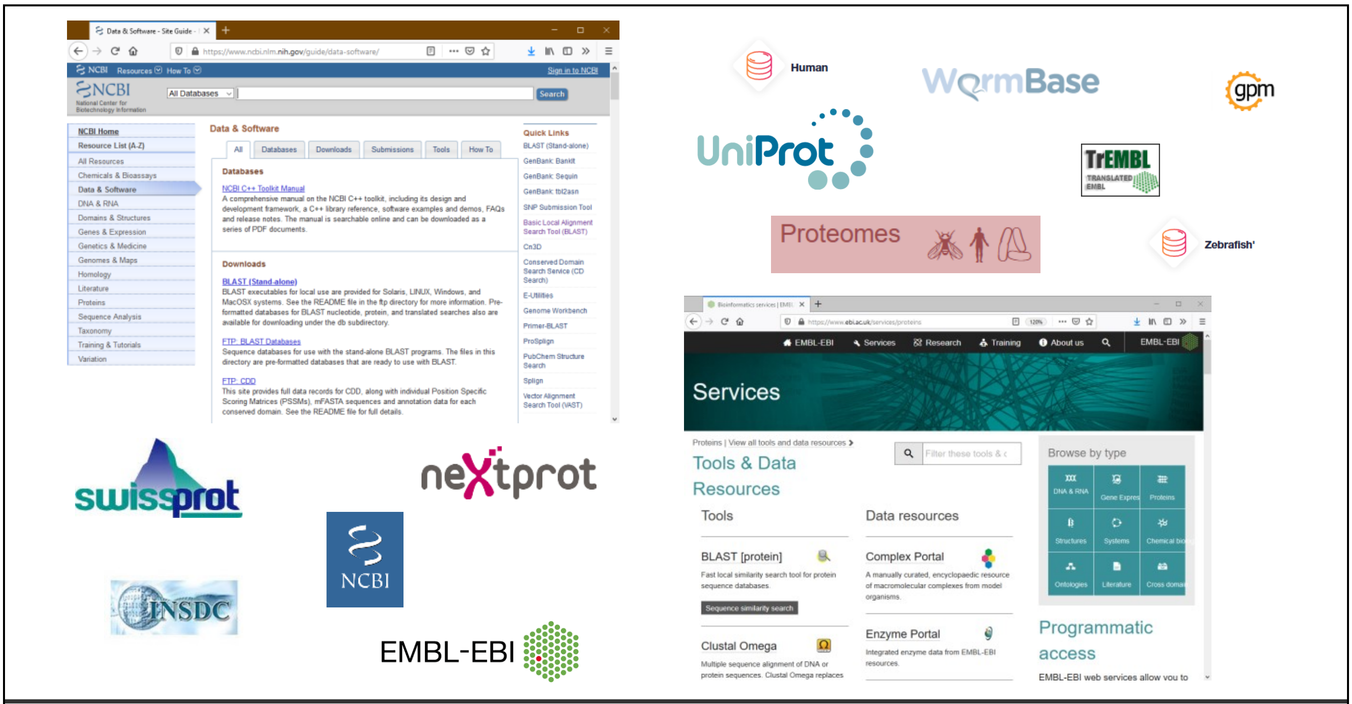
Sequence Database Administration

MASCOT

MATRIX SCIENCE



Databases

NCBI C++ Toolkit Manual
A comprehensive manual on the NCBI C++ toolkit, including its design and development framework, a C++ library reference, software examples and demos, FAQs and release notes. The manual is searchable online and can be downloaded as a series of PDF documents.

Downloads

BLAST (Stand-alone)
BLAST executables for local use are provided for Solaris, LINUX, Windows, and MacOSX systems. See the README file in the ftp directory for more information. Pre-formatted databases for BLAST nucleotide, protein, and translated searches also are available for downloading under the db subdirectory.

FTP: BLAST Databases
Sequence databases for use with the stand-alone BLAST programs. The files in this directory are pre-formatted databases that are ready to use with BLAST.

FTP: CDD
This site provides full data records for CDD, along with individual Position Specific Scoring Matrices (PSSMs), mFASTA sequences and annotation data for each conserved domain. See the README file for full details.

BLAST [protein]
Fast local similarity search tool for protein sequence databases.

Clustal Omega
Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces

Complex Portal
A manually curated, encyclopaedic resource of macromolecular complexes from model organisms.

Enzyme Portal
Integrated enzyme data from EMBL-EBI resources.

EMBL-EBI web services allow you to

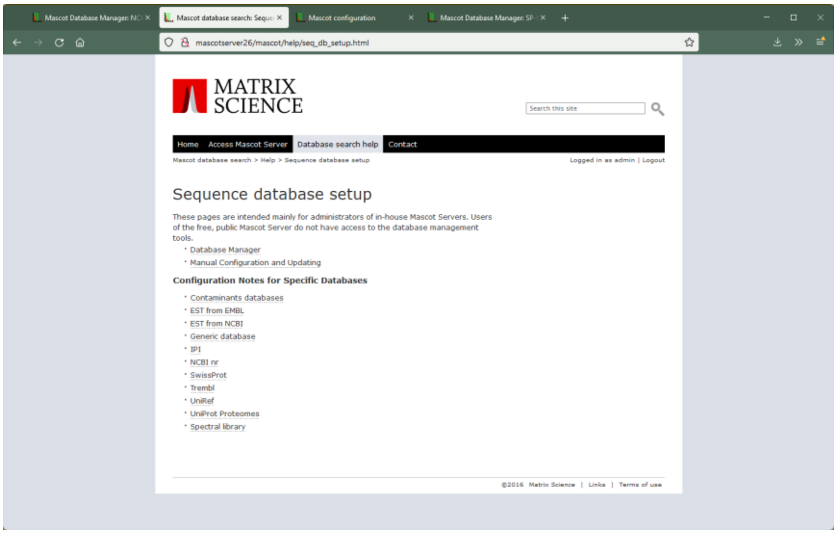
When you install Mascot, it includes a copy of the SwissProt protein database. However, it is almost certain that you and your colleagues will want to search other databases as well. There are very many to choose from, and Mascot allows you to have as many databases as you wish on-line for searching at any one time.

Matrix Science doesn’t supply sequence databases. Most databases are public domain, and there are a few sites that provide comprehensive database repositories. Two of the best known are NCBI and EBI. Here, you can download nr, SwissProt, neXtProt, Trembl, UniRef100, etc.

For specialised databases, such as individual genomes, you may have to track down the FTP site of the group that is doing the sequencing.

This topic described the general procedure for adding a new database to Mascot and keeping it up-to-date.

Sequence Database Requirements



Mascot Database Manager: NC
Mascot database search: Seque
Mascot configuration
Mascot Database Manager: SP
mascotserver26/mascot/help/seq_db_setup.html
MATRIX SCIENCE
Search this site
Home
Access Mascot Server
Database search help
Contact
Mascot database search > Help > Sequence database setup
Logged in as admin | Logout

Sequence database setup

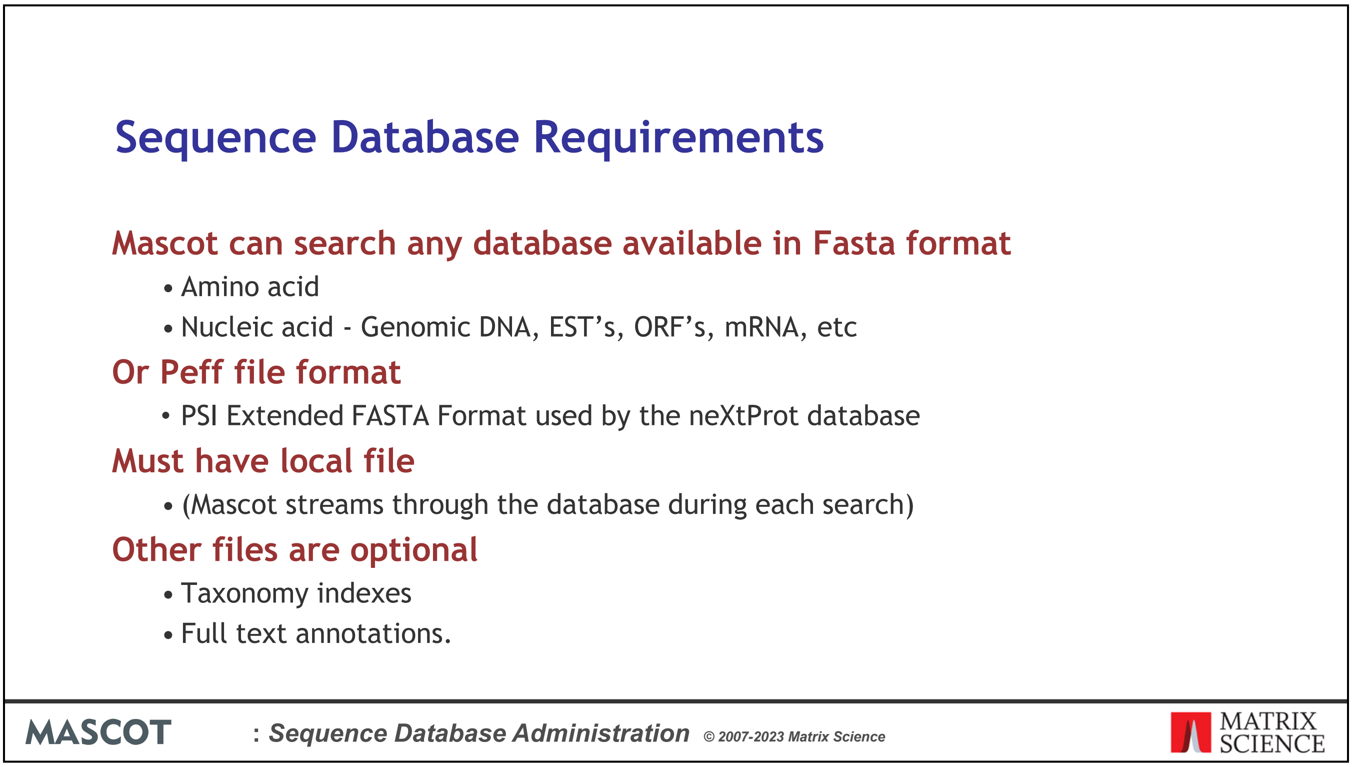
These pages are intended mainly for administrators of in-house Mascot Servers. Users of the free, public Mascot Server do not have access to the database management tools.

- Database Manager
- Manual Configuration and Updating

Configuration Notes for Specific Databases

- Contaminants databases
- EST from EMBL
- EST from NCBI
- Generic database
- IPI
- NCBI nr
- SwissProt
- Trembl
- UniRef
- UniProt Proteomes
- Spectral library

For the latest information about the major public databases, refer to the help pages on the Matrix Science public web site. The help pages in your in-house copy of Mascot are similar, but become progressively out-of-date.



Sequence Database Requirements

Mascot can search any database available in Fasta format

- Amino acid
- Nucleic acid - Genomic DNA, EST’s, ORF’s, mRNA, etc

Or Peff file format

- PSI Extended FASTA Format used by the neXtProt database

Must have local file

- (Mascot streams through the database during each search)

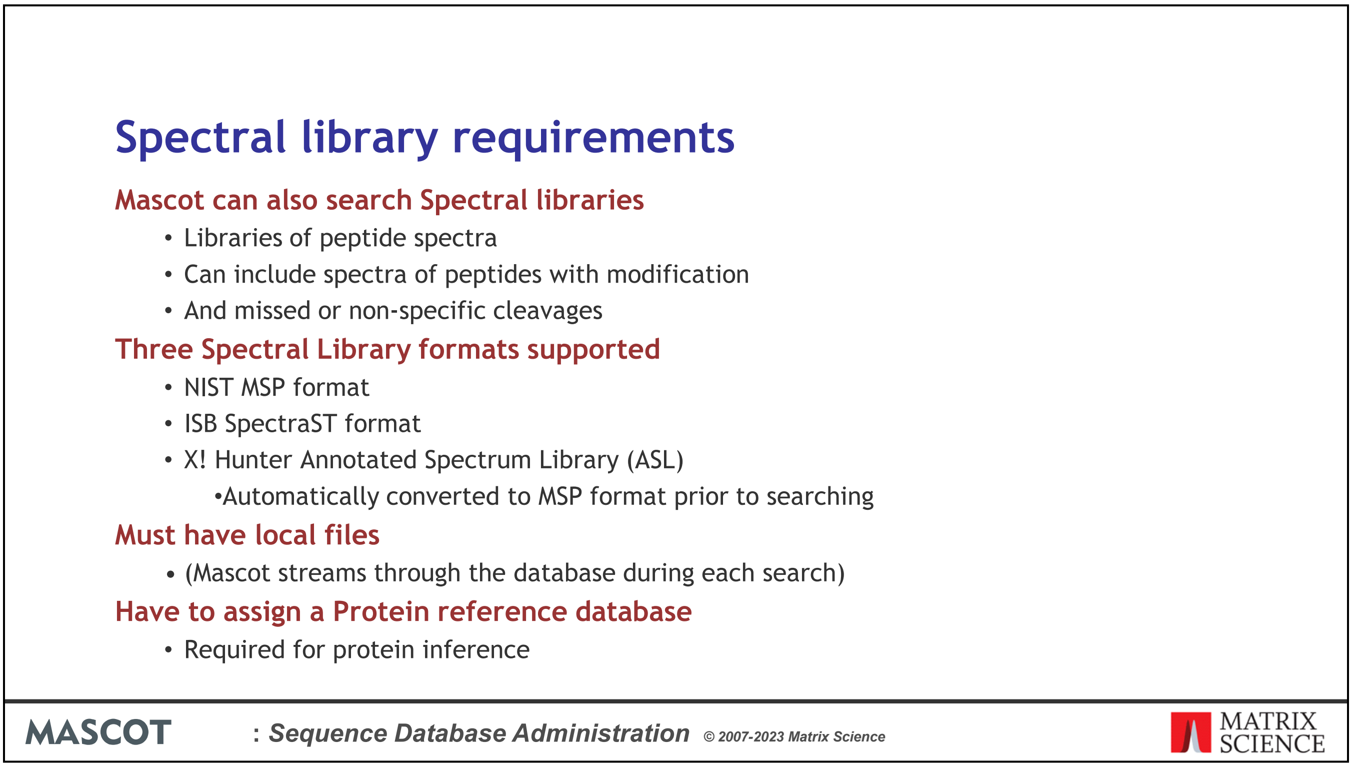
Other files are optional

- Taxonomy indexes
- Full text annotations.

To perform Mascot searches against a database, at a minimum, we need a FASTA file.

If the database contains nucleic acid sequences, there is no need to pre-translate the sequences. Mascot performs a 6 frame translation during each search. Nucleic acid databases come in several flavours. They may be described as genomic DNA, Expressed Sequence Tags, Open Reading Frames, messenger RNA, etc. As far as Mascot is concerned, the main differences are the quality and length of the individual entries. If the database contains entries from multiple organisms, and you want to be able to filter searches by taxonomy, this will require some additional files, which vary from database to database

Some databases, such as SwissProt, also come with 'full text' files, containing comprehensive annotations.



Spectral library requirements

Mascot can also search Spectral libraries

- Libraries of peptide spectra
- Can include spectra of peptides with modification
- And missed or non-specific cleavages

Three Spectral Library formats supported

- NIST MSP format
- ISB SpectraST format
- X! Hunter Annotated Spectrum Library (ASL)
    - Automatically converted to MSP format prior to searching

Must have local files

- (Mascot streams through the database during each search)

Have to assign a Protein reference database

- Required for protein inference

Mascot Server can also search Spectral Libraries. The spectral libraries need to be of peptides; libraries of nucleic acid, sugar or lipid oligomers are not currently supported.

The spectral library can include peptides with modifications and missed or non-specific cleavages.

There are three supported Spectral library formats:

The first one is the most frequently used National Institute of Standards and Technology (NIST) MS *Search* Program (MSP) format.

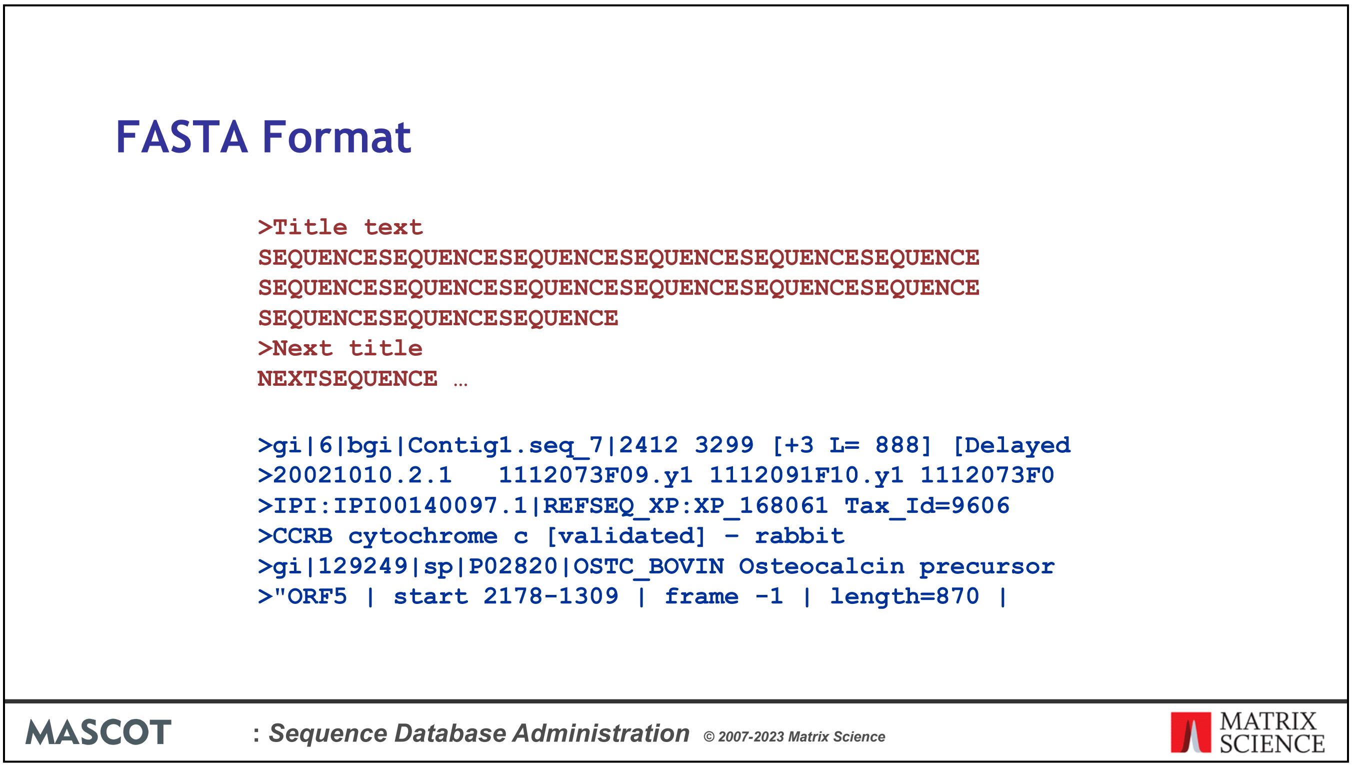
The second one is Institute for Systems Biology SpectraST format.

The final format the X! Hunter Annotated Spectrum Library (ASL) format that is an extension of the MSP format. This database is automatically converted to a MSP database prior to searching.

The spectral library files need to downloaded to the Mascot Server but otherwise do not need any additional supporting files.

When configuring a spectral library you need to specify a Protein reference database that is used for protein inference. SwissProt is commonly used along with a taxonomy filter but individual proteome databases can also be used.

Spectral libraries can be download from the internet or created from Mascot search results. In this talk we will only cover spectral libraries downloaded from the internet. Libraries created from search results are covered in the Spectral Library talk.



FASTA Format

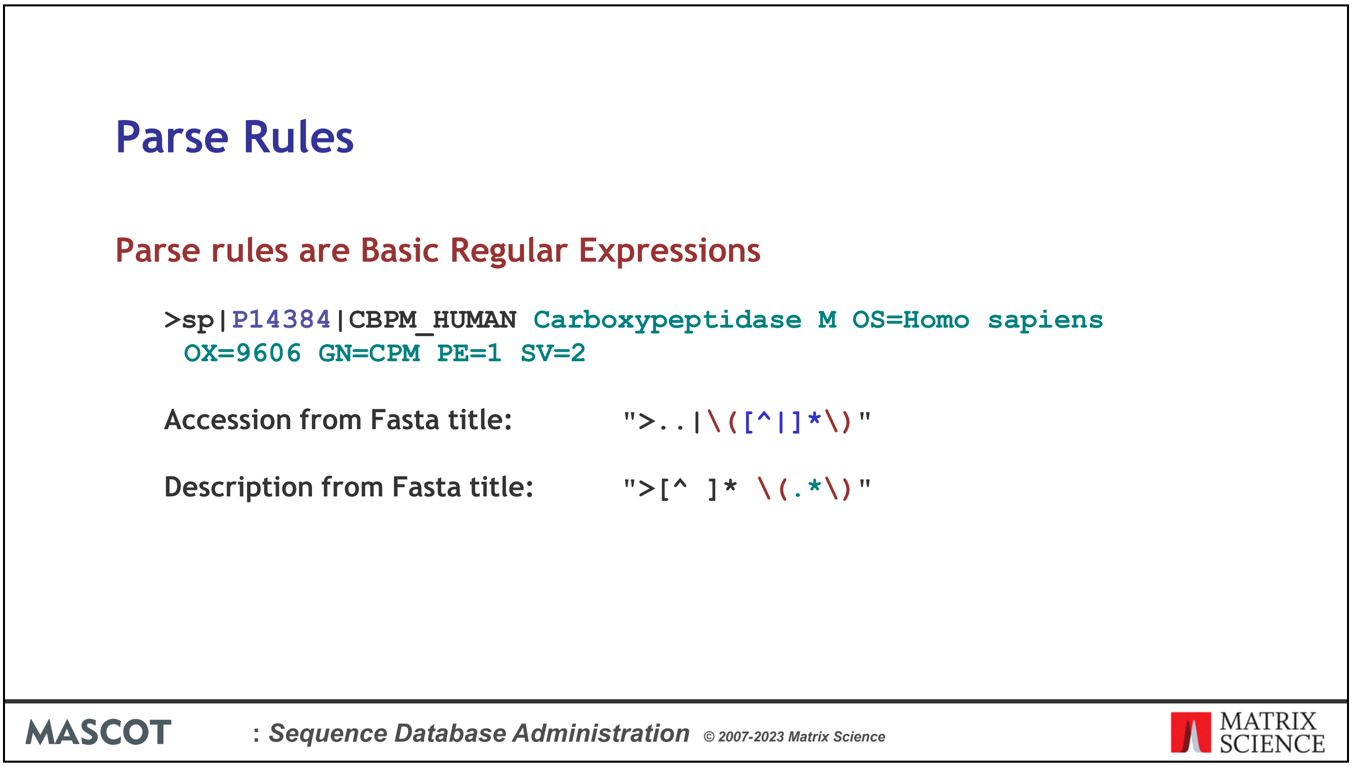
```
>Title text
SEQUENCESEQUENCESEQUENCESEQUENCESEQUENCESEQUENCE
SEQUENCESEQUENCESEQUENCESEQUENCESEQUENCESEQUENCE
SEQUENCESEQUENCESEQUENCE
>Next title
NEXTSEQUENCE …

>gi|6|bgi|Contig1.seq_7|2412 3299 [+3 L= 888] [Delayed
>20021010.2.1   1112073F09.y1 1112091F10.y1 1112073F0
>IPI:IPI00140097.1|REFSEQ_XP:XP_168061 Tax_Id=9606
>CCRB cytochrome c [validated] - rabbit
>gi|129249|sp|P02820|OSTC_BOVIN Osteocalcin precursor
>"ORF5 | start 2178-1309 | frame -1 | length=870 |
```

Perhaps this is a good moment to clarify exactly what we mean by a FASTA file.

FASTA is a very popular standard because it is so simple. On the down-side, it isn't much of a standard ... almost anything goes.

FASTA specifies that there will be a title line, starting with a 'greater than' character, followed by one or more lines containing the sequence in 1 letter code.

The problem is the lack of a well defined syntax for the title line. Here are a handful of examples of FASTA title lines. As you can see, there isn't much similarity. For a Mascot search, we need to find a short, unique identifier or accession string for each sequence. As you can see from these examples, the position of the identifier and the delimiters (e.g. spaces, pipe symbols, commas) varies considerably.



Parse Rules

Parse rules are Basic Regular Expressions

```
>sp|P14384|CBPM_HUMAN Carboxypeptidase M OS=Homo sapiens
 OX=9606 GN=CPM PE=1 SV=2
```

Accession from Fasta title: `">..|\([^|]*\)"`

Description from Fasta title: `">[^ ]* \(.*\)"`

The way Mascot handles this is to use regular expressions to describe how to parse information from the title lines in any particular database. Regular expressions will be familiar to anyone with a Unix background, but there may be a bit of a learning curve for someone with more of a Windows or Mac background.

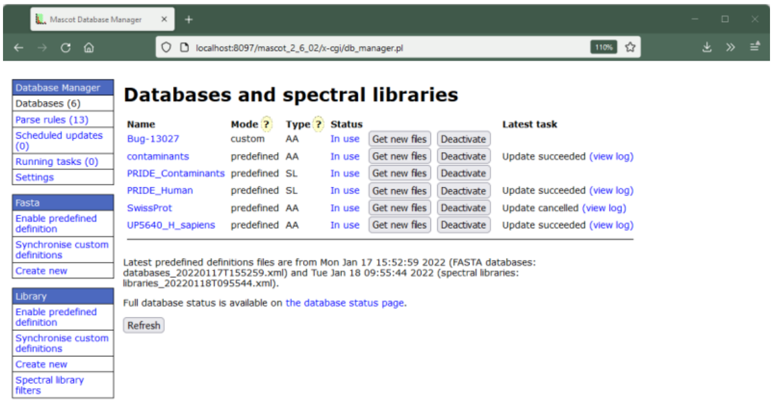
Here, for example, we have a title line from the UniProt Proteome human database. Let's say that we want to use P14384 as the unique accession string and everything after the first space should be treated as the description.

The regular expressions, or parse rules, used to extract this information look like this.

The string we want to extract is always within back-slashed parentheses. For the accession, we show the first few characters as literal text. We then say that we want to take all the following characters, stopping when we hit either a pipe symbol, a space, or a period. In fact, it is the period which applies in this example. The contents of the square brackets are known as a character class, and the circumflex at the beginning means 'not'. The asterisk means 'as many as available'.

For the description, we discard everything up to and including the first space. This is done using a character class of 'not a space' followed by one literal space. Then, we use back-slashed parentheses, take everything to the end of the title. The period matches to any character, so .* matches to all the remaining text.

Database Manager



Databases and spectral libraries

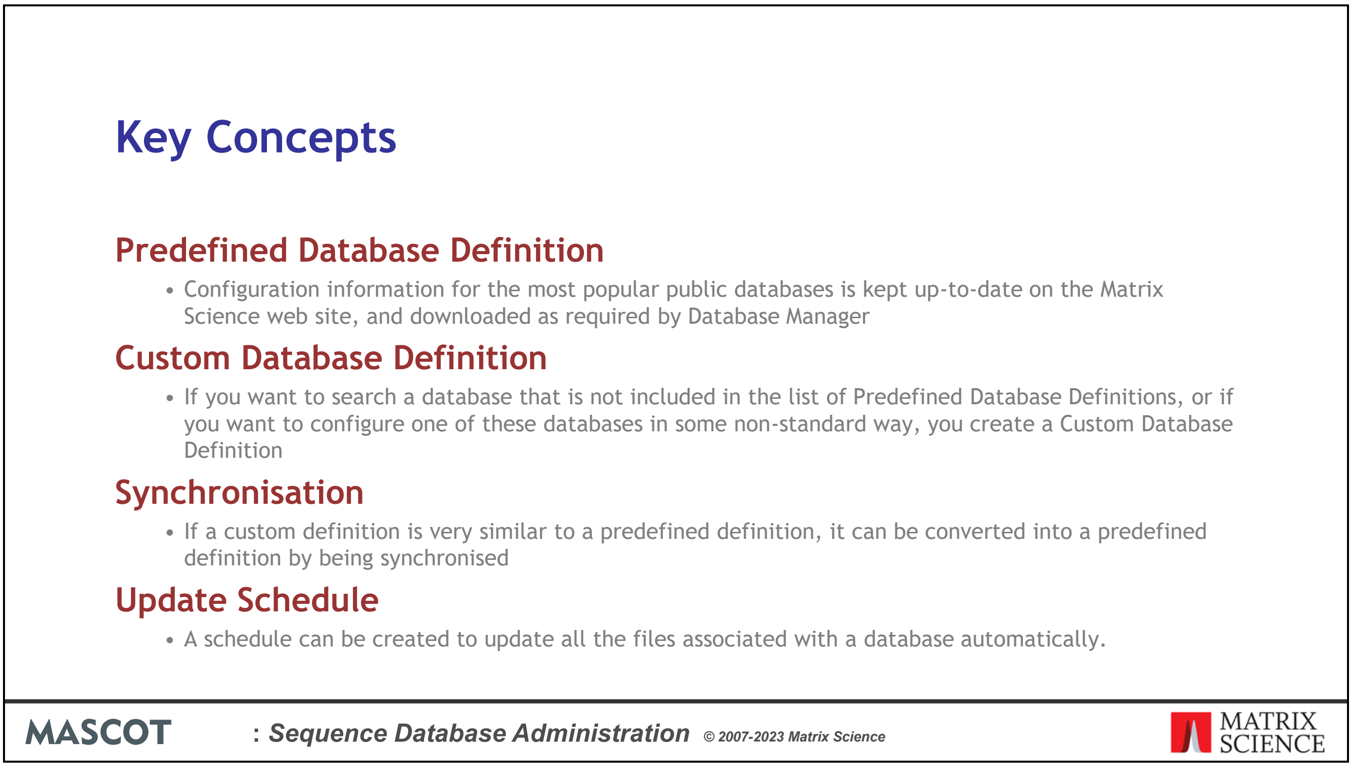
Name	Mode ?	Type ?	Status			Latest task
Bug-13027	custom	AA	In use	Get new files	Deactivate	
contaminants	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
PRIDE_Contaminants	predefined	SL	In use	Get new files	Deactivate	
PRIDE_Human	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
SwissProt	predefined	AA	In use	Get new files	Deactivate	Update cancelled (view log)
UP5640_H_sapiens	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)

Latest predefined definitions files are from Mon Jan 17 15:52:59 2022 (FASTA databases: databases_20220117T155259.xml) and Tue Jan 18 09:55:44 2022 (spectral libraries: libraries_20220118T095544.xml).

Full database status is available on the database status page.

In Mascot 2.4, we introduced Database Manager, which handles both database configuration and the downloading of files from external servers. This replaced two utilities in Mascot 2.3 and earlier: the browser-based Database Maintenance, used for configuration, and the command line Database Update, used for downloading. If you are still using Mascot 2.3 or earlier, you will probably find the archived, 2.3 version of this presentation provides more practical information.

The file formats and download locations of sequence databases change from time to time. One of the smart features of Database Manager is that database configurations for many public databases are updated automatically, by downloading configuration data from the Matrix Science web site.



Key Concepts

Predefined Database Definition

- Configuration information for the most popular public databases is kept up-to-date on the Matrix Science web site, and downloaded as required by Database Manager

Custom Database Definition

- If you want to search a database that is not included in the list of Predefined Database Definitions, or if you want to configure one of these databases in some non-standard way, you create a Custom Database Definition

Synchronisation

- If a custom definition is very similar to a predefined definition, it can be converted into a predefined definition by being synchronised

Update Schedule

- A schedule can be created to update all the files associated with a database automatically.

Let’s review a few of the important terms used in Database Manager:

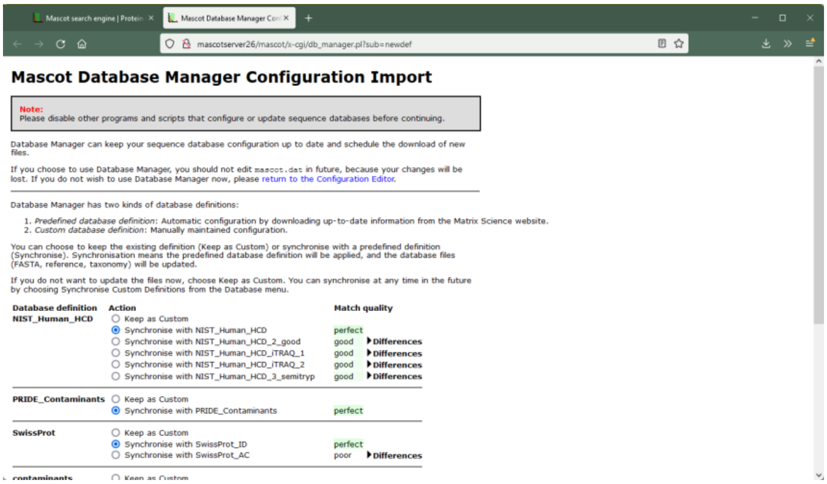
A Predefined Database Definition is one in which the configuration information is kept up-to-date on the Matrix Science web site, and downloaded as required by Database Manager. You don't need to know file URLs or worry about parse rules, etc. for a Predefined Database.

If you want to search a database that is not included in the list of Predefined Database Definitions, or if you want to configure one of these databases in some non-standard way, you create a Custom Database Definition.

If a Custom Database Definition is very similar to a Predefined Database Definition, it can be converted into a predefined definition by being synchronised. The advantage of doing this is that the configuration will then be kept up-to-date automatically.

An Update Schedule can be created to update all the files associated with a database automatically. Maybe once each week or each month. Files will only be downloaded if a new version is available.

Initialisation



Mascot search engine | Protein
Mascot Database Manager Con
mascotserver26/mascot/x-cgi/db_manager.pl?sub=newdef

Mascot Database Manager Configuration Import

Note:
Please disable other programs and scripts that configure or update sequence databases before continuing.

Database Manager can keep your sequence database configuration up to date and schedule the download of new files.

If you choose to use Database Manager, you should not edit `mascot.dat` in future, because your changes will be lost. If you do not wish to use Database Manager now, please return to the Configuration Editor.

Database Manager has two kinds of database definitions:

1. *Predefined database definition*: Automatic configuration by downloading up-to-date information from the Matrix Science website.
2. *Custom database definition*: Manually maintained configuration.

You can choose to keep the existing definition (Keep as Custom) or synchronise with a predefined definition (Synchronise). Synchronisation means the predefined database definition will be applied, and the database files (FASTA, reference, taxonomy) will be updated.

If you do not want to update the files now, choose Keep as Custom. You can synchronise at any time in the future by choosing Synchronise Custom Definitions from the Database menu.

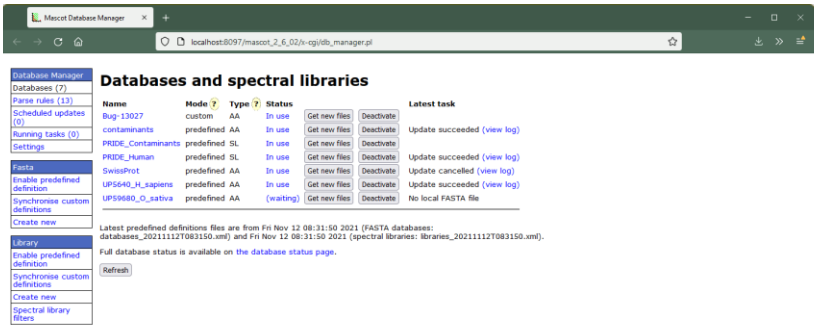
Database definition	Action	Match quality	
NIST_Human_HCD	○ Keep as Custom		
	● Synchronise with NIST_Human_HCD	perfect	
	○ Synchronise with NIST_Human_HCD_2_good	good	▸Differences
	○ Synchronise with NIST_Human_HCD_iTRAQ_1	good	▸Differences
	○ Synchronise with NIST_Human_HCD_iTRAQ_2	good	▸Differences
	○ Synchronise with NIST_Human_HCD_3_semitryp	good	▸Differences
PRIDE_Contaminants	○ Keep as Custom		
	● Synchronise with PRIDE_Contaminants	perfect	
SwissProt	○ Keep as Custom		
	● Synchronise with SwissProt_ID	perfect	
	○ Synchronise with SwissProt_AC	poor	▸Differences
contaminants	○ Keep as Custom		

Database Manager must be allowed exclusive control of database configuration. Editing mascot.dat outside of Database Manager will just cause confusion because Database Manager re-writes mascot.dat whenever a configuration changes. If you prefer to configure sequence databases manually, by editing mascot.dat, never run Database Manager.

The first time Database Manager is run, it tries to match existing database definitions against predefined definitions and reports the quality of the match as none, poor, good, or perfect. For poor or good matches, the differences can be inspected. Usually, these arise because the existing definition is out-of-date in some respect. You can choose whether to synchronise an existing definition, making it predefined, or keep it as a custom definition.

If the Mascot Server is not allowed to access the Internet, choose *Keep as Custom* because synchronisation of any definition requires the database files to be updated.

Initialisation



Databases and spectral libraries

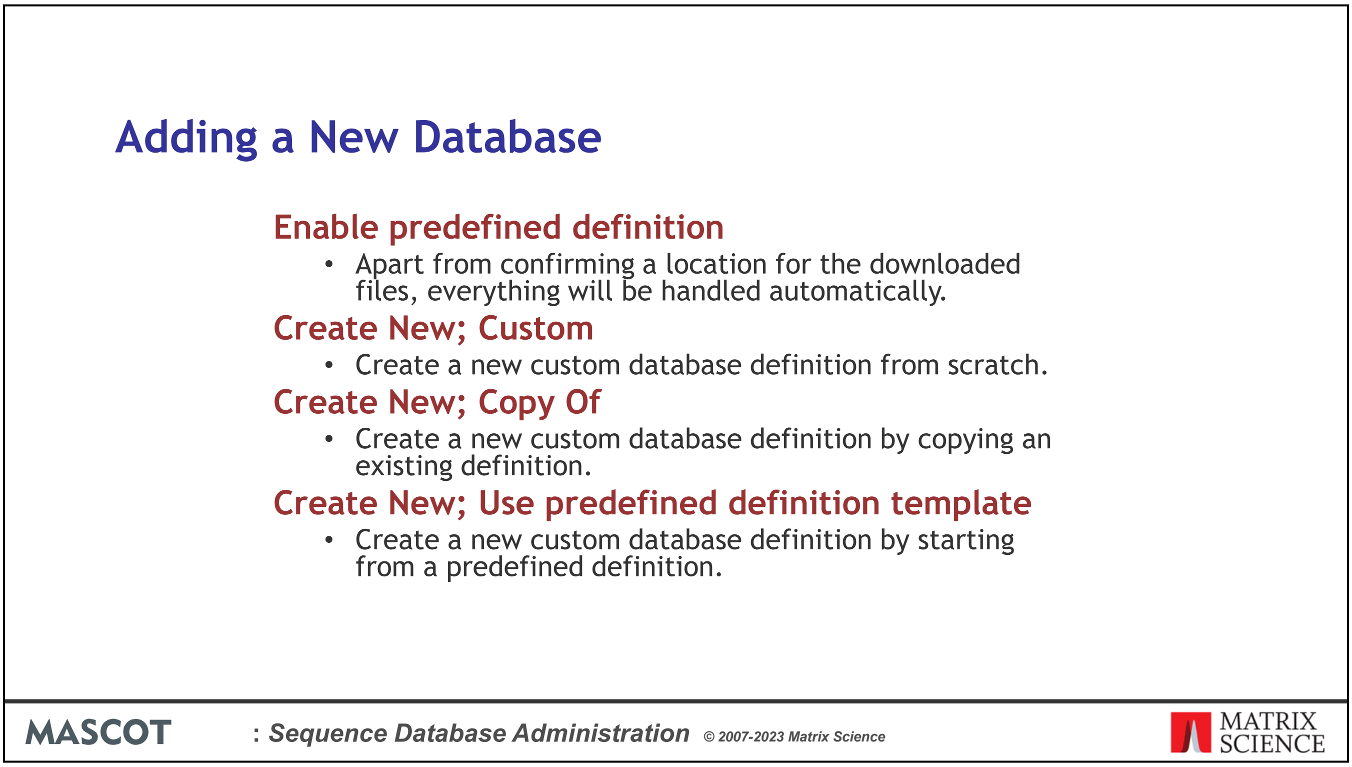
Name	Mode ?	Type ?	Status			Latest task
Bug-13027	custom	AA	In use	Get new files	Deactivate	
contaminants	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
PRIDE_Contaminants	predefined	SL	In use	Get new files	Deactivate	
PRIDE_Human	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
SwissProt	predefined	AA	In use	Get new files	Deactivate	Update cancelled (view log)
UP5640_H_sapiens	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
UP59680_O_sativa	predefined	AA	(waiting)	Get new files	Deactivate	No local FASTA file

Latest predefined definitions files are from Fri Nov 12 08:31:50 2021 (FASTA databases: databases_20211112T083150.xml) and Fri Nov 12 08:31:50 2021 (spectral libraries: libraries_20211112T083150.xml).

Full database status is available on the database status page.

Refresh

Having made your selections, choose *Import* to proceed. The list of Databases will be displayed, with status information for those that have been synchronised and are being updated.



Adding a New Database

Enable predefined definition

- Apart from confirming a location for the downloaded files, everything will be handled automatically.

Create New; Custom

- Create a new custom database definition from scratch.

Create New; Copy Of

- Create a new custom database definition by copying an existing definition.

Create New; Use predefined definition template

- Create a new custom database definition by starting from a predefined definition.

You can add new databases in four different ways:

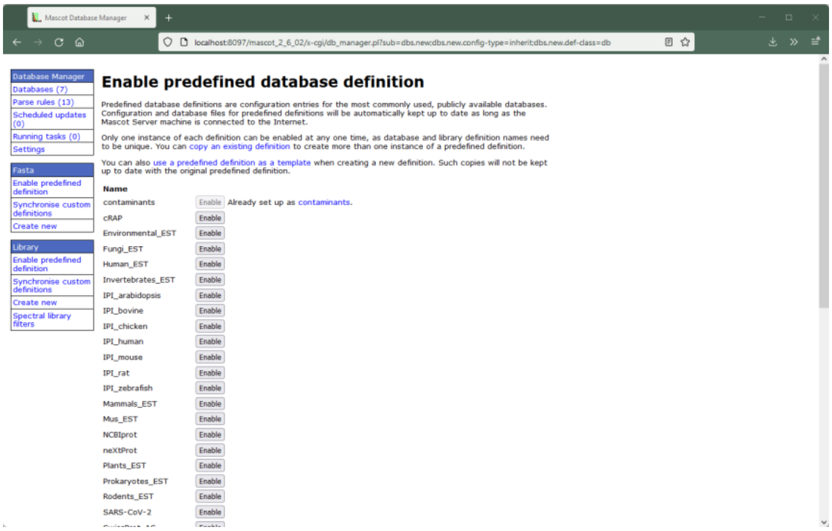
1. Enable predefined definition: Apart from confirming a location for the downloaded files, everything will be handled automatically. Only one instance of each predefined definition can be enabled at any one time, as database names must be unique. If you want something similar to a predefined database, but with configuration changes, choose the final option: Use predefined definition template.

2. Create New; Custom: Create a new custom database definition from scratch.

3. Create New; Copy Of: Create a new custom database definition by copying an existing definition. You will be required to enter a new database name and given the choice of copying the existing database files.

4. Create New; Use predefined definition template: Create a new custom database definition by starting from a predefined definition. The differences between this and enabling a predefined definition are (i) you can make changes to the configuration, (ii) the definition will not be kept up-to-date automatically.

Enable Predefined Definition


Mascot Database Manager
localhost:8097/mascot_2_6_02/x-cgi/db_manager.pl?sub=dbs.new;dbs.new.config-type=inherit;dbs.new.def-class=db
Database Manager
Databases (7)
Parse rules (13)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Enable predefined database definition

Predefined database definitions are configuration entries for the most commonly used, publicly available databases. Configuration and database files for predefined definitions will be automatically kept up to date as long as the Mascot Server machine is connected to the Internet.

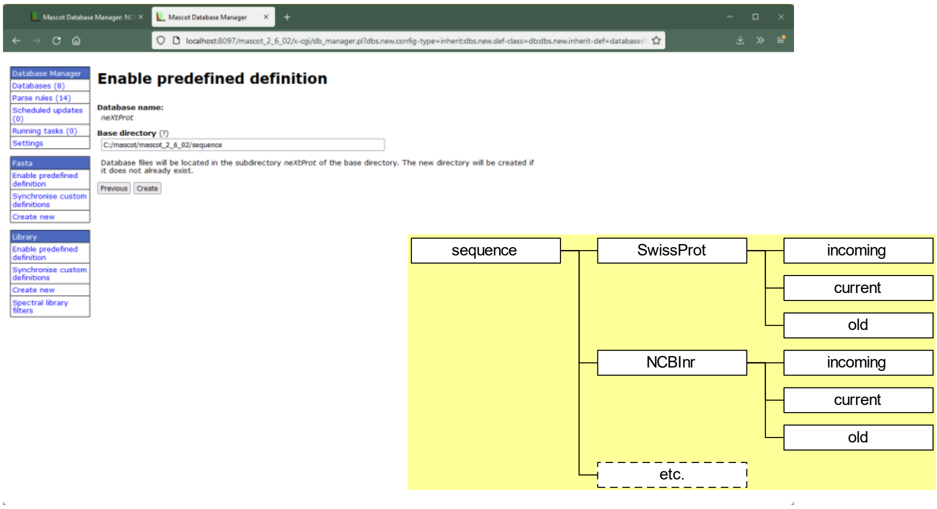
Only one instance of each definition can be enabled at any one time, as database and library definition names need to be unique. You can copy an existing definition to create more than one instance of a predefined definition.

You can also use a predefined definition as a template when creating a new definition. Such copies will not be kept up to date with the original predefined definition.

Name		
contaminants	Enable	Already set up as contaminants.
cRAP	Enable	
Environmental_EST	Enable	
Fungi_EST	Enable	
Human_EST	Enable	
Invertebrates_EST	Enable	
IPI_arabidopsis	Enable	
IPI_bovine	Enable	
IPI_chicken	Enable	
IPI_human	Enable	
IPI_mouse	Enable	
IPI_rat	Enable	
IPI_zebrafish	Enable	
Mammals_EST	Enable	
Mus_EST	Enable	
NCBIprot	Enable	
neXtProt	Enable	
Plants_EST	Enable	
Prokaryotes_EST	Enable	
Rodents_EST	Enable	
SARS-CoV-2	Enable	

Let’s look at the first of these options: enabling a predefined definition, using neXtProt as the example.
Scroll down to neXtProt and click Enable.

Enable Predefined Definition



Mascot Database Manager: NC
Mascot Database Manager
localhost:8097/mascot_2_6_02/x-cgi/db_manager.pl?dbs.new.config-type=inherit;dbs.new.def-class=db;dbs.new.inherit-def=databases
Database Manager
Databases (8)
Parse rules (14)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Previous
Create
sequence
SwissProt
incoming
current
old
NCBInr
incoming
current
old
etc.

Enable predefined definition

Database name:
neXtProt

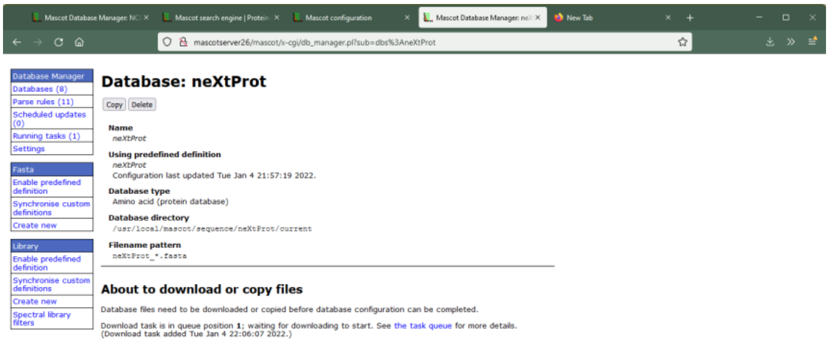
Base directory (?)
C:/mascot/mascot_2_6_02/sequence

Database files will be located in the subdirectory *neXtProt* of the base directory. The new directory will be created if it does not already exist.

The default location for the local copies of the sequence database files is specified in Database Manager settings. You can also change the location for the new database here. New directories will be created automatically, unless they already exist. For each database, there is a directory with the same name as the database. Under this directory are three sub-directories. The incoming directory provides a workspace for downloading and processing a new database file. The current directory contains the active database, and this is where Mascot Monitor creates the compressed files that will be memory mapped. The old directory is where the immediate past database files are archived … just in case.

Choose Create.

Enable Predefined Definition



Mascot Database Manager: NC
Mascot search engine | Protein
Mascot configuration
Mascot Database Manager: ne
New Tab
mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs%3AneXtProt
Database Manager
Databases (8)
Parse rules (11)
Scheduled updates (0)
Running tasks (1)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete

Database: neXtProt

Name
neXtProt

Using predefined definition
neXtProt
Configuration last updated Tue Jan 4 21:57:19 2022.

Database type
Amino acid (protein database)

Database directory
`/usr/local/mascot/sequence/neXtProt/current`

Filename pattern
`neXtProt_*.fasta`

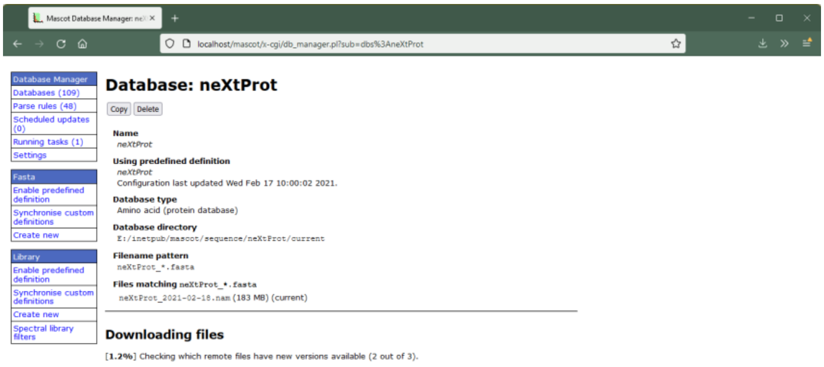
About to download or copy files

Database files need to be downloaded or copied before database configuration can be completed.

Download task is in queue position **1**; waiting for downloading to start. See the task queue for more details.
(Download task added Tue Jan 4 22:06:07 2022.)

All the files will now be downloaded automatically. For neXtProt, this is the Fasta file plus the files that are needed to create a taxonomy index. The lower part of the page is updated with status information. You don’t have to leave this page open; you can close the browser and return later.

Enable Predefined Definition


Mascot Database Manager: ne
localhost/mascot/x-cgi/db_manager.pl?sub=dbs%3AneXtProt
Database Manager
Databases (109)
Parse rules (48)
Scheduled updates (0)
Running tasks (1)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete

Database: neXtProt

Name
neXtProt

Using predefined definition
neXtProt
Configuration last updated Wed Feb 17 10:00:02 2021.

Database type
Amino acid (protein database)

Database directory
`E:/inetpub/mascot/sequence/neXtProt/current`

Filename pattern
`neXtProt_*.fasta`

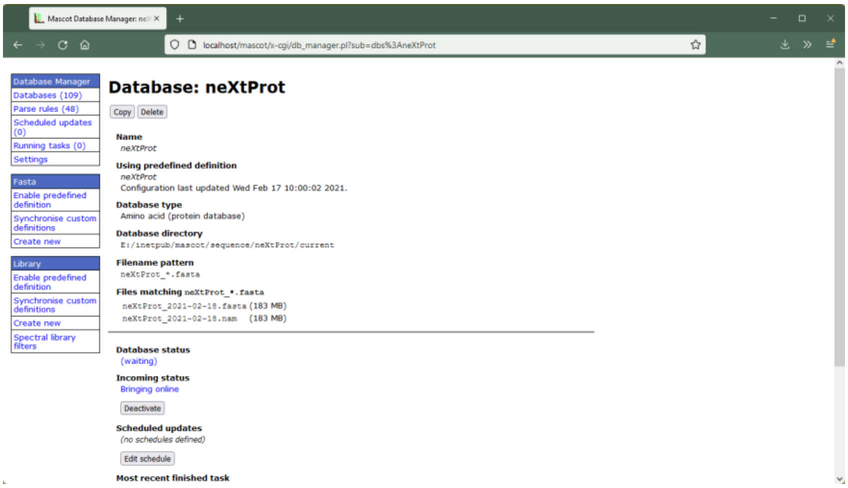
Files matching `neXtProt_*.fasta`
`neXtProt_2021-02-18.nam` (183 MB) (current)

Downloading files

[**1.2%**] Checking which remote files have new versions available (2 out of 3).

Some database files are very large, and downloads can fail for all sorts of reasons. Database Manager tries each download 5 times before giving up. If you have persistent problems, check the support page on our web site to see if there are any known issues.

Enable Predefined Definition



Mascot Database Manager: ne
localhost/mascot/x-cgi/db_manager.pl?sub=dbs%3AneXtProt
Database Manager
Databases (109)
Parse rules (48)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Deactivate
Edit schedule

Database: neXtProt

Name
neXtProt

Using predefined definition
neXtProt
Configuration last updated Wed Feb 17 10:00:02 2021.

Database type
Amino acid (protein database)

Database directory
`E:/inetpub/mascot/sequence/neXtProt/current`

Filename pattern
`neXtProt_*.fasta`

Files matching `neXtProt_*.fasta`
`neXtProt_2021-02-18.fasta` (183 MB)
`neXtProt_2021-02-18.nam` (183 MB)

Database status
(waiting)

Incoming status
Bringing online

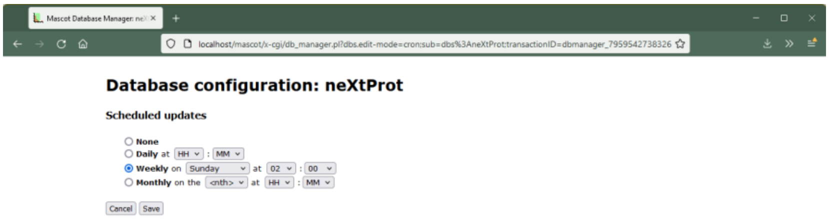
Scheduled updates
(no schedules defined)

Most recent finished task

Assuming the download is successful, this page will be displayed as the new files are compressed and the database is brought online. As soon as the new database shows as 'In Use', it is ready for searching.

To setup automatic updates of the database files, choose Edit Schedule from here or via the Scheduled Updates side menu.

Scheduled Updates



localhost/mascot/x-cgi/db_manager.pl?dbs.edit-mode=cron;sub=dbs%3AneXtProt;transactionID=dbmanager_7959542738326

Database configuration: neXtProt

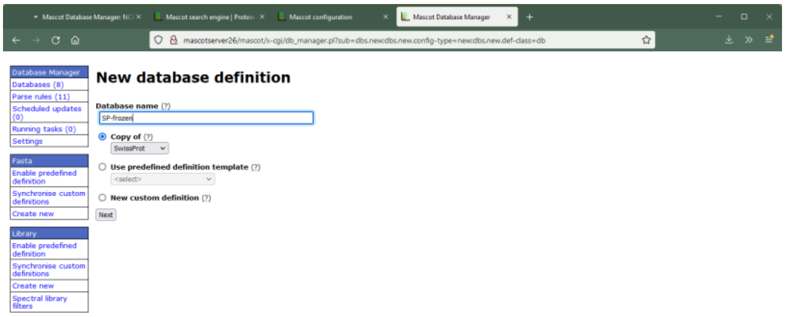
Scheduled updates

○ **None**
○ **Daily** at HH : MM
● **Weekly** on Sunday at 02 : 00
○ **Monthly** on the <nth> at HH : MM

It is usually best to download the files at a quiet time, like the middle of the night or at the weekend.

Note that keeping the definition up-to-date and keeping the database files up-to-date are two different things. A predefined database definition is kept up to date automatically while a custom database definition is not. The only requirement for keeping the files up to date is that the definition includes URLs for downloading the required files. Files are not updated by default; you have to save a schedule for the database specifying how often to look for new files. If no new Fasta file is available at the scheduled time, nothing will be downloaded.

Create New - Copy of



Mascot Database Manager: NC ×
Mascot search engine | Protein ×
Mascot configuration ×
Mascot Database Manager ×
+
mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs.new:dbs.new.config-type=new:dbs.new.def-class=db
Database Manager
Databases (8)
Parse rules (11)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

New database definition

Database name (?)
SP-frozen

● **Copy of** (?)
SwissProt

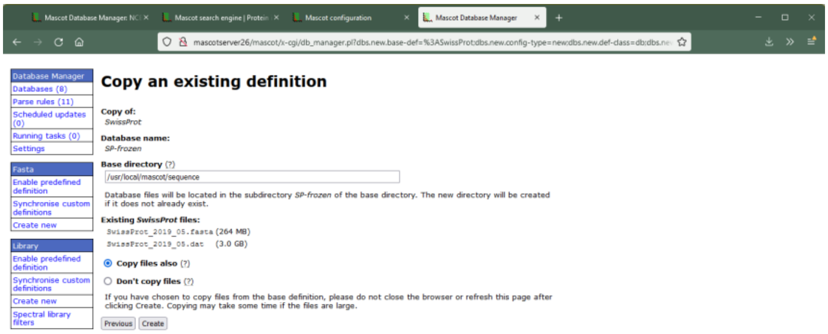
○ **Use predefined definition template** (?)
<select>

○ **New custom definition** (?)

Next

Creating a new database by starting from a copy of an existing database is usually more convenient than starting from scratch. It is also a good way to preserve a copy of a particular version of a database. Imagine you have SwissProt configured to be automatically updated every month or so. If you want to keep a copy of the current version, so that it can be used for all searches during a year long project, choose SwissProt from the drop down list and give it a suitable name so that it won’t be confused with the ‘live’ version.

Create New - Copy of


Mascot Database Manager: NC
Mascot search engine | Protein
Mascot configuration
Mascot Database Manager
mascotserver26/mascot/x-cgi/db_manager.pl?dbs.new.base-def=%3ASwissProt;dbs.new.config-type=new;dbs.new.def-class=db;dbs.ne
Database Manager
Databases (8)
Parse rules (11)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Previous
Create

Copy an existing definition

Copy of:
SwissProt

Database name:
SP-frozen

Base directory (?)
/usr/local/mascot/sequence

Database files will be located in the subdirectory *SP-frozen* of the base directory. The new directory will be created if it does not already exist.

Existing *SwissProt* files:

```
SwissProt_2019_05.fasta (264 MB)
SwissProt_2019_05.dat   (3.0 GB)
```

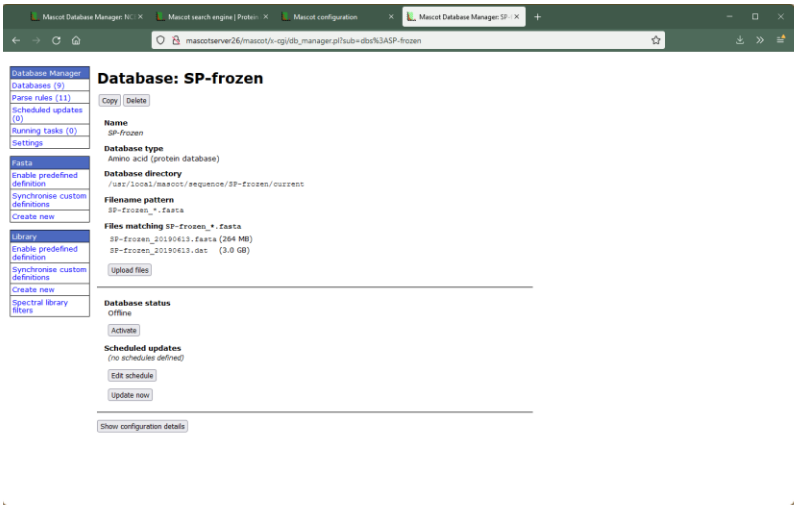
● **Copy files also** (?)

○ **Don't copy files** (?)

If you have chosen to copy files from the base definition, please do not close the browser or refresh this page after clicking Create. Copying may take some time if the files are large.

You are given the option to copy the existing files.

Create New - Copy of


Mascot Database Manager: NC
Mascot search engine | Protein
Mascot configuration
Mascot Database Manager: SP-
mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs%3ASP-frozen
Database Manager
Databases (9)
Parse rules (11)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Upload files
Activate
Edit schedule
Update now
Show configuration details

Database: SP-frozen

Name
SP-frozen

Database type
Amino acid (protein database)

Database directory
`/usr/local/mascot/sequence/SP-frozen/current`

Filename pattern
`SP-frozen_*.fasta`

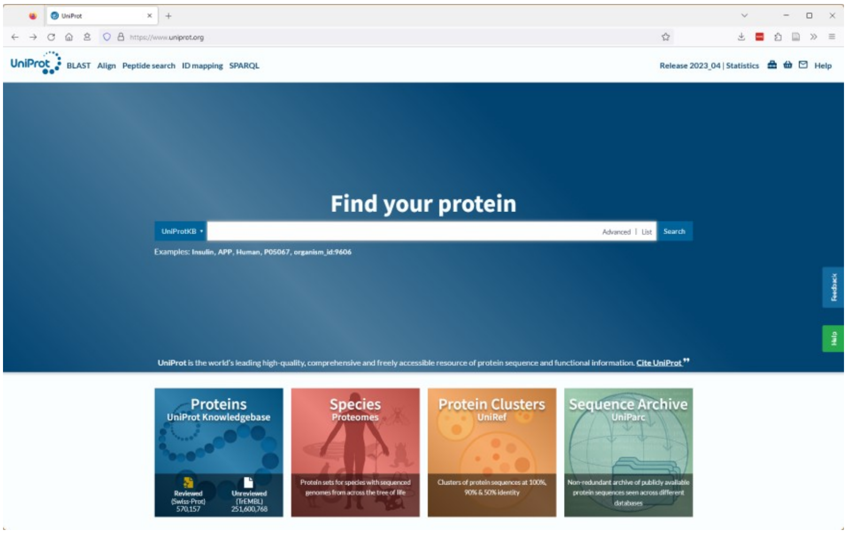
Files matching `SP-frozen_*.fasta`
`SP-frozen_20190613.fasta` (264 MB)
`SP-frozen_20190613.dat` (3.0 GB)

Database status
Offline

Scheduled updates
(no schedules defined)

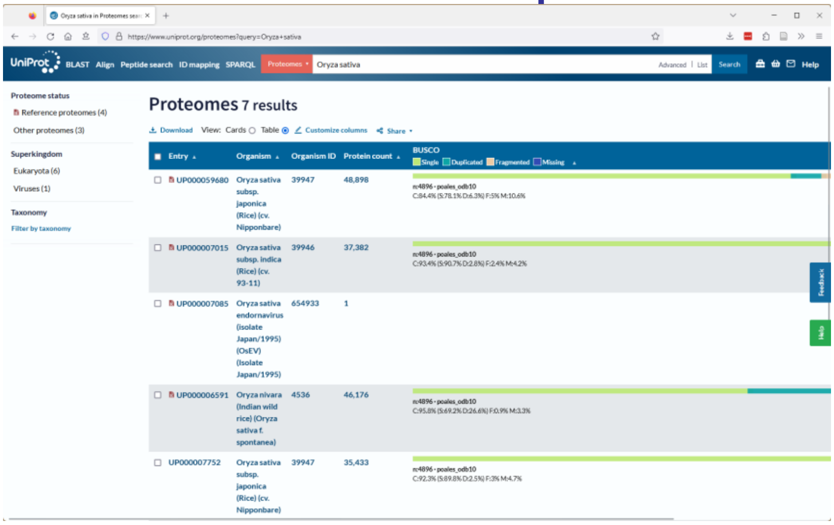
Unless you wish to make some change to the configuration, all you need to do now is to choose Activate. When submitting searches for the year-long project, you choose SP-frozen rather than SwissProt.

Create New – Predefined as Template


UniProt
https://www.uniprot.org
UniProt
BLAST
Align
Peptide search
ID mapping
SPARQL
Release 2023_04 | Statistics
Help
Find your protein
UniProtKB
Advanced | List
Search
Examples: Insulin, APP, Human, P05067, organism_id:9606
Feedback
Help
UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt
Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 570,157
Unreviewed (TrEMBL) 251,600,768
Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life
Protein Clusters
UniRef
Clusters of protein sequences at 100%, 90% & 50% identity
Sequence Archive
UniParc
Non-redundant archive of publicly available protein sequences seen across different databases

To illustrate how a predefined definition can be used as a template, we'll set up a database for the Uniprot proteome of rice. In a browser, go to the Uniprot web site, www.uniprot.org, and follow the 'Complete Proteome' links. You could search on rice or, if you remember the latin name, oryza.

Create New – Predefined as Template



Proteome status

- Reference proteomes (4)
- Other proteomes (3)

Superkingdom

- Eukaryota (6)
- Viruses (1)

Taxonomy

Filter by taxonomy

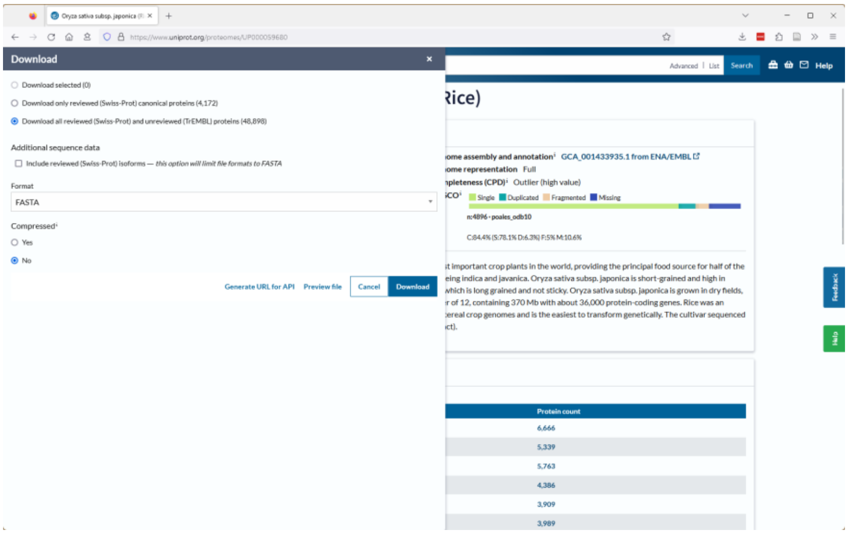
Proteomes 7 results

Download View: Cards ○ Table ● Customize columns Share ▾

■	Entry ▴	Organism ▴	Organism ID	Protein count ▴	BUSCO 🟩Single 🟦Duplicated 🟧Fragmented 🟪Missing ▴
□	UP000059680	Oryza sativa subsp. japonica (Rice) (cv. Nipponbare)	39947	48,898	n:4896 - poales_odb10 C:84.4% (S:78.1% D:6.3%) F:5% M:10.6%
□	UP000007015	Oryza sativa subsp. indica (Rice) (cv. 93-11)	39946	37,382	n:4896 - poales_odb10 C:93.4% (S:90.7% D:2.8%) F:2.4% M:4.2%
□	UP000007085	Oryza sativa endornavirus (isolate Japan/1995) (OsEV) (Isolate Japan/1995)	654933	1	
□	UP000006591	Oryza nivara (Indian wild rice) (Oryza sativa f. spontanea)	4536	46,176	n:4896 - poales_odb10 C:95.8% (S:69.2% D:26.6%) F:0.9% M:3.3%
□	UP000007752	Oryza sativa subsp. japonica (Rice) (cv. Nipponbare)	39947	35,433	n:4896 - poales_odb10 C:92.3% (S:89.8% D:2.5%) F:3% M:4.7%

There is a choice of fourteen including a number of virus that infects rice. We’ll choose *Oryza sativa subsp. japonica* with Proteome ID UP000059680.

Create New - Predefined as Template



Oryza sativa subsp. japonica (R

https://www.uniprot.org/proteomes/UP000059680

Download

- ○ Download selected (0)
- ○ Download only reviewed (Swiss-Prot) canonical proteins (4,172)
- ● Download all reviewed (Swiss-Prot) and unreviewed (TrEMBL) proteins (48,898)

Additional sequence data

☐ Include reviewed (Swiss-Prot) isoforms — *this option will limit file formats to FASTA*

Format: FASTA

Compressed

- ○ Yes
- ● No

Rice)

ome assembly and annotation GCA_001433935.1 from ENA/EMBL

ome representation Full

npleteness (CPD) Outlier (high value)

SCO 🟩 Single 🟦 Duplicated 🟧 Fragmented 🟦 Missing

n:4896 · poales_odb10

C:84.4% (S:78.1% D:6.3%) F:5% M:10.6%

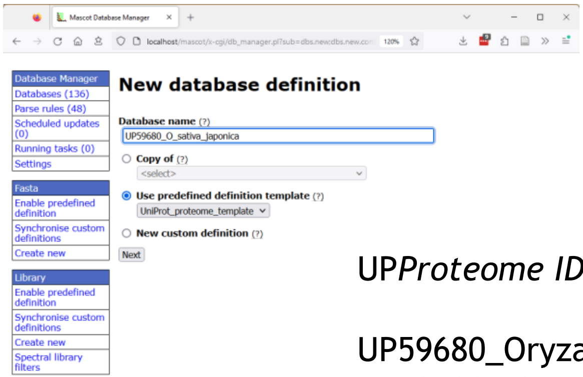
st important crop plants in the world, providing the principal food source for half of the
eing indica and javanica. Oryza sativa subsp. japonica is short-grained and high in
which is long grained and not sticky. Oryza sativa subsp. japonica is grown in dry fields,
r of 12, containing 370 Mb with about 36,000 protein-coding genes. Rice was an
ereal crop genomes and is the easiest to transform genetically. The cultivar sequenced
ct).

Protein count
6,666
5,339
5,763
4,386
3,909
3,989

You could download the file manually. If so, click on the proteome ID link, then a download button, and select all proteins in Fasta format.

A better option is to set up a download URL, because this will allow us to configure automatic updating of the database files. First, make a note of the proteome ID for this strain – 59680.

Create New - Predefined as Template


Mascot Database Manager
localhost/mascot/x-cgi/db_manager.pl?sub=dbs.new;dbs.new.con
120%
Database Manager
Databases (136)
Parse rules (48)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

New database definition

Database name (?)
UP59680_O_sativa_japonica

○ Copy of (?)
<select>

● Use predefined definition template (?)
UniProt_proteome_template

○ New custom definition (?)

Next

UP*Proteome ID_Species_name*

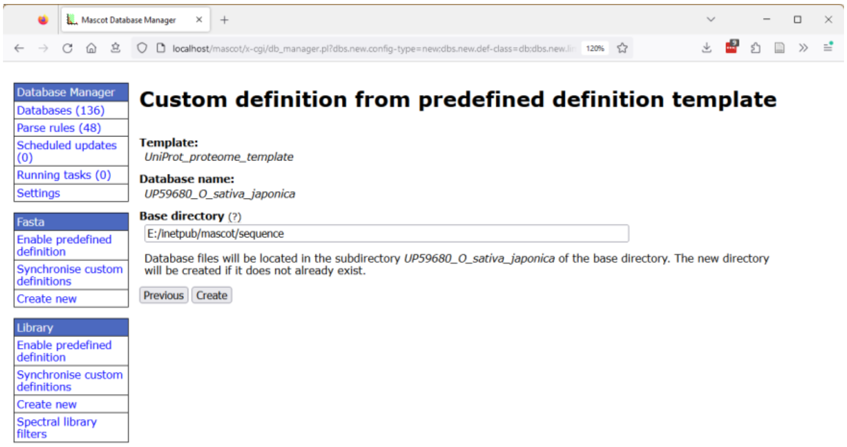
UP59680_Oryza_sativa
UP59680_ Oryza _sativa_japonica
UP59680_O_sativa_japonica

In Database Manager, choose Create new. Enter a suitable name.

For Uniprot database Mascot Server is currently using the naming convention “UP Proteome ID underscore Species name”. And typically, just the first letter of the genus is used along with the species epithet. In this example it is a proteome for a sub species and there is a trinomen rather than a binomial name. Any of these three examples would be suitable. I have used the last example to be consistent with the other uniprot proteome databases. You can of course call it what ever name you like as long as there are no spaces in it.

Next select the uniprot proteome predefined definition as a template.

Create New - Predefined as Template



Mascot Database Manager
localhost/mascot/x-cgi/db_manager.pl?dbs.new.config-type=new;dbs.new.def-class=db;dbs.new.li
120%
Database Manager
Databases (136)
Parse rules (48)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Custom definition from predefined definition template

Template:
UniProt_proteome_template

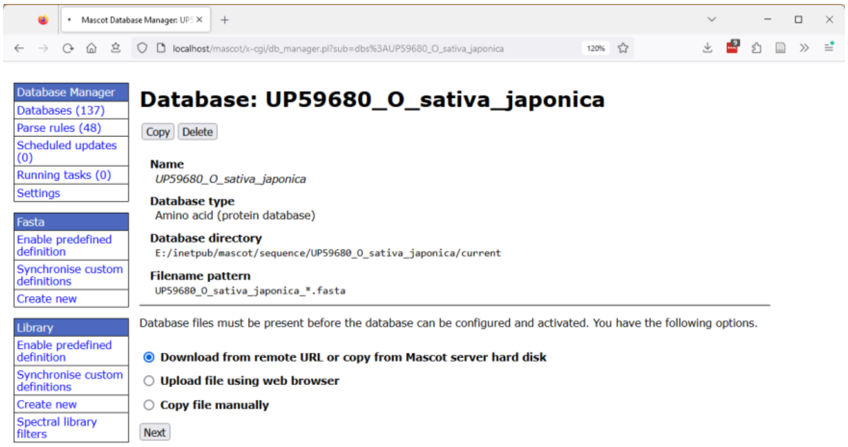
Database name:
UP59680_O_sativa_japonica

Base directory (?)
E:/inetpub/mascot/sequence

Database files will be located in the subdirectory *UP59680_O_sativa_japonica* of the base directory. The new directory will be created if it does not already exist.

Choose Create.

Create New - Predefined as Template


Mascot Database Manager: UP5
localhost/mascot/x-cgi/db_manager.pl?sub=dbs%3AUP59680_O_sativa_japonica
120%
Database Manager
Databases (137)
Parse rules (48)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Next

Database: UP59680_O_sativa_japonica

Name
UP59680_O_sativa_japonica

Database type
Amino acid (protein database)

Database directory
`E:/inetpub/mascot/sequence/UP59680_O_sativa_japonica/current`

Filename pattern
`UP59680_O_sativa_japonica_*.fasta`

Database files must be present before the database can be configured and activated. You have the following options.

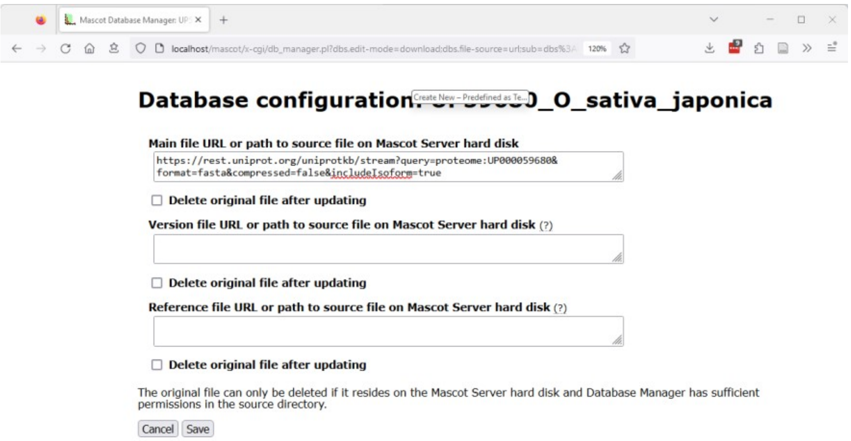
● **Download from remote URL or copy from Mascot server hard disk**

○ **Upload file using web browser**

○ **Copy file manually**

If we had chosen to download the Fasta manually, we could follow the instructions to upload or copy the Fasta file to the target directory and rename it to match the Filename pattern. Since we want to schedule automatic file updates, we choose instead to download from a remote URL and, in the next page, choose Setup download URL.

Create New – Predefined as Template



Mascot Database Manager: UP
localhost/mascot/x-cgi/db_manager.pl?dbs.edit-mode=download;dbs.file-source=url;sub=dbs%3A
120%
Create New – Predefined as Te...

Database configuration 0_O_sativa_japonica

Main file URL or path to source file on Mascot Server hard disk

```
https://rest.uniprot.org/uniprotkb/stream?query=proteome:UP000059680&
format=fasta&compressed=false&includeIsoform=true
```

□ **Delete original file after updating**

Version file URL or path to source file on Mascot Server hard disk (?)

□ **Delete original file after updating**

Reference file URL or path to source file on Mascot Server hard disk (?)

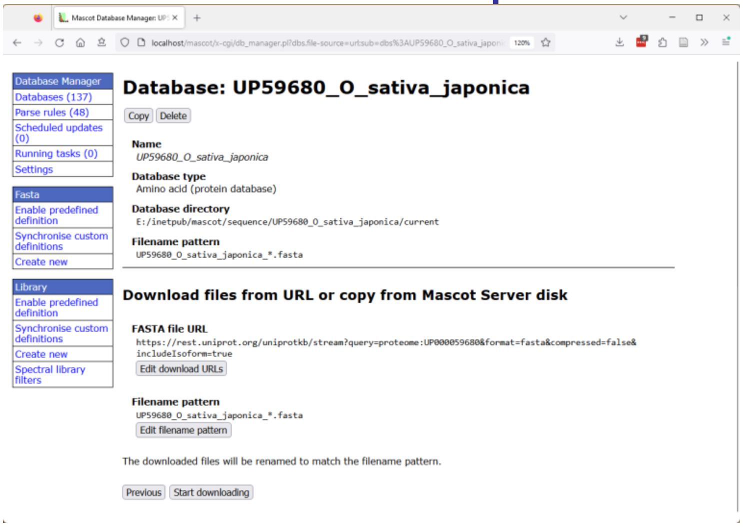
□ **Delete original file after updating**

The original file can only be deleted if it resides on the Mascot Server hard disk and Database Manager has sufficient permissions in the source directory.

Cancel | Save

The format for the URL can be found on the Mascot help page for UniProt proteomes under Sequence database setup. You just need to change the UniProt proteome ID to the one for rice that you noted earlier. There is no reference file to download … we'll link out to Uniprot to get annotation text for the Mascot Protein View report. There is no version file either, so each update will be identified using an ISO datestamp.

Create New - Predefined as Template



Mascot Database Manager: UP
localhost/mascot/x-cgi/db_manager.pl?dbs.file-source=url:sub=dbs%3AUP59680_O_sativa_japoni
120%
Database Manager
Databases (137)
Parse rules (48)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Edit download URLs
Edit filename pattern
Previous
Start downloading

Database: UP59680_O_sativa_japonica

Name
UP59680_O_sativa_japonica

Database type
Amino acid (protein database)

Database directory
`E:/inetpub/mascot/sequence/UP59680_O_sativa_japonica/current`

Filename pattern
`UP59680_O_sativa_japonica_*.fasta`

Download files from URL or copy from Mascot Server disk

FASTA file URL

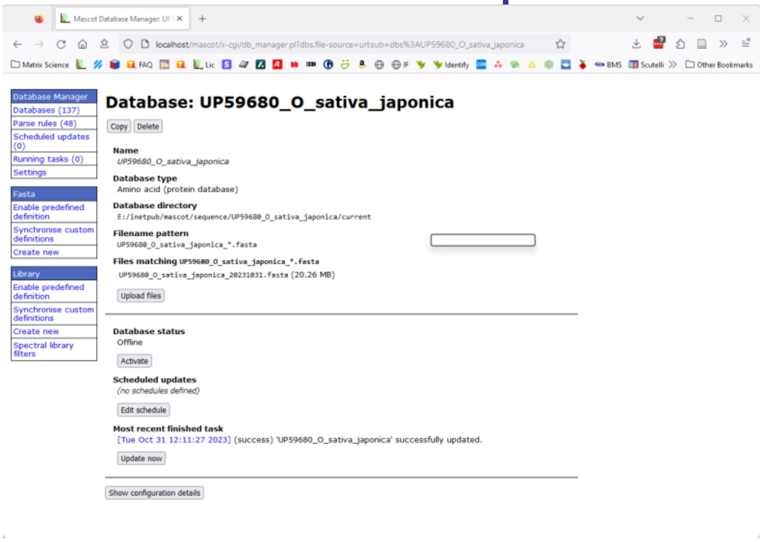
```
https://rest.uniprot.org/uniprotkb/stream?query=proteome:UP000059680&format=fasta&compressed=false&
includeIsoform=true
```

Filename pattern
`UP59680_O_sativa_japonica_*.fasta`

The downloaded files will be renamed to match the filename pattern.

Save, and we’re ready to start downloading.

Create New - Predefined as Template



Mascot Database Manager: UP5
localhost/mascot/x-cgi/db_manager.pl?dbs.file-source=url:sub=dbs%3AUP59680_O_sativa_japonica
Matrix Science
FAQ
Lic
IF
Identify
BMS
Scutelli
Other Bookmarks
Database Manager
Databases (137)
Parse rules (48)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Upload files
Activate
Edit schedule
Update now
Show configuration details

Database: UP59680_O_sativa_japonica

Name
UP59680_O_sativa_japonica

Database type
Amino acid (protein database)

Database directory
`E:/inetpub/mascot/sequence/UP59680_O_sativa_japonica/current`

Filename pattern
`UP59680_O_sativa_japonica_*.fasta`

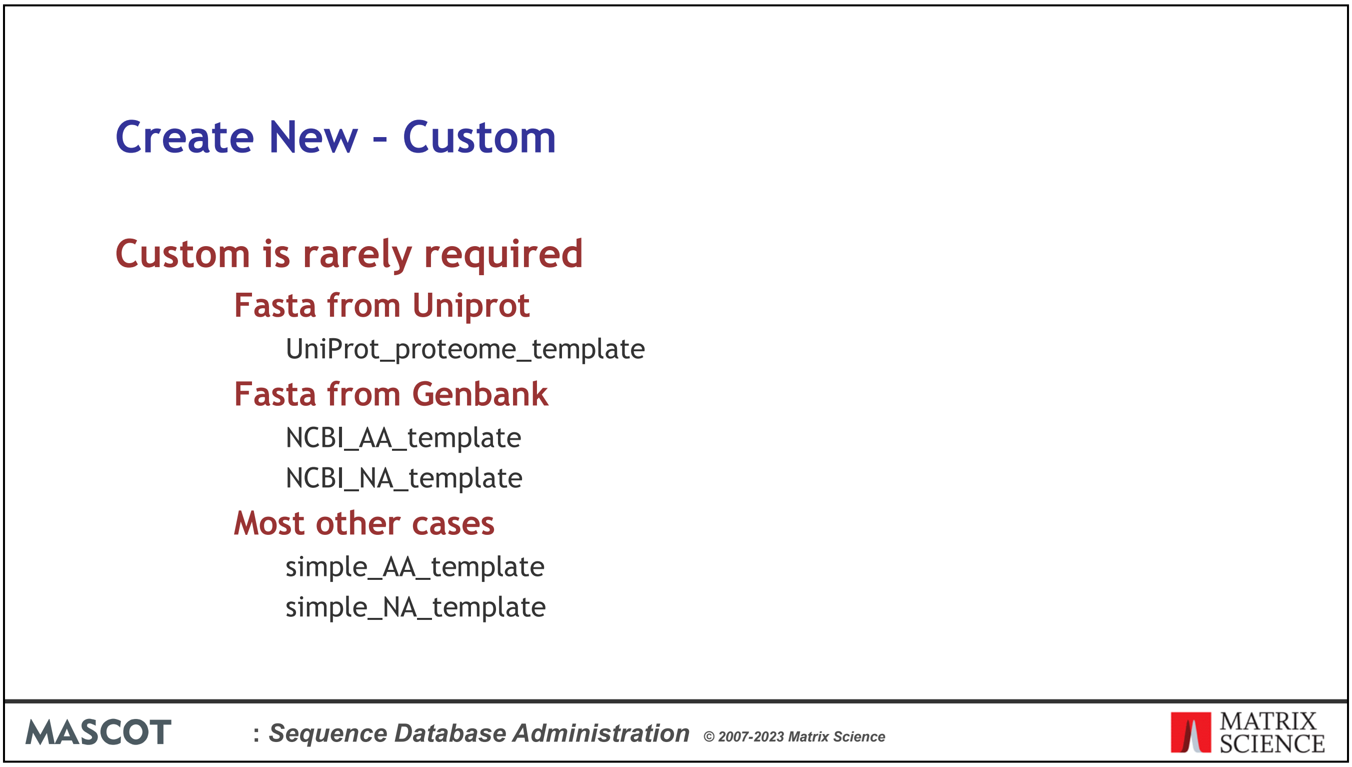
Files matching `UP59680_O_sativa_japonica_*.fasta`
`UP59680_O_sativa_japonica_20231031.fasta` (20.26 MB)

Database status
Offline

Scheduled updates
(no schedules defined)

Most recent finished task
[Tue Oct 31 12:11:27 2023] (success) 'UP59680_O_sativa_japonica' successfully updated.

Once the download is complete, the page is updated. Assuming we don’t want to inspect or modify the configuration, just two things left to do. Make the database active in Mascot and create an update schedule.



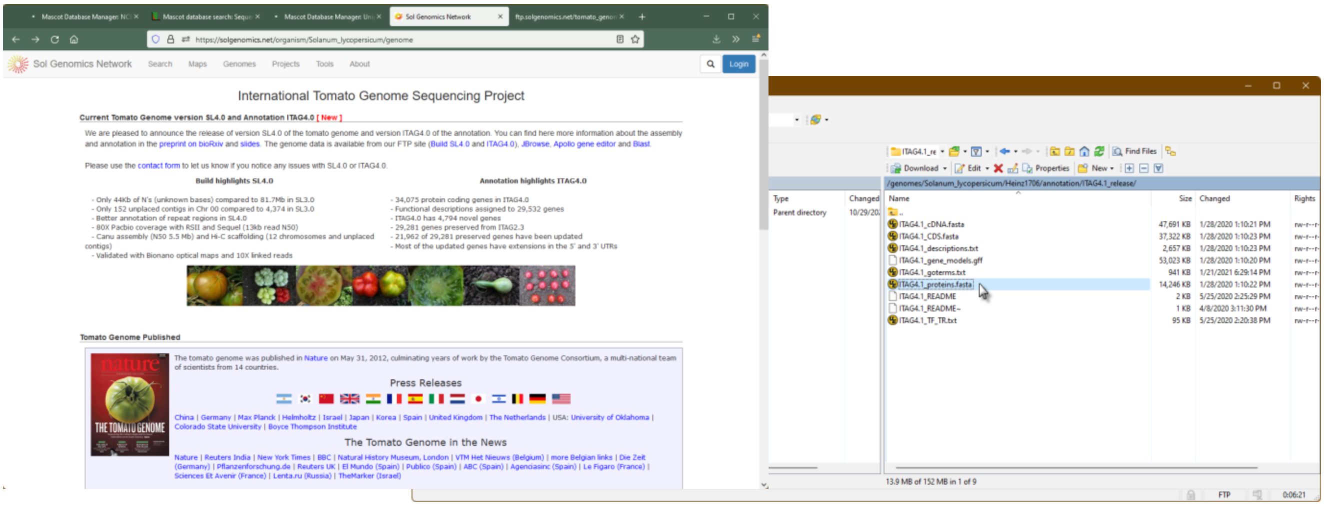
Create New - Custom

Custom is rarely required

- Fasta from Uniprot
  - UniProt_proteome_template
- Fasta from Genbank
  - NCBI_AA_template
  - NCBI_NA_template
- Most other cases
  - simple_AA_template
  - simple_NA_template

In most cases, it is faster to start from one of the predefined definitions, and modify it, than choose custom.

Create New - Custom



https://solgenomics.net/organism/Solanum_lycopersicum/genome

Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager: Uni
Sol Genomics Network
ftp.solgenomics.net/tomato_genon
Sol Genomics Network
Search
Maps
Genomes
Projects
Tools
About
Login

International Tomato Genome Sequencing Project

Current Tomato Genome version SL4.0 and Annotation ITAG4.0 [New]

We are pleased to announce the release of version SL4.0 of the tomato genome and version ITAG4.0 of the annotation. You can find here more information about the assembly and annotation in the preprint on bioRxiv and slides. The genome data is available from our FTP site (Build SL4.0 and ITAG4.0), JBrowse, Apollo gene editor and Blast.

Please use the contact form to let us know if you notice any issues with SL4.0 or ITAG4.0.

Build highlights SL4.0

- Only 44Kb of N's (unknown bases) compared to 81.7Mb in SL3.0
- Only 152 unplaced contigs in Chr 00 compared to 4,374 in SL3.0
- Better annotation of repeat regions in SL4.0
- 80X Pacbio coverage with RSII and Sequel (13kb read N50)
- Canu assembly (N50 5.5 Mb) and Hi-C scaffolding (12 chromosomes and unplaced contigs)
- Validated with Bionano optical maps and 10X linked reads

Annotation highlights ITAG4.0

- 34,075 protein coding genes in ITAG4.0
- Functional descriptions assigned to 29,532 genes
- ITAG4.0 has 4,794 novel genes
- 29,281 genes preserved from ITAG2.3
- 21,962 of 29,281 preserved genes have been updated
- Most of the updated genes have extensions in the 5' and 3' UTRs

Tomato Genome Published

The tomato genome was published in Nature on May 31, 2012, culminating years of work by the Tomato Genome Consortium, a multi-national team of scientists from 14 countries.

Press Releases

China | Germany | Max Planck | Helmholtz | Israel | Japan | Korea | Spain | United Kingdom | The Netherlands | USA: University of Oklahoma | Colorado State University | Boyce Thompson Institute

The Tomato Genome in the News

Nature | Reuters India | New York Times | BBC | Natural History Museum, London | VTM Het Nieuws (Belgium) | more Belgian links | Die Zeit (Germany) | Pflanzenforschung.de | Reuters UK | El Mundo (Spain) | Publico (Spain) | ABC (Spain) | Agenciasinc (Spain) | Le Figaro (France) | Sciences Et Avenir (France) | Lenta.ru (Russia) | TheMarker (Israel)

/genomes/Solanum_lycopersicum/Heinz1706/annotation/ITAG4.1_release/

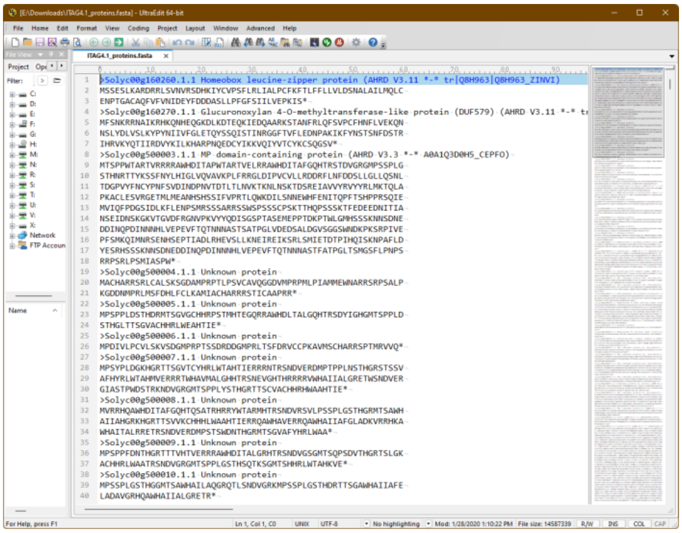
Type	Changed
Parent directory	10/29/20

Name	Size	Changed	Rights
..			
ITAG4.1_cDNA.fasta	47,691 KB	1/28/2020 1:10:21 PM	rw-r--r-
ITAG4.1_CDS.fasta	37,322 KB	1/28/2020 1:10:23 PM	rw-r--r-
ITAG4.1_descriptions.txt	2,657 KB	1/28/2020 1:10:23 PM	rw-r--r-
ITAG4.1_gene_models.gff	53,023 KB	1/28/2020 1:10:20 PM	rw-r--r-
ITAG4.1_goterms.txt	941 KB	1/21/2021 6:29:14 PM	rw-r--r-
ITAG4.1_proteins.fasta	14,246 KB	1/28/2020 1:10:22 PM	rw-r--r-
ITAG4.1_README	2 KB	5/25/2020 2:25:29 PM	rw-r--r-
ITAG4.1_README~	1 KB	4/8/2020 3:11:30 PM	rw-r--r-
ITAG4.1_TF_TR.txt	95 KB	5/25/2020 2:20:38 PM	rw-r--r-

13.9 MB of 152 MB in 1 of 9

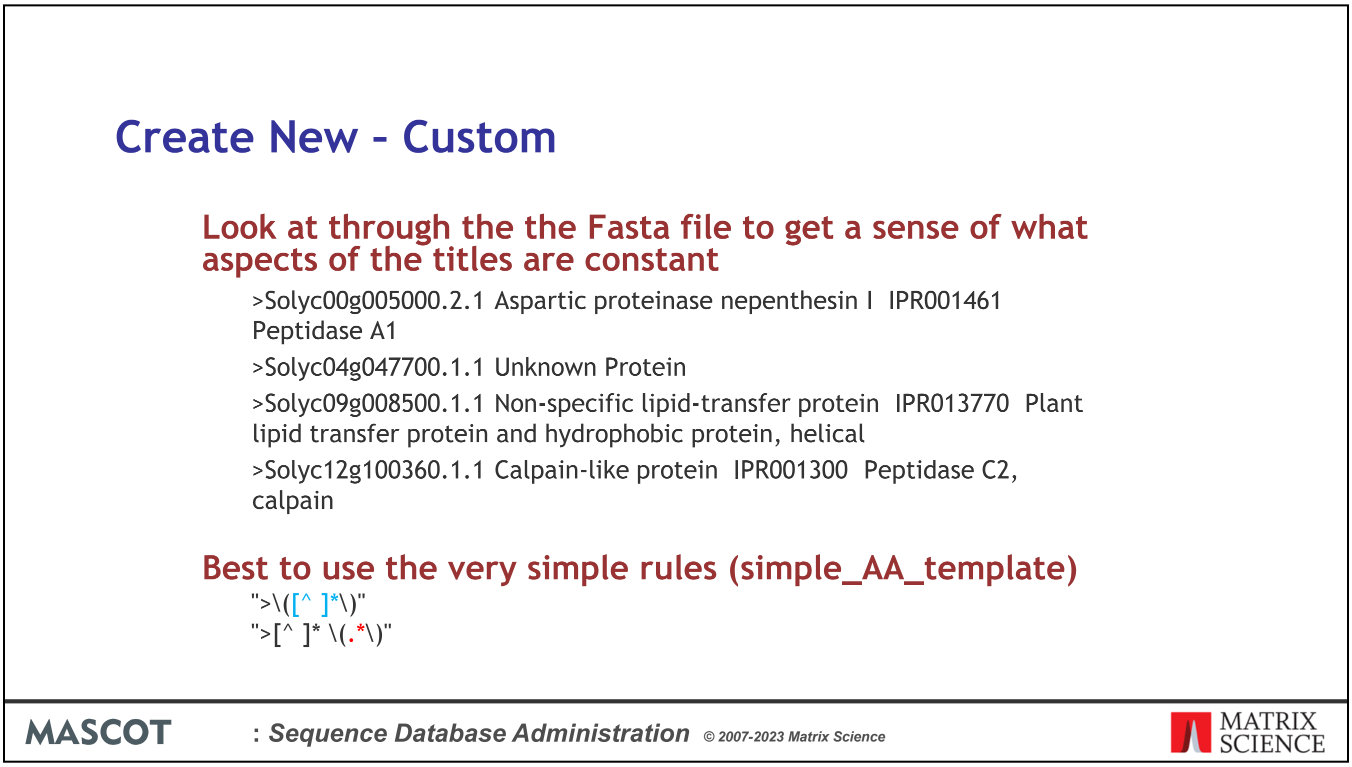
But, to illustrate, lets configure a protein Fasta for the most recent version of the tomato genome. Choose a file from the FTP site and download it.

Create New - Custom


[E:\Downloads\ITAG4.1_proteins.fasta] - UltraEdit 64-bit
File Home Edit Format View Coding Project Layout Window Advanced Help
ITAG4.1_proteins.fasta
Project
Filter:
Network
FTP Account
Name
For Help, press F1
Ln 1, Col 1, C0
UNIX
UTF-8
No highlighting
Mod: 1/28/2020 1:10:22 PM
File size: 14587339
R/W
INS
COL
CAP

```
>Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)
MSSESLKARDRRLSVNVRSDHKIYCVPSFLRLIALPCFKFTLFFLLVLDSNALAILMQLC
ENPTGACAQFVFVNIDEYFDDDASLLPFGFSIILVEPKIS*
>Solyc00g160270.1.1 Glucuronoxylan 4-O-methyltransferase-like protein (DUF579) (AHRD V3.11 *-* tr
MFSNKRRNAIKRHKQNHEQGKDLKDTEQKIEDQAARKSTANFRLQFSVPCFHNFLVEKQN
NSLYDLVSLKYPYNIIVFGLETQYSSQISTINRGGFTVFLEDNPAKIKFYNSTSNFDSTR
IHRVKYQTIIRDVYKILKHARPNQEDCYIKKVQIYVTCYKCSQGSV*
>Solyc00g500003.1.1 MP domain-containing protein (AHRD V3.3 *-* A0A1Q3D0H5_CEPFO)
MTSPPWTARTVRRRRAWHDITAPWTARTVELRRAWHDITAFGQHTRSTDVGRGMPSSPLG
STHNRTTYKSSFNYLHIGLVQVAVKPLFRRGLDIPVCVLLRDDRFLNFDDSLLGLLQSNL
TDGPVYFNCYPNFSVDINDPNVTDTLTLNVKTKNLNSKTDSREIAVVYRVYYRLMKTQLA
PKACLESVRGETMLMEANHSHSSIFVPRTLQWKDILSNNEWHFENITQPFTSHPPRSQIE
MVIQFPDGSIDLKFLENPSMRSSSARRSSWSPSSSCPSKTTHQPSSSKTFEDEEDNITIA
NSEIDNSKGKVTGVDFRGNVPKVYYQDISGSPTASEMEPPTDKPTWLGMHSSSKNNSDNE
DDINQPDINNNHLVEPEVFTQTNNNASTSATPGLVDEDSALDGVSGGSWNDKPKSRPIVE
PFSMKQIMNRSENHSEPTIADLRHEVSLLKNEIREIKSRLSMIETDTPIHQISKNPAFLD
YESRHSSSKNNSDNEDDINQPDINNNHLVEPEVFTQTNNNASTFATPGLTSMGSFLPNPS
RRPSRLPSMIASPW*
>Solyc00g500004.1.1 Unknown protein
MACHARRSRLCALSKSGDAMPRPTLPSVCAVQGGDVMPRPMLPIAMMEWNARRSRPSALP
KGDDNMPRLMSFDHLFCLKAMIACHARRRSTICAAPRR*
>Solyc00g500005.1.1 Unknown protein
MPSPPLDSTHDRMTSGVGCHHRPSTMHTEGQRRAWHDLTALGQHTRSDYIGHGMTSPPLD
STHGLTTSGVACHHRLWEAHTIE*
>Solyc00g500006.1.1 Unknown protein
MPDIVLPCVLSKVSDGMPRPTSSDRDDGMPRLTSFDRVCCPKAVMSCHARRSPTMRVVQ*
>Solyc00g500007.1.1 Unknown protein
MPSYPLDGKHGRTTSGVTCYHRLWTAHTIERRRNTRSNDVERDMPTPPLNSTHGRSTSSV
AFHYRLWTAHMVERRRTWHAVMALGHHTRSNEVGHTHRRRRVWHAIIALGRETWSNDVER
GIASTPWDSTRKNDVGRGMTSPPLYSTHGRTTSCVACHHRHWAAHTIE*
>Solyc00g500008.1.1 Unknown protein
MVRRHQAWHDITAFGQHTQSATRHRRYWTARMHTRSNDVRSVLPSSPLGSTHGRMTSAWH
AIIAHGRKHGRTTSVVKCHHHLWAAHTIERRQAWHAVERRQAWHAIIAFGLADKVRRHKA
WHAITALRRETRSNDVERDMPSTSWDNTHGRMTSGVAFYHRLWAA*
>Solyc00g500009.1.1 Unknown protein
MPSPPFDNTHGRTTTVHTVERRRAWHDITALGRHTRSNDVGSGMTSQPSDVTHGRTSLGK
ACHHRLWAATRSNDVGRGMTSPPLGSTHSQTKSGMTSHHRLWTAHKVE*
>Solyc00g500010.1.1 Unknown protein
MPSSPLGSTHGGMTSAWHAILAQGRQTLSNDVGRKMPSSPLGSTHDRTTSGAWHAIIAFE
LADAVGRHQAWHAIIALGRETR*
```

The first thing to do with any unknown Fasta file is open it in a text editor that can handle large files and take a look at the syntax of the title lines. If you don’t have a suitable text editor, you can use more at a command prompt. Make sure you look at different places in the file because it may be a merge of entries from different sources.



Create New - Custom

Look at through the the Fasta file to get a sense of what aspects of the titles are constant

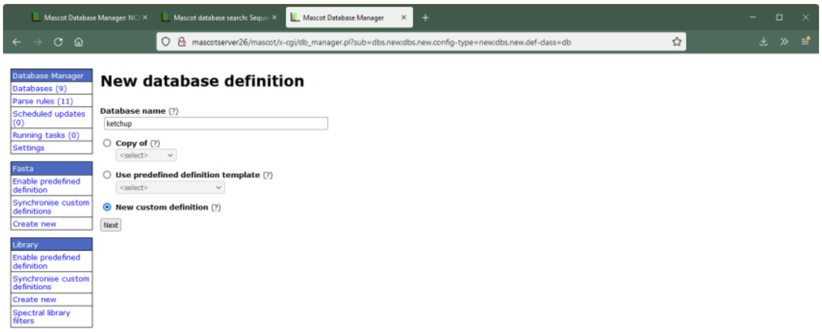
```
>Solyc00g005000.2.1 Aspartic proteinase nepenthesin I  IPR001461
Peptidase A1
>Solyc04g047700.1.1 Unknown Protein
>Solyc09g008500.1.1 Non-specific lipid-transfer protein  IPR013770  Plant
lipid transfer protein and hydrophobic protein, helical
>Solyc12g100360.1.1 Calpain-like protein  IPR001300  Peptidase C2,
calpain
```

Best to use the very simple rules (simple_AA_template)

```
">\([^ ]*\)"
">[^ ]* \(.*\)"
```

As is often the case, a simple rule that takes everything between the ">" symbol and the first space as the accession is the safest choice. Everything after the first space can be treated as the description. These rules are pre-defined in Database Manager and you could set up this database by using simple_AA_template as the template. But, to illustrate the flexibility of Database Manager, we'll follow the custom definition route, and even create a new parse rule.

Create New - Custom



Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager

mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs.new:dbs.new.config-type=new:dbs.new.def-class=db

Database Manager
Databases (9)
Parse rules (11)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

New database definition

Database name (?)
ketchup

○ Copy of (?)
<select>

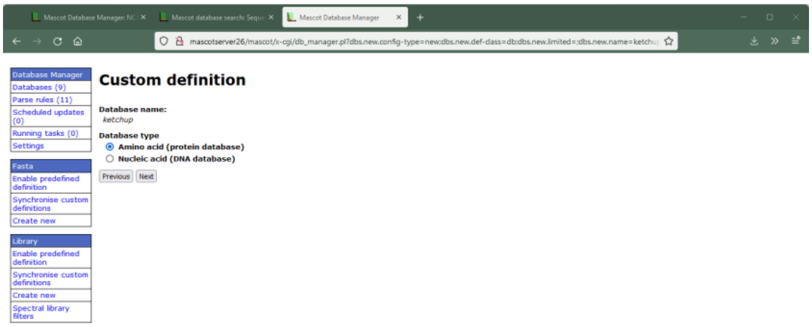
○ Use predefined definition template (?)
<select>

● New custom definition (?)

Next

So, we Create new, enter a suitable name, and choose Custom.

Create New - Custom



mascotserver26/mascot/x-cgi/db_manager.pl?dbs.new.config-type=new;dbs.new.def-class=db;dbs.new.limited=;dbs.new.name=ketchup

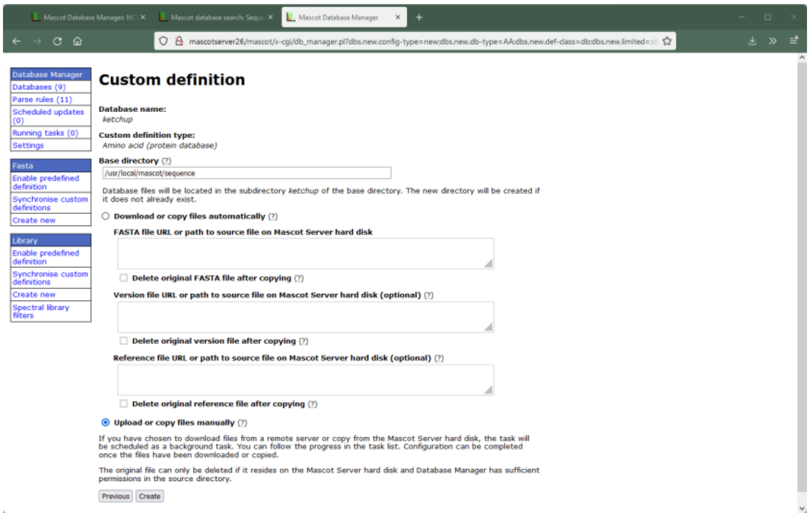
Custom definition

Database name:
ketchup

Database type
● **Amino acid (protein database)**
○ **Nucleic acid (DNA database)**

This is an Amino Acid database and we already have the Fasta file. Next.

Create New - Custom



Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager
mascotserver26/mascot/x-cgi/db_manager.pl?dbs.new.config-type=new;dbs.new.db-type=AA;dbs.new.def-class=db;dbs.new.limited=
Database Manager
Databases (9)
Parse rules (11)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Custom definition

Database name:
ketchup

Custom definition type:
Amino acid (protein database)

Base directory (?)
/usr/local/mascot/sequence

Database files will be located in the subdirectory *ketchup* of the base directory. The new directory will be created if it does not already exist.

○ **Download or copy files automatically** (?)

FASTA file URL or path to source file on Mascot Server hard disk

□ **Delete original FASTA file after copying** (?)

Version file URL or path to source file on Mascot Server hard disk (optional) (?)

□ **Delete original version file after copying** (?)

Reference file URL or path to source file on Mascot Server hard disk (optional) (?)

□ **Delete original reference file after copying** (?)

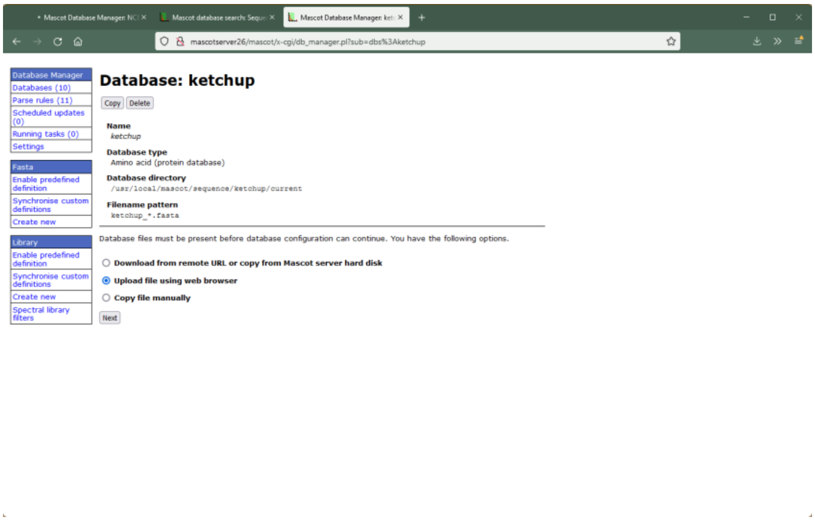
● **Upload or copy files manually** (?)

If you have chosen to download files from a remote server or copy from the Mascot Server hard disk, the task will be scheduled as a background task. You can follow the progress in the task list. Configuration can be completed once the files have been downloaded or copied.

The original file can only be deleted if it resides on the Mascot Server hard disk and Database Manager has sufficient permissions in the source directory.

Set these options and choose Create.

Create New – Custom


Mascot Database Manager: NC ×
Mascot database search: Seque ×
Mascot Database Manager: ket ×
mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs%3Aketchup
Database Manager
Databases (10)
Parse rules (11)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Next

Database: ketchup

Name
ketchup

Database type
Amino acid (protein database)

Database directory
`/usr/local/mascot/sequence/ketchup/current`

Filename pattern
`ketchup_*.fasta`

Database files must be present before database configuration can continue. You have the following options.

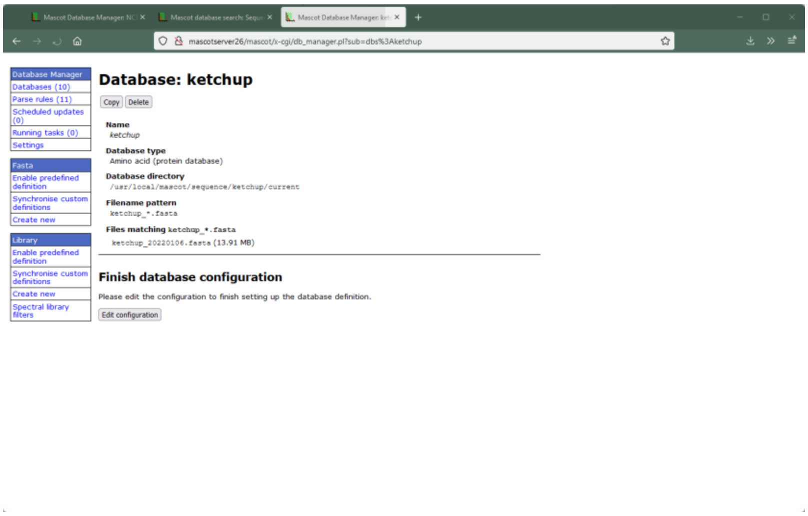
○ **Download from remote URL or copy from Mascot server hard disk**

● **Upload file using web browser**

○ **Copy file manually**

We’ve chosen to rename the Fasta file and upload it via the web browser.

Create New - Custom


Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager: ket
mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs%3Aketchup
Database Manager
Databases (10)
Parse rules (11)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Edit configuration

Database: ketchup

Name
ketchup

Database type
Amino acid (protein database)

Database directory
`/usr/local/mascot/sequence/ketchup/current`

Filename pattern
`ketchup_*.fasta`

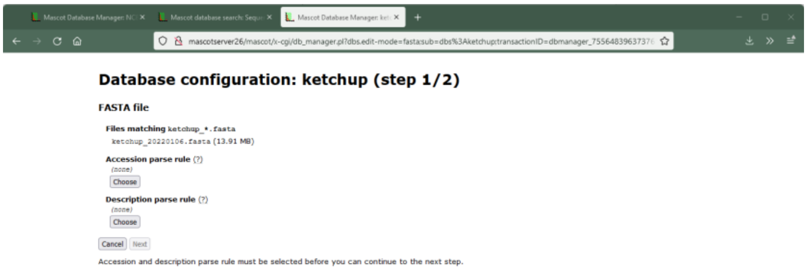
Files matching `ketchup_*.fasta`
`ketchup_20220106.fasta` (13.91 MB)

Finish database configuration

Please edit the configuration to finish setting up the database definition.

Once the file is uploaded and the name matches the wild card pattern, it will be recognized, and we can proceed to edit the configuration. If Database Manager doesn't recognize the presence of the Fasta, maybe there is a typo in the database name or maybe you have the file permissions / security settings set so that a CGI process cannot read the file.

Create New - Custom



Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager: ket
mascotserver26/mascot/x-cgi/db_manager.pl?dbs.edit-mode=fasta;sub=dbs%3Aketchup;transactionID=dbmanager_75564839637376

Database configuration: ketchup (step 1/2)

FASTA file

Files matching `ketchup_*.fasta`
`ketchup_20220106.fasta` (13.91 MB)

Accession parse rule (?)
(none)
Choose

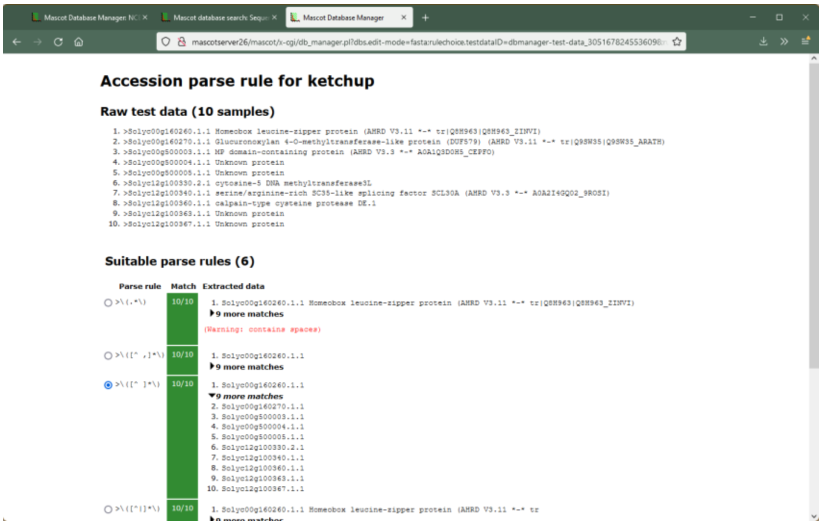
Description parse rule (?)
(none)
Choose

Cancel Next

Accession and description parse rule must be selected before you can continue to the next step.

First thing we have to do is choose parse rules for accession and description.

Create New - Custom


Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager
mascotserver26/mascot/x-cgi/db_manager.pl?dbs.edit-mode=fasta:rulechoice.testdataID=dbmanager-test-data_3051678245536098

Accession parse rule for ketchup

Raw test data (10 samples)

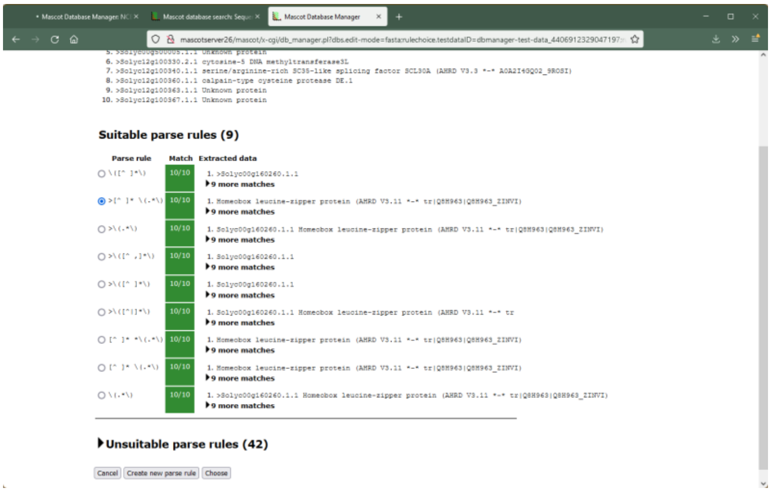
```
 1. >Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)
 2. >Solyc00g160270.1.1 Glucuronoxylan 4-O-methyltransferase-like protein (DUF579) (AHRD V3.11 *-* tr|Q9SW35|Q9SW35_ARATH)
 3. >Solyc00g500003.1.1 MP domain-containing protein (AHRD V3.3 *-* A0A1Q3D0H5_CEPFO)
 4. >Solyc00g500004.1.1 Unknown protein
 5. >Solyc00g500005.1.1 Unknown protein
 6. >Solyc12g100330.2.1 cytosine-5 DNA methyltransferase3L
 7. >Solyc12g100340.1.1 serine/arginine-rich SC35-like splicing factor SCL30A (AHRD V3.3 *-* A0A2I4GQ02_9ROSI)
 8. >Solyc12g100360.1.1 calpain-type cysteine protease DE.1
 9. >Solyc12g100363.1.1 Unknown protein
10. >Solyc12g100367.1.1 Unknown protein
```

Suitable parse rules (6)

Parse rule	Match	Extracted data
○ >\(.*\)	10/10	1. Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI) ▸9 more matches (Warning: contains spaces)
○ >\([^ ,]*\)	10/10	1. Solyc00g160260.1.1 ▸9 more matches
● >\([^]*\)	10/10	1. Solyc00g160260.1.1 ▾*9 more matches* 2. Solyc00g160270.1.1 3. Solyc00g500003.1.1 4. Solyc00g500004.1.1 5. Solyc00g500005.1.1 6. Solyc12g100330.2.1 7. Solyc12g100340.1.1 8. Solyc12g100360.1.1 9. Solyc12g100363.1.1 10. Solyc12g100367.1.1
○ >\([^|]*\)	10/10	1. Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr ▸9 more matches

All the existing parse rules are tested against ten of the title lines, five from the start of the file and five from the end. Four parse rules give matches to all ten. We need to study the matches and choose the one that pulls out a suitable accession string. I think its fairly obvious that the one we've selected is suitable.

Create New - Custom



• Mascot Database Manager: NC ×
Mascot database search: Seque ×
Mascot Database Manager ×
mascotserver26/mascot/x-cgi/db_manager.pl?dbs.edit-mode=fasta:rulechoice.testdataID=dbmanager-test-data_440691232904719

```
 5. >Solyc00g500005.1.1 Unknown protein
 6. >Solyc12g100330.2.1 cytosine-5 DNA methyltransferase3L
 7. >Solyc12g100340.1.1 serine/arginine-rich SC35-like splicing factor SCL30A (AHRD V3.3 *-* A0A2I4GQ02_9ROSI)
 8. >Solyc12g100360.1.1 calpain-type cysteine protease DE.1
 9. >Solyc12g100363.1.1 Unknown protein
10. >Solyc12g100367.1.1 Unknown protein
```

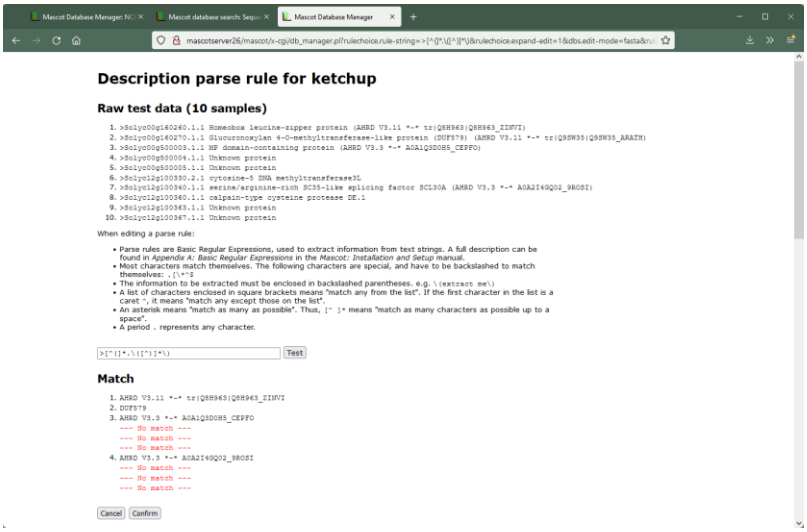
Suitable parse rules (9)

<table>
<tr><th>Parse rule</th><th>Match</th><th>Extracted data</th></tr>
<tr><td>○ \([^]*\)</td><td>10/10</td><td>1. >Solyc00g160260.1.1
▸9 more matches</td></tr>
<tr><td>● >[^]* \(.*\)</td><td>10/10</td><td>1. Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)
▸9 more matches</td></tr>
<tr><td>○ >\(.*\)</td><td>10/10</td><td>1. Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)
▸9 more matches</td></tr>
<tr><td>○ >\([^ ,]*\)</td><td>10/10</td><td>1. Solyc00g160260.1.1
▸9 more matches</td></tr>
<tr><td>○ >\([^]*\)</td><td>10/10</td><td>1. Solyc00g160260.1.1
▸9 more matches</td></tr>
<tr><td>○ >\([^|]*\)</td><td>10/10</td><td>1. Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr
▸9 more matches</td></tr>
<tr><td>○ [^]* *\(.*\)</td><td>10/10</td><td>1. Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)
▸9 more matches</td></tr>
<tr><td>○ [^]* \(.*\)</td><td>10/10</td><td>1. Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)
▸9 more matches</td></tr>
<tr><td>○ \(.*\)</td><td>10/10</td><td>1. >Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)
▸9 more matches</td></tr>
</table>

▸**Unsuitable parse rules (42)**

When it comes to the description parse rule, the first one listed would be fine but, as a challenge, we’ll devise a new rule to extract only the Interpro reference and the text that follows. We click on ‘Create new parse rule’.

Create New - Custom



Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager
mascotserver26/mascot/x-cgi/db_manager.pl?rulechoice.rule-string=>[^(]*.\([^)]*\)&rulechoice.expand-edit=1&dbs.edit-mode=fasta&ru

Description parse rule for ketchup

Raw test data (10 samples)

1. `>Solyc00g160260.1.1 Homeobox leucine-zipper protein (AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI)`
2. `>Solyc00g160270.1.1 Glucuronoxylan 4-O-methyltransferase-like protein (DUF579) (AHRD V3.11 *-* tr|Q9SW35|Q9SW35_ARATH)`
3. `>Solyc00g500003.1.1 MP domain-containing protein (AHRD V3.3 *-* A0A1Q3D0H5_CEPFO)`
4. `>Solyc00g500004.1.1 Unknown protein`
5. `>Solyc00g500005.1.1 Unknown protein`
6. `>Solyc12g100330.2.1 cytosine-5 DNA methyltransferase3L`
7. `>Solyc12g100340.1.1 serine/arginine-rich SC35-like splicing factor SCL30A (AHRD V3.3 *-* A0A2I4GQ02_9ROSI)`
8. `>Solyc12g100360.1.1 calpain-type cysteine protease DE.1`
9. `>Solyc12g100363.1.1 Unknown protein`
10. `>Solyc12g100367.1.1 Unknown protein`

When editing a parse rule:

- Parse rules are Basic Regular Expressions, used to extract information from text strings. A full description can be found in *Appendix A: Basic Regular Expressions* in the *Mascot: Installation and Setup* manual.
- Most characters match themselves. The following characters are special, and have to be backslashed to match themselves: `.[\*^$`
- The information to be extracted must be enclosed in backslashed parentheses. e.g. `\(extract me\)`
- A list of characters enclosed in square brackets means "match any from the list". If the first character in the list is a caret `^`, it means "match any except those on the list".
- An asterisk means "match as many as possible". Thus, `[^ ]*` means "match as many characters as possible up to a space".
- A period `.` represents any character.

`>[^(]*.\([^)]*\)` Test

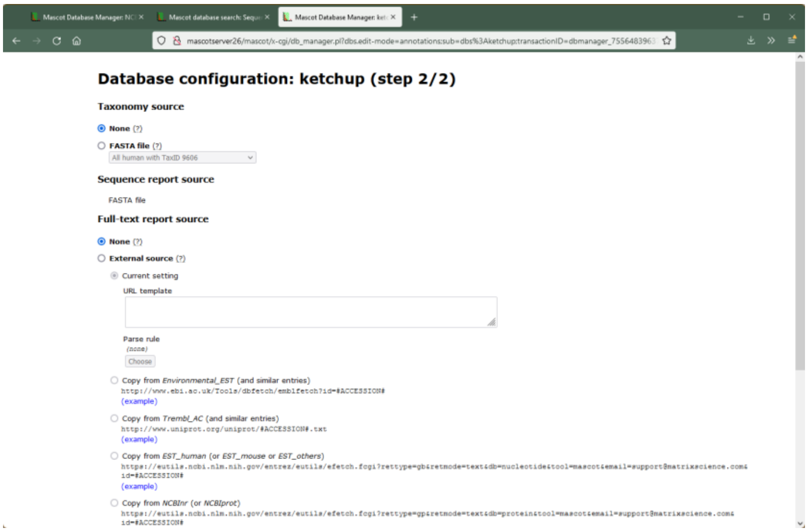
Match

1. `AHRD V3.11 *-* tr|Q8H963|Q8H963_ZINVI`
2. `DUF579`
3. `AHRD V3.3 *-* A0A1Q3D0H5_CEPFO`
   `--- No match ---`
   `--- No match ---`
   `--- No match ---`
4. `AHRD V3.3 *-* A0A2I4GQ02_9ROSI`
   `--- No match ---`
   `--- No match ---`
   `--- No match ---`

Cancel Confirm

There is a brief reminder of how basic regular expressions work, and you can use trial and error; pressing test until your rule succeeds. This screen shot shows a rule that does not work. It locks on to the text after the open parentheses, then takes everything from that point to the closing of the parentheses. This is not a good rule, because not all titles include a reference accession number in parentheses, but let's not worry about that right now. We confirm our choices and move on to the remainder of the configuration.

Create New - Custom



Mascot Database Manager: NC
Mascot database search: Seque
Mascot Database Manager: ketc

mascotserver26/mascot/x-cgi/db_manager.pl?dbs.edit-mode=annotations;sub=dbs%3Aketchup;transactionID=dbmanager_7556483963

Database configuration: ketchup (step 2/2)

Taxonomy source

● **None** (?)

○ **FASTA file** (?)

All human with TaxID 9606

Sequence report source

FASTA file

Full-text report source

● **None** (?)

○ **External source** (?)

● Current setting

URL template

Parse rule

(none)

○ Copy from *Environmental_EST* (and similar entries)

`http://www.ebi.ac.uk/Tools/dbfetch/emblfetch?id=#ACCESSION#`

(example)

○ Copy from *Trembl_AC* (and similar entries)

`http://www.uniprot.org/uniprot/#ACCESSION#.txt`

(example)

○ Copy from *EST_human* (or *EST_mouse* or *EST_others*)

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?rettype=gb&retmode=text&db=nucleotide&tool=mascot&email=support@matrixscience.com&id=#ACCESSION#`

(example)

○ Copy from *NCBInr* (or *NCBIprot*)

`https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?rettype=gp&retmode=text&db=protein&tool=mascot&email=support@matrixscience.com&id=#ACCESSION#`

The next step deals with taxonomy and annotations.

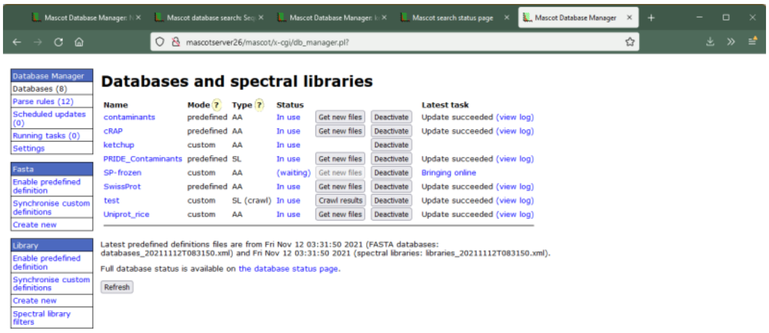
If this was a comprehensive database, containing entries from many different organisms, we might want to select a rule for determining the taxonomy of each entry. There is a drop down list of rules for the major public databases. Most of these require additional files, which can be downloaded automatically by Database Manager whenever the Fasta file is updated. It is extremely unlikely that you will need to create a new taxonomy definition. But, for completeness, the syntax is described in Chapter 9 of the Installation & Setup manual.

Annotation text usually comes from some type of web service. Mascot submits a request using an accession string and the text is returned and embedded in the Protein View report. This form allows you to select from existing URLs or create a new one. In rare cases, the full text for the entire database is downloaded as a local file. The only common examples of this are the SwissProt and Trembl DAT files.

For the tomato genome, as is often the case with simple, single organism databases, we don’t need to worry about taxonomy and there isn’t a suitable source for full-text annotation reports.

At the bottom of this page, we can choose ‘Save and Finish’ then, on the next page, ‘Activate’.

Create New - Custom



Databases and spectral libraries

Name	Mode ?	Type ?	Status			Latest task
contaminants	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
cRAP	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
ketchup	custom	AA	In use		Deactivate	
PRIDE_Contaminants	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
SP-frozen	custom	AA	(waiting)	Get new files	Deactivate	Bringing online
SwissProt	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
test	custom	SL (crawl)	In use	Crawl results	Deactivate	Update succeeded (view log)
Uniprot_rice	custom	AA	In use	Get new files	Deactivate	Update succeeded (view log)

Latest predefined definitions files are from Fri Nov 12 03:31:50 2021 (FASTA databases: databases_20211112T083150.xml) and Fri Nov 12 03:31:50 2021 (spectral libraries: libraries_20211112T083150.xml).

Full database status is available on the database status page.

Refresh

All being well, a short time later, our new database will show as ‘In Use’. You’ll notice that there is no option to update ‘ketchup’ because we copied the file manually.

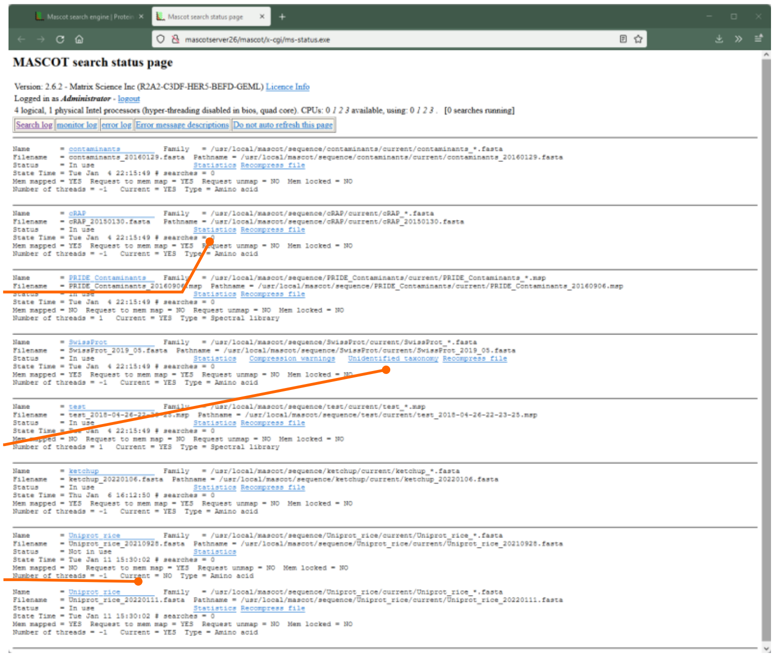
If there are problems, and the database fails to reach ‘In use’, you’ll need to follow the Status link to Database Status .

Database Status

Statistics

Unidentified taxonomy

“Old” & “New”


Mascot search engine | Protein
Mascot search status page
mascotserver26/mascot/x-cgi/ms-status.exe

MASCOT search status page

Version: 2.6.2 - Matrix Science Inc (R2A2-C3DF-HER5-BEFD-GEML) Licence Info
Logged in as ***Administrator*** - logout
4 logical, 1 physical Intel processors (hyper-threading disabled in bios, quad core). CPUs: 0 *1 2 3* available, using: 0 *1 2 3* .   [0 searches running]

Search log | monitor log | error log | Error message descriptions | Do not auto refresh this page

```
Name       = contaminants          Family   = /usr/local/mascot/sequence/contaminants/current/contaminants_*.fasta
Filename   = contaminants_20160129.fasta  Pathname = /usr/local/mascot/sequence/contaminants/current/contaminants_20160129.fasta
Status     = In use                     Statistics Recompress file
State Time = Tue Jan  4 22:15:49 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = YES  Type = Amino acid

Name       = cRAP                  Family   = /usr/local/mascot/sequence/cRAP/current/cRAP_*.fasta
Filename   = cRAP_20150130.fasta   Pathname = /usr/local/mascot/sequence/cRAP/current/cRAP_20150130.fasta
Status     = In use                     Statistics Recompress file
State Time = Tue Jan  4 22:15:49 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = YES  Type = Amino acid

Name       = PRIDE_Contaminants    Family   = /usr/local/mascot/sequence/PRIDE_Contaminants/current/PRIDE_Contaminants_*.msp
Filename   = PRIDE_Contaminants_20160906.msp  Pathname = /usr/local/mascot/sequence/PRIDE_Contaminants/current/PRIDE_Contaminants_20160906.msp
Status     = In use                     Statistics Recompress file
State Time = Tue Jan  4 22:15:49 # searches = 0
Mem mapped = NO  Request to mem map = NO  Request unmap = NO  Mem locked = NO
Number of threads = 1   Current = YES  Type = Spectral library

Name       = SwissProt             Family   = /usr/local/mascot/sequence/SwissProt/current/SwissProt_*.fasta
Filename   = SwissProt_2019_05.fasta  Pathname = /usr/local/mascot/sequence/SwissProt/current/SwissProt_2019_05.fasta
Status     = In use                     Statistics   Compression warnings   Unidentified taxonomy Recompress file
State Time = Tue Jan  4 22:15:49 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = YES  Type = Amino acid

Name       = test                  Family   = /usr/local/mascot/sequence/test/current/test_*.msp
Filename   = test_2018-04-26-22-23-25.msp  Pathname = /usr/local/mascot/sequence/test/current/test_2018-04-26-22-23-25.msp
Status     = In use                     Statistics Recompress file
State Time = Tue Jan  4 22:15:49 # searches = 0
Mem mapped = NO  Request to mem map = NO  Request unmap = NO  Mem locked = NO
Number of threads = 1   Current = YES  Type = Spectral library

Name       = ketchup               Family   = /usr/local/mascot/sequence/ketchup/current/ketchup_*.fasta
Filename   = ketchup_20220106.fasta  Pathname = /usr/local/mascot/sequence/ketchup/current/ketchup_20220106.fasta
Status     = In use                     Statistics Recompress file
State Time = Thu Jan  6 16:12:50 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = YES  Type = Amino acid

Name       = Uniprot_rice          Family   = /usr/local/mascot/sequence/Uniprot_rice/current/Uniprot_rice_*.fasta
Filename   = Uniprot_rice_20210928.fasta  Pathname = /usr/local/mascot/sequence/Uniprot_rice/current/Uniprot_rice_20210928.fasta
Status     = Not in use                 Statistics
State Time = Tue Jan 11 15:30:02 # searches = 0
Mem mapped = NO  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = NO  Type = Amino acid

Name       = Uniprot_rice          Family   = /usr/local/mascot/sequence/Uniprot_rice/current/Uniprot_rice_*.fasta
Filename   = Uniprot_rice_20220111.fasta  Pathname = /usr/local/mascot/sequence/Uniprot_rice/current/Uniprot_rice_20220111.fasta
Status     = In use                     Statistics Recompress file
State Time = Tue Jan 11 15:30:02 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = YES  Type = Amino acid
```

Database status provides an overview of all the active databases. It also provides links to other pages of useful information.

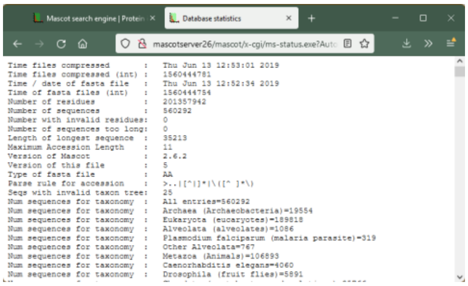
Initially, there will be a single information block for each database on this page. When a database is updated, a second information block is added. One is for the new or incoming database, the other is for the old or outgoing. If all is well, one of the pair will have the status of “In use”, and the other “Not in use”.

If there is a problem, the status will be an error message and a ‘compression warning’ link added to the relevant error messages.

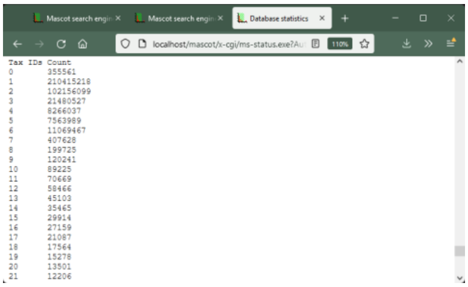
The database statistics are very useful for diagnosing problems and checking up on the health of a database.

Database Statistics

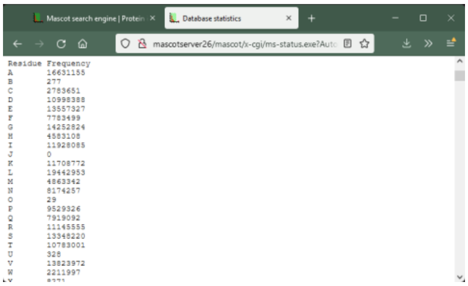
- Is the number of entries correct?
- Any invalid codes?
- Any entries “too long”?
- Is an AA database all ACGT?
- If using taxonomy, is the success rate > 99%?


Mascot search engine | Protein
Database statistics
mascotserver26/mascot/x-cgi/ms-status.exe?Auto

```
Time files compressed       :   Thu Jun 13 12:53:01 2019
Time files compressed (int) :   1560444781
Time / date of fasta file   :   Thu Jun 13 12:52:34 2019
Time of fasta files (int)   :   1560444754
Number of residues          :   201357942
Number of sequences         :   560292
Number with invalid residues:   0
Number of sequences too long:   0
Length of longest sequence  :   35213
Maximum Accession Length    :   11
Version of Mascot           :   2.6.2
Version of this file        :   5
Type of fasta file          :   AA
Parse rule for accession    :   >..|[^|]*|\([^ ]*\)
Seqs with invalid taxon tree:   25
Num sequences for taxonomy  :   All entries=560292
Num sequences for taxonomy  :   Archaea (Archaeobacteria)=19554
Num sequences for taxonomy  :   Eukaryota (eucaryotes)=189818
Num sequences for taxonomy  :   Alveolata (alveolates)=1086
Num sequences for taxonomy  :   Plasmodium falciparum (malaria parasite)=319
Num sequences for taxonomy  :   Other Alveolata=767
Num sequences for taxonomy  :   Metazoa (Animals)=106893
Num sequences for taxonomy  :   Caenorhabditis elegans=4060
Num sequences for taxonomy  :   Drosophila (fruit flies)=5891
```



```
Tax IDs Count
0       355561
1       210415218
2       102156099
3       21480527
4       8266037
5       7563989
6       11069467
7       407628
8       199725
9       120241
10      89225
11      70669
12      58466
13      45103
14      35465
15      29914
16      27159
17      21087
18      17564
19      15278
20      13501
21      12206
```


Mascot search engine | Protein
Database statistics
mascotserver26/mascot/x-cgi/ms-status.exe?Auto

```
Residue Frequency
A       16631155
B       277
C       2783651
D       10998388
E       13557327
F       7783499
G       14252824
H       4583108
I       11928085
J       0
K       11708772
L       19442953
M       4863342
N       8174257
O       29
P       9529326
Q       7919092
R       11145555
S       13348220
T       10783001
U       328
V       13823972
W       2211997
```

For example, does the number of entries look about right? Sometimes, a download may be truncated and the problem go undetected.

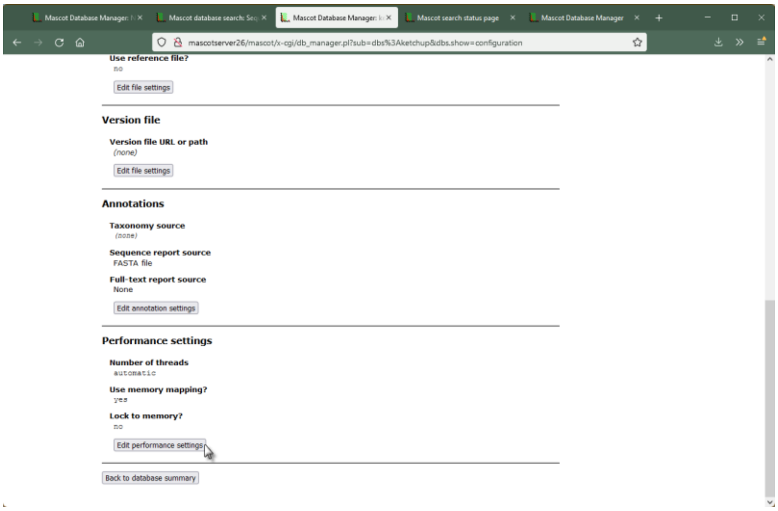
Are there any invalid characters in the sequences? If there are, this should definitely be investigated.

Mascot has a parameter, MaxSequenceLen, to set the length of the longest sequence. The default is 80,000. The higher this value, the more memory Mascot uses, so it should not be set to a ridiculously high value. If any sequences are “too long”, then you need to increase MaxSequenceLen to something a little greater than the length of the longest sequence. If you are trying to search an assembled genome, you might want to consider searching shorter sequences instead, such as a database of contigs.

If your protein database seems to be composed entirely of A, C, G, and T, then it may be worth double checking that you downloaded the correct file.

Although it is rarely possible to achieve 100% accuracy for taxonomy, you certainly want the accuracy to be better than 99%. Otherwise, the results could be misleading. Near the bottom of the stats file is a list of the number of entries with 0, 1, 2, etc., taxonomy identifiers. From time to time, check that the number of entries with no taxonomy identifiers is well below 1% of the database. Here, it is a very healthy 0.03%.

Configuration - Performance


Mascot Database Manager: N
Mascot database search: Seq
Mascot Database Manager: k
Mascot search status page
Mascot Database Manager
mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs%3Aketchup&dbs.show=configuration

Use reference file?
no

Edit file settings

Version file

Version file URL or path
(none)

Edit file settings

Annotations

Taxonomy source
(none)

Sequence report source
FASTA file

Full-text report source
None

Edit annotation settings

Performance settings

Number of threads
automatic

Use memory mapping?
yes

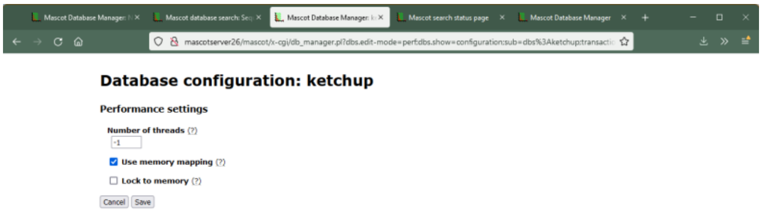
Lock to memory?
no

Edit performance settings

Back to database summary

If you look at the configuration details for any database, there is a section for Performance Settings.

Configuration - Performance


Mascot Database Manager: N
Mascot database search: Seq
Mascot Database Manager: k
Mascot search status page
Mascot Database Manager
mascotserver26/mascot/x-cgi/db_manager.pl?dbs.edit-mode=perf;dbs.show=configuration;sub=dbs%3Aketchup;transactio

Database configuration: ketchup

Performance settings

Number of threads (?)
-1

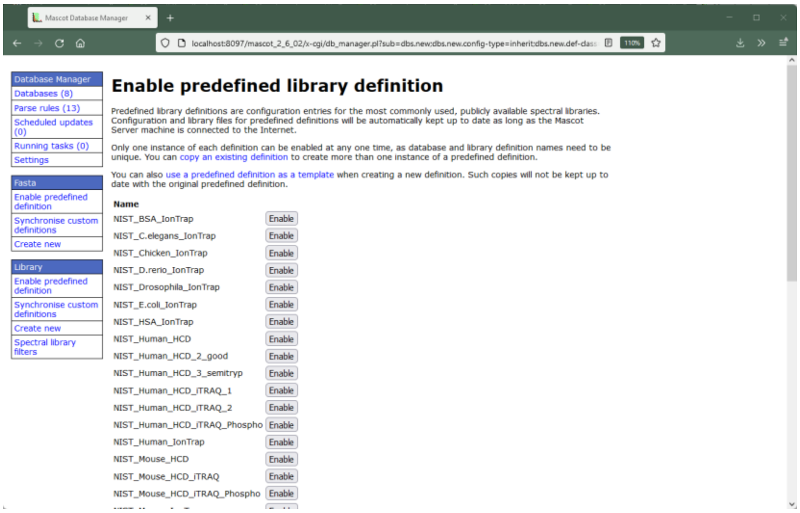
☑ **Use memory mapping** (?)

☐ **Lock to memory** (?)

A Mascot search can use multiple threads, so as to make use of all the logical processors covered by the licence. Usually, it is best to leave threads set to -1, which means automatic. If you want to restrict the number of threads on a non-cluster (SMP) system, you can do so by setting a value of 1 or more. Each CPU in the Mascot licence allows use of up to 4 cores, which requires 8 threads for a hyperthreaded processor or 4 otherwise. On a cluster system, the number of threads is set for each search node in a separate configuration file, nodelist.txt.

Database files should always be memory mapped because this gives the fastest access times. Memory mapped files can be locked in memory, but only if the computer has sufficient RAM. Having a database locked in memory means that it can never be swapped out to disk, ensuring there is never a lag as the database is read from disk. If you try to lock databases into RAM when there isn't room, this will not be a major problem. The locking will fail, generate an error message, and Mascot will carry on regardless. A more serious problem is when there is just sufficient RAM to lock the databases, but none left over for searches or other applications. In this case, the whole system will slow down and the hard disk will be observed to be "thrashing". Eventually, the system is likely to hang or crash. In general, it is better to let the operating system manage which files are held in memory and not lock any databases into memory.

Enable Predefined Definition - Spectral Library


Mascot Database Manager
localhost:8097/mascot_2_6_02/x-cgi/db_manager.pl?sub=dbs.new;dbs.new.config-type=inherit;dbs.new.def-class
110%
Database Manager
Databases (8)
Parse rules (13)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Enable predefined library definition

Predefined library definitions are configuration entries for the most commonly used, publicly available spectral libraries. Configuration and library files for predefined definitions will be automatically kept up to date as long as the Mascot Server machine is connected to the Internet.

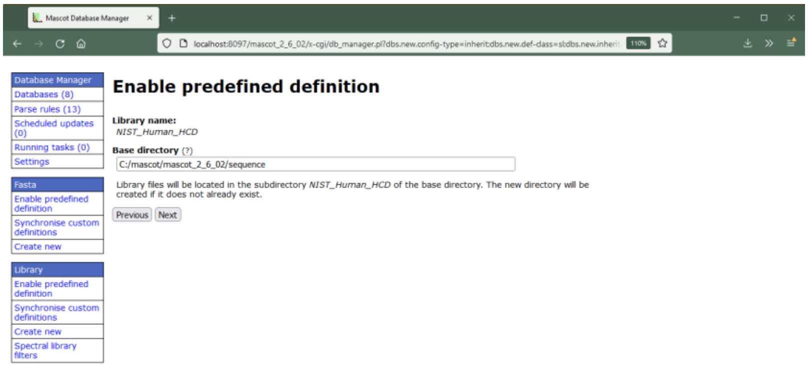
Only one instance of each definition can be enabled at any one time, as database and library definition names need to be unique. You can copy an existing definition to create more than one instance of a predefined definition.

You can also use a predefined definition as a template when creating a new definition. Such copies will not be kept up to date with the original predefined definition.

Name	
NIST_BSA_IonTrap	Enable
NIST_C.elegans_IonTrap	Enable
NIST_Chicken_IonTrap	Enable
NIST_D.rerio_IonTrap	Enable
NIST_Drosophila_IonTrap	Enable
NIST_E.coli_IonTrap	Enable
NIST_HSA_IonTrap	Enable
NIST_Human_HCD	Enable
NIST_Human_HCD_2_good	Enable
NIST_Human_HCD_3_semitryp	Enable
NIST_Human_HCD_iTRAQ_1	Enable
NIST_Human_HCD_iTRAQ_2	Enable
NIST_Human_HCD_iTRAQ_Phospho	Enable
NIST_Human_IonTrap	Enable
NIST_Mouse_HCD	Enable
NIST_Mouse_HCD_iTRAQ	Enable
NIST_Mouse_HCD_iTRAQ_Phospho	Enable

For spectral libraries activating Predefined Definition is almost the same as activating a fasta definition. Click on the enable button.

Enable Predefined Definition - Spectral Library


Mascot Database Manager
localhost:8097/mascot_2_6_02/x-cgi/db_manager.pl?dbs.new.config-type=inherit;dbs.new.def-class=s;dbs.new.inherit
110%
Database Manager
Databases (8)
Parse rules (13)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Enable predefined definition

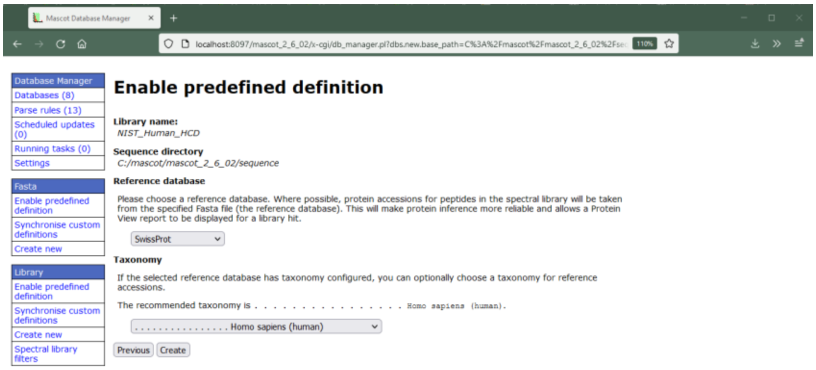
Library name:
NIST_Human_HCD

Base directory (?)
C:/mascot/mascot_2_6_02/sequence

Library files will be located in the subdirectory *NIST_Human_HCD* of the base directory. The new directory will be created if it does not already exist.

Save the directory location as per normal.

Enable Predefined Definition - Spectral Library


Mascot Database Manager
localhost:8097/mascot_2_6_02/x-cgi/db_manager.pl?dbs.new.base_path=C%3A%2Fmascot%2Fmascot_2_6_02%2Fse
110%
Database Manager
Databases (8)
Parse rules (13)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Previous
Create

Enable predefined definition

Library name:
NIST_Human_HCD

Sequence directory
C:/mascot/mascot_2_6_02/sequence

Reference database

Please choose a reference database. Where possible, protein accessions for peptides in the spectral library will be taken from the specified Fasta file (the reference database). This will make protein inference more reliable and allows a Protein View report to be displayed for a library hit.

SwissProt

Taxonomy

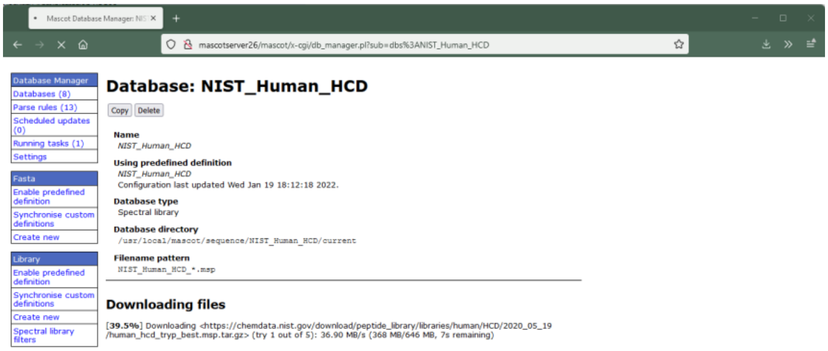
If the selected reference database has taxonomy configured, you can optionally choose a taxonomy for reference accessions.

The recommended taxonomy is . . . . Homo sapiens (human).

. . . . Homo sapiens (human)

And the next step is different, as the spectral library requires a reference library. For the predefined definitions they use SwissProt, a database that is installed on the Mascot Server by default, and the appropriate taxonomy. You can accept these predefined settings and enable the definition.

Enable Predefined Definition - Spectral Library



Mascot Database Manager: NIS
mascotserver26/mascot/x-cgi/db_manager.pl?sub=dbs%3ANIST_Human_HCD
Database Manager
Databases (8)
Parse rules (13)
Scheduled updates (0)
Running tasks (1)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete

Database: NIST_Human_HCD

Name
NIST_Human_HCD

Using predefined definition
NIST_Human_HCD
Configuration last updated Wed Jan 19 18:12:18 2022.

Database type
Spectral library

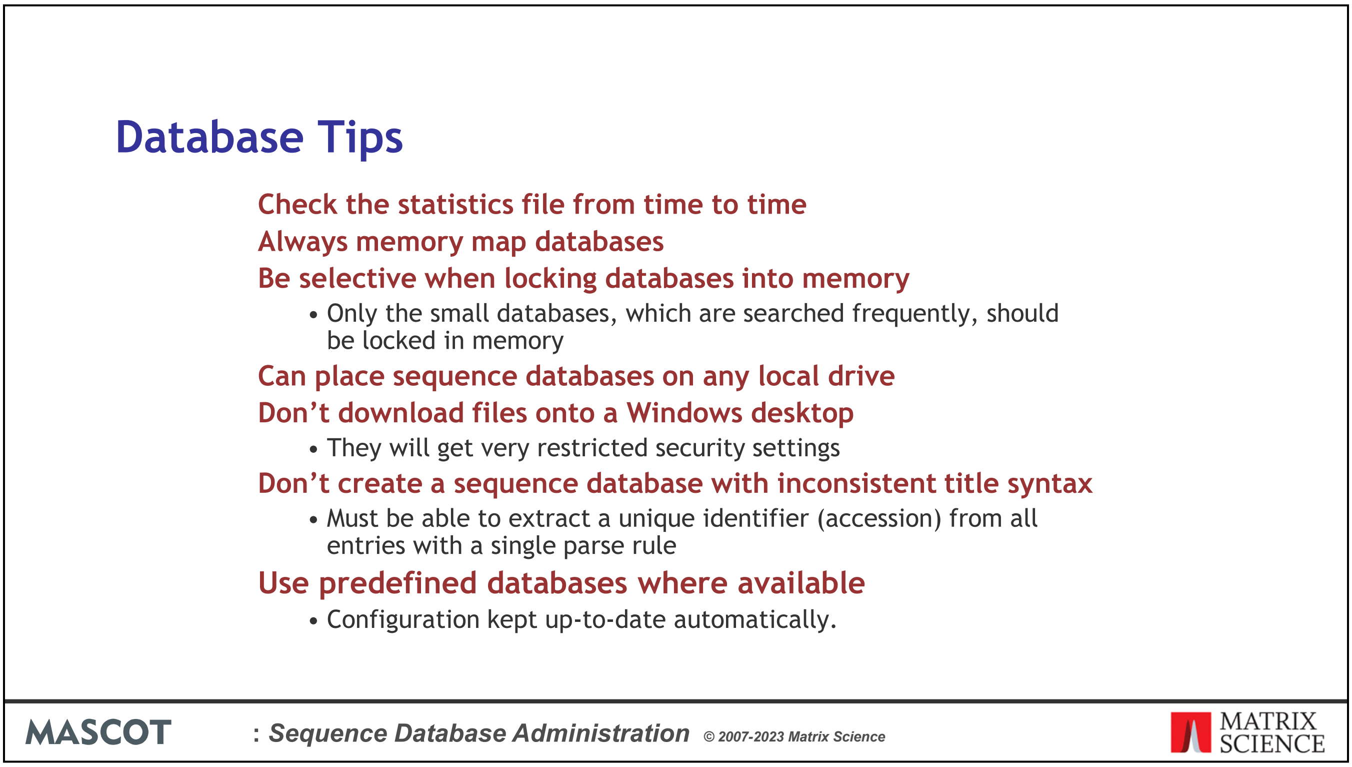
Database directory
`/usr/local/mascot/sequence/NIST_Human_HCD/current`

Filename pattern
`NIST_Human_HCD_*.msp`

Downloading files

[**39.5%**] Downloading <https://chemdata.nist.gov/download/peptide_library/libraries/human/HCD/2020_05_19/human_hcd_tryp_best.msp.tar.gz> (try 1 out of 5): 36.90 MB/s (368 MB/646 MB, 7s remaining)

And Mascot Server will download and uncompress the spectral library then activate it.
Adding a new Spectral Library definition and creating a Spectral Library from search results is covered in the Spectral Library talk.



Database Tips

Check the statistics file from time to time

Always memory map databases

Be selective when locking databases into memory

- Only the small databases, which are searched frequently, should be locked in memory

Can place sequence databases on any local drive

Don’t download files onto a Windows desktop

- They will get very restricted security settings

Don’t create a sequence database with inconsistent title syntax

- Must be able to extract a unique identifier (accession) from all entries with a single parse rule

Use predefined databases where available

- Configuration kept up-to-date automatically.

This slide recaps some important tips.

Check the statistics file from time to time, particularly after configuring a new database.

Always memory map database files to make access as fast as possible, but be selective about locking databases into memory. Only the smaller databases, which are searched regularly, should be locked in memory.

You can place sequence database files on any local drive. Under Unix, you can use NFS mounted drives as long as the connection is fast and stable.

If you download files manually, don’t download to your Windows desktop. Chances are that Database Manager won’t be able to see the file because it becomes private to your Windows login. If this happens, add the local Users group to the security settings for each file and give the Users group full control.

If you create your own Fasta file, use a consistent title syntax. It must be possible to extract a unique identifier (accession) from all entries with a single parse rule

Use predefined databases where available because this means the configuration is kept up-to-date automatically.