYRIFYHDTPEGES
YDIMFNPNHHIKEAMEMKPTTVWHQRQCGNIVYWTISQWQKNCWVA
```

Protein name	Sequence	Remove
myprotein	KCLYNTDIRDNSVVHQHPGSLPYEMYHPWYQA:	<input type="button" value="↑"/>

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



Options

Is trypsin present in your database ? Yes No Add ?

Are keratins present in your database ? Yes No Add ?

Does your database contain target & decoy entries ? Yes No Add ?

Receive an e-mail at the end of the generation ? Yes No ?

Automatically remove duplicate entries ? Yes No ?

Description ?

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.

Database name ?

Date (yyyy-mm-dd) ?

Database	rattus_DCpSP_ABU_20110719.fasta
File size	8.99 MB
Upload progress	<div style="width: 0%;"></div>

In which format is the database ?

<input checked="" type="radio"/> Swiss-Prot	<input type="radio"/> UniProt	<input type="radio"/> NCBI
---	-------------------------------	----------------------------

Is trypsin present in your database ?

<input type="radio"/> Yes	<input checked="" type="radio"/> No	<input type="button" value="Add"/> ?
---------------------------	-------------------------------------	--------------------------------------

Are keratins present in your database ?

<input type="radio"/> Yes	<input checked="" type="radio"/> No	<input type="button" value="Add"/> ?
---------------------------	-------------------------------------	--------------------------------------

Does your database contain target & decoy entries ?

<input type="radio"/> Yes	<input checked="" type="radio"/> No	<input type="button" value="Add"/> ?
---------------------------	-------------------------------------	--------------------------------------

Receive an e-mail at the end of the upload ?

<input type="radio"/> Yes	<input checked="" type="radio"/> No	<input type="button" value="Add"/> ?
---------------------------	-------------------------------------	--------------------------------------

Automatically remove duplicate entries ?

<input checked="" type="radio"/> Yes	<input type="radio"/> No	<input type="button" value="Add"/> ?
--------------------------------------	--------------------------	--------------------------------------

Description ?



<https://msda.unistra.fr>

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 NCBI nr 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 name ?

Date (yyyy-mm-dd) ?

Select the databases you want to merge ?

Select	Database	Decoy	Trypsin	Keratins	Date	Ref.
<input type="checkbox"/>	droso	✗	✓	✓	2012-10-08	SP
<input type="checkbox"/>	humanNR	✓	✓	✓	2011-10-06	NC
<input type="checkbox"/>	humanSP	✓	✓	✓	2011-10-05	SP
<input type="checkbox"/>	rattus	✓	✓	✓	2011-10-06	SP
<input type="checkbox"/>	SWISSPROT	✗	✓	✓	2013-05-13	SP
<input type="checkbox"/>	TAIRwb	✗	✓	✓	2013-04-25	UN
<input type="checkbox"/>	wb	✗	✓	✓	2012-06-22	SP

Is trypsin present in your database ? ?

Are keratins present in your database ? ?

Does your database contain target & decoy entries ? ?

Receive an e-mail at the end of the merge ? ?

Automatically remove duplicate entries ? ?

Description ?

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



<https://msda.unistra.fr>

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