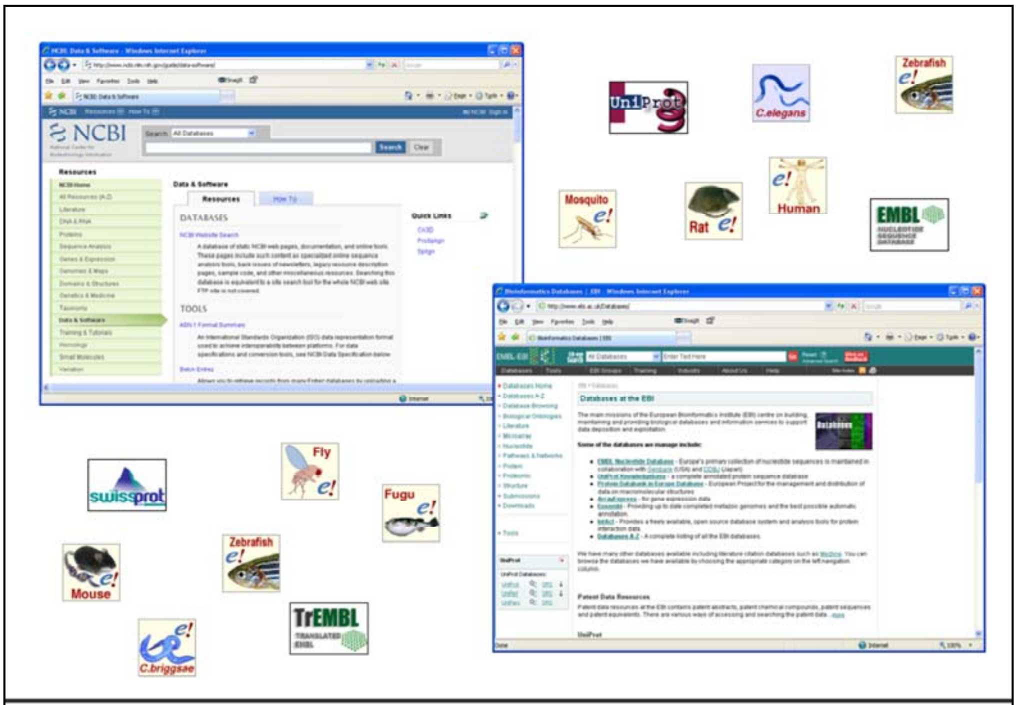
Sequence Database Administration



MASCOT : *Sequence Database Administration* ©2007-2014 Matrix Science MATRIX SCIENCE

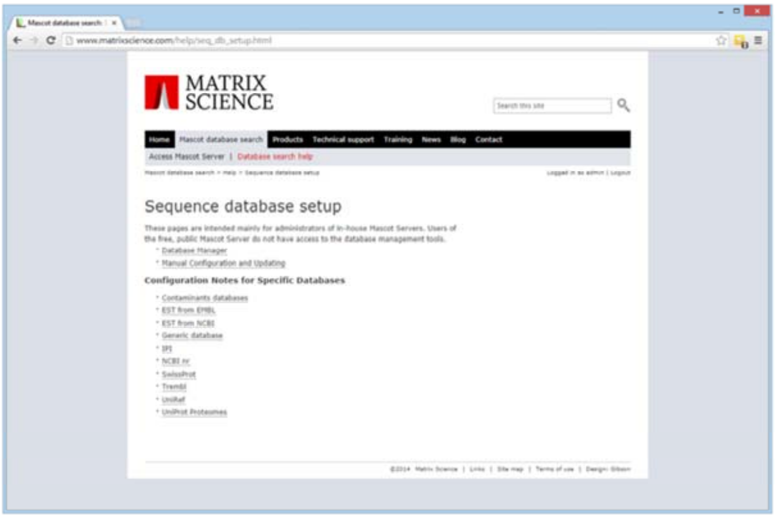
When you install Mascot, it includes a copy of the Swiss-Prot protein database. However, it is almost certain that you and your colleagues will want to search other databases as well. There are very many to choose from, and Mascot allows you to have as many databases as you wish on-line for searching 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, Swiss-Prot, EMBL, 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 search
www.matrixscience.com/help/seq_db_setup.html
MATRIX SCIENCE
Search this site
Home
Mascot database search
Products
Technical support
Training
News
Blog
Contact
Access Mascot Server | Database search help
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.

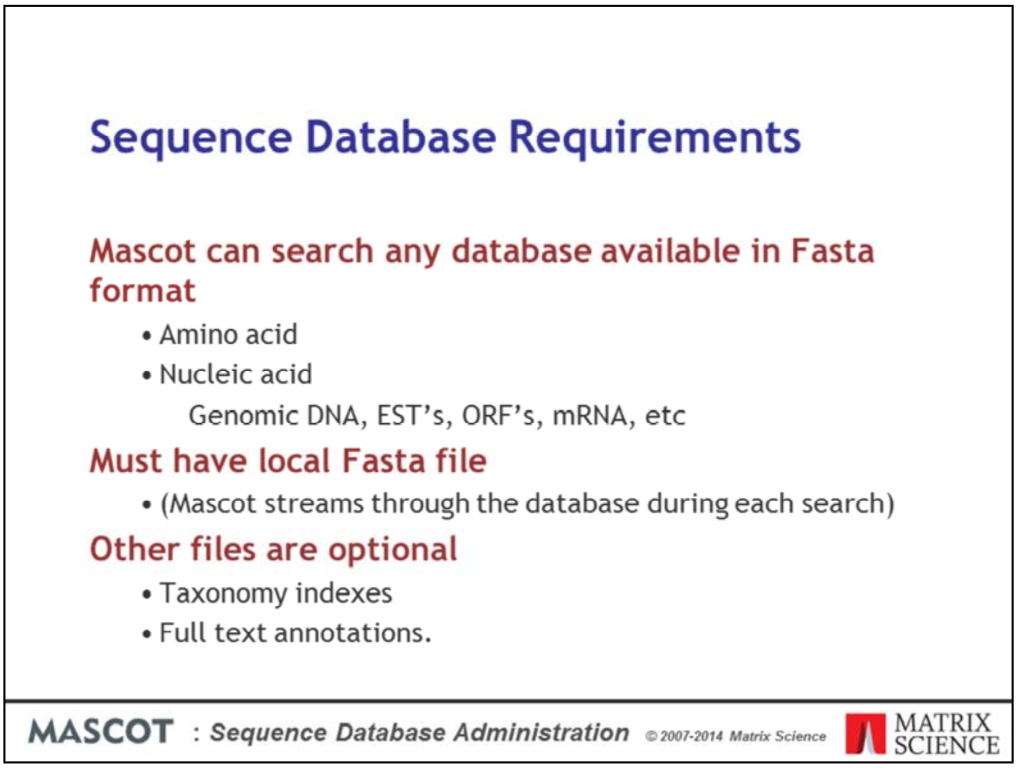
- Database Manager
- Manual Configuration and Updating

Configuration Notes for Specific Databases

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

©2014 Matrix Science | Links | Site map | Terms of use | Design: Gibson

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

Must have local Fasta file

- (Mascot streams through the database during each search)

Other files are optional

- Taxonomy indexes
- Full text annotations.

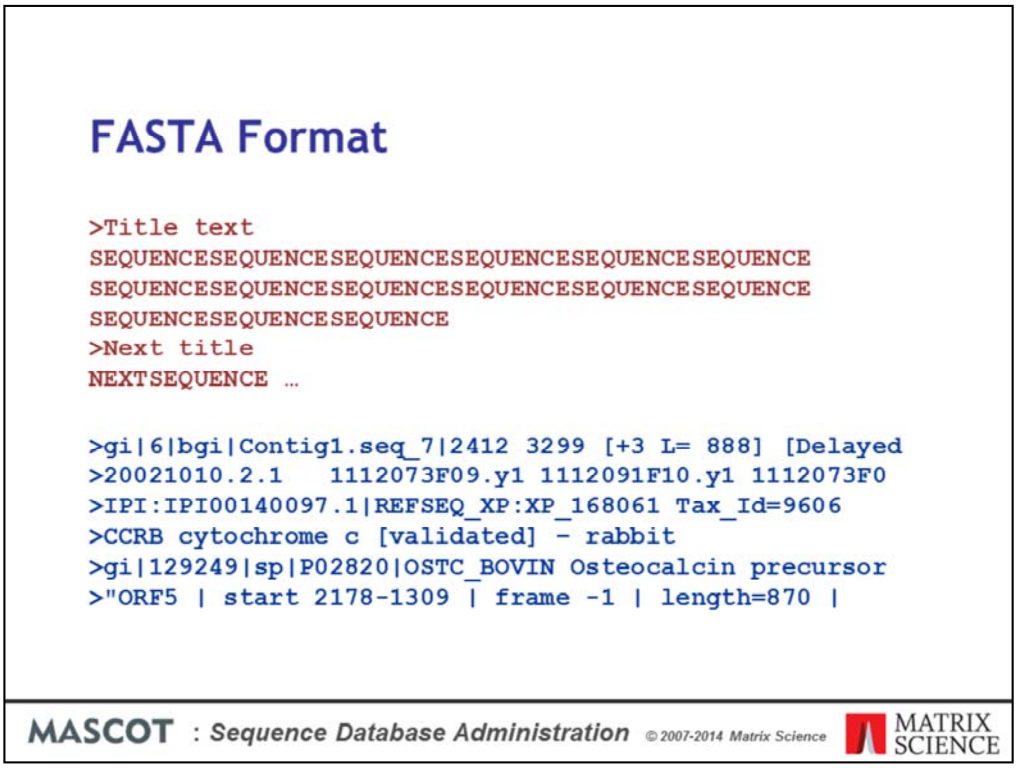
MASCOT : *Sequence Database Administration* © 2007-2014 Matrix Science MATRIX SCIENCE

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. The relative merits of searching protein, EST and DNA sequences are discussed in Choudhary *et. al. Matching peptide mass spectra to EST and genomic DNA databases*. Trends in Biotechnology, 19, S17-S22 (2001)

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.



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 |
```

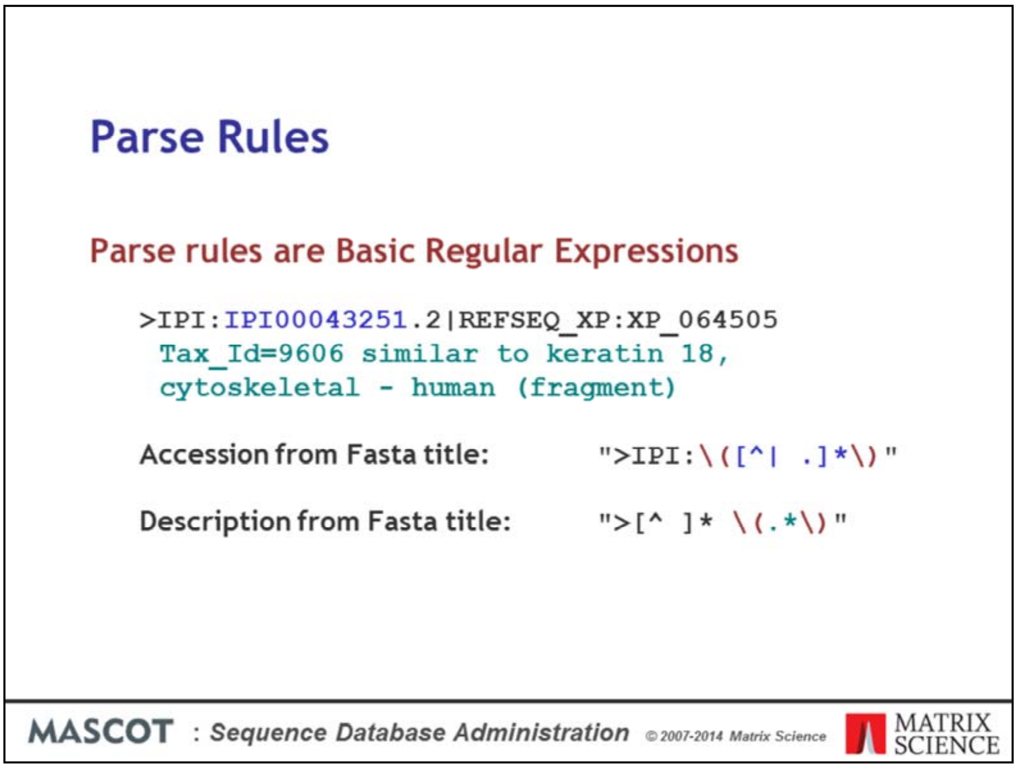
MASCOT : *Sequence Database Administration* © 2007-2014 Matrix Science MATRIX SCIENCE

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

```
>IPI:IPI00043251.2|REFSEQ_XP:XP_064505
 Tax_Id=9606 similar to keratin 18,
 cytoskeletal - human (fragment)
```

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

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

MASCOT : *Sequence Database Administration* © 2007-2014 Matrix Science MATRIX SCIENCE

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.

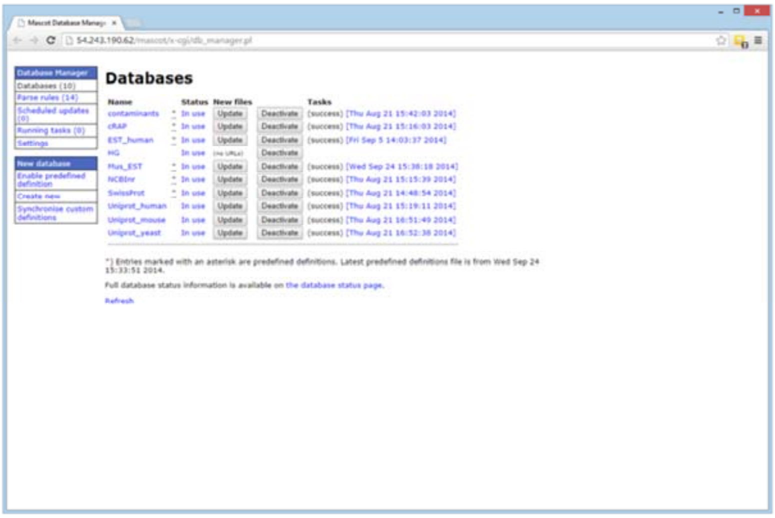
Here, for example, we have a title line from the IPI human database. Let's say that we want to use IPI00043251 as the unique accession string and everything after the first space should be treated as the description.

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

Name		Status	New files		Tasks
contaminants	*	In use	Update	Deactivate	(success) [Thu Aug 21 15:42:03 2014]
cRAP	*	In use	Update	Deactivate	(success) [Thu Aug 21 15:16:03 2014]
EST_human	*	In use	Update	Deactivate	(success) [Fri Sep 5 14:03:37 2014]
HG		In use	(no URLs)	Deactivate	
Mus_EST	*	In use	Update	Deactivate	(success) [Wed Sep 24 15:38:18 2014]
NCBInr	*	In use	Update	Deactivate	(success) [Thu Aug 21 15:15:39 2014]
SwissProt	*	In use	Update	Deactivate	(success) [Thu Aug 21 14:48:54 2014]
Uniprot_human		In use	Update	Deactivate	(success) [Thu Aug 21 15:19:11 2014]
Uniprot_mouse		In use	Update	Deactivate	(success) [Thu Aug 21 16:51:49 2014]
Uniprot_yeast		In use	Update	Deactivate	(success) [Thu Aug 21 16:52:38 2014]

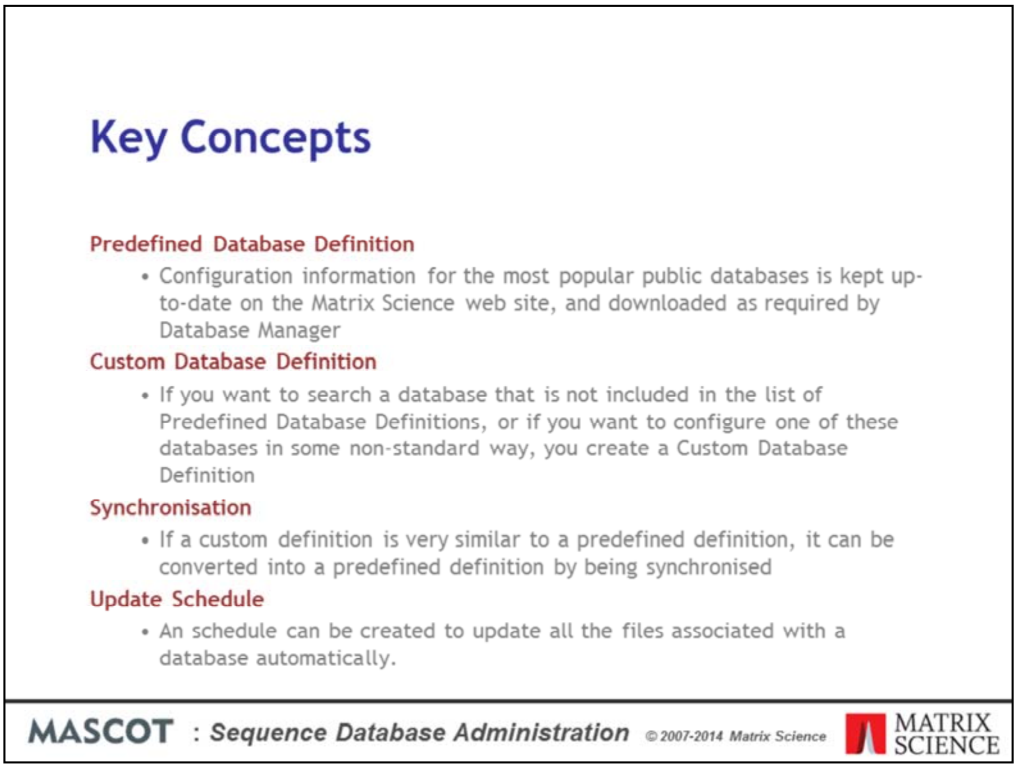
*) Entries marked with an asterisk are predefined definitions. Latest predefined definitions file is from Wed Sep 24 15:33:51 2014.

Full database status information is available on the database status page.

Refresh

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

- An 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:

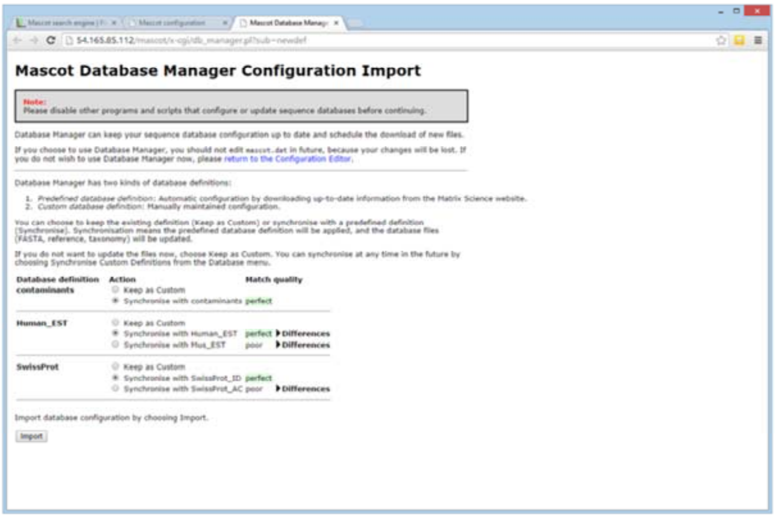
A Predefined Database Definition is one in which the configuration information is kept up-to-date on the Matrix Science web site, and downloaded as required by Database Manager. You don't need to know file URLs or worry about parse rules, etc. for a Predefined Database.

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 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	
contaminants	○ Keep as Custom		
	● Synchronise with contaminants	perfect	
Human_EST	○ Keep as Custom		
	● Synchronise with Human_EST	perfect	▸ **Differences**
	○ Synchronise with Mus_EST	poor	▸ **Differences**
SwissProt	○ Keep as Custom		
	● Synchronise with SwissProt_ID	perfect	
	○ Synchronise with SwissProt_AC	poor	▸ **Differences**

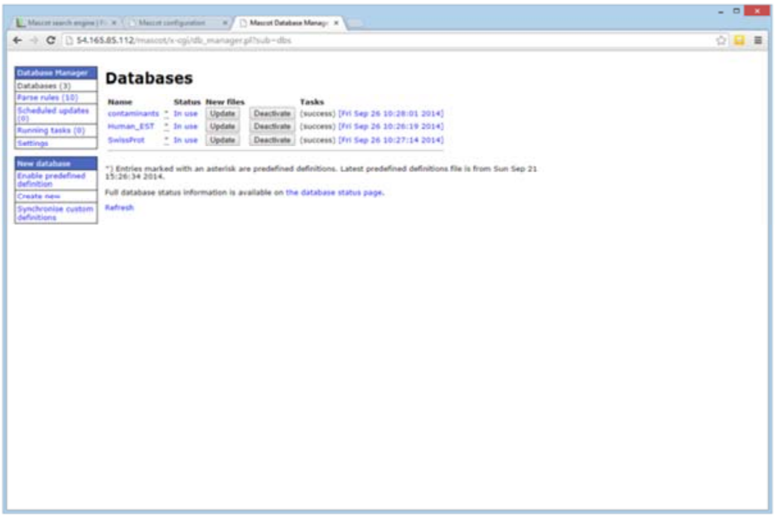
Import database configuration by choosing Import.

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

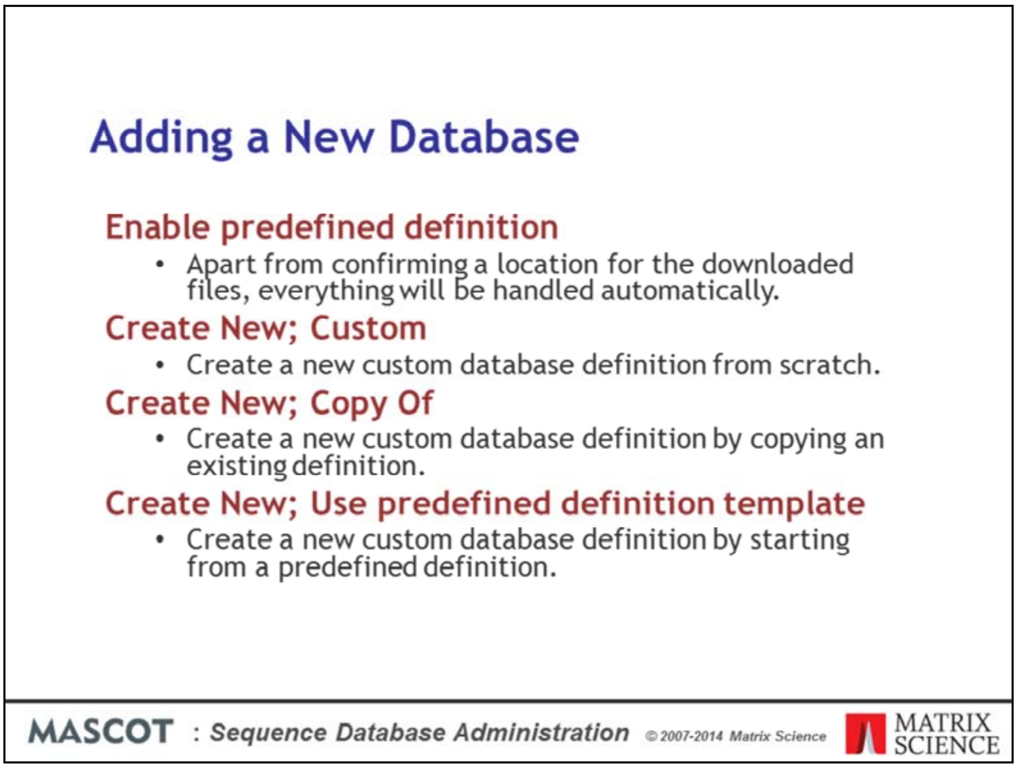
Name		Status	New files		Tasks
contaminants	*	In use	Update	Deactivate	(success) [Fri Sep 26 10:28:01 2014]
Human_EST	*	In use	Update	Deactivate	(success) [Fri Sep 26 10:26:19 2014]
SwissProt	*	In use	Update	Deactivate	(success) [Fri Sep 26 10:27:14 2014]

*) Entries marked with an asterisk are predefined definitions. Latest predefined definitions file is from Sun Sep 21 15:26:34 2014.

Full database status information 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:

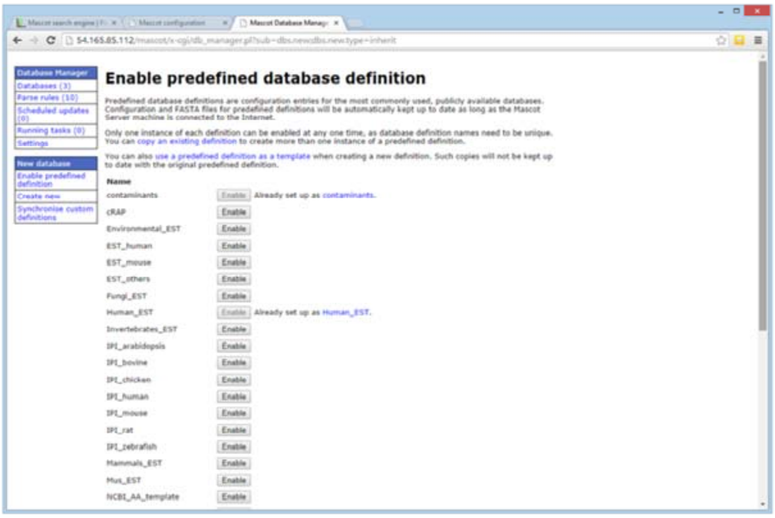
1. Enable predefined definition: Apart from confirming a location for the downloaded files, everything will be handled automatically. Only one instance of each predefined definition can be enabled at any one time, as database names must be unique. If you want something similar to a predefined database,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



Enable predefined database definition

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

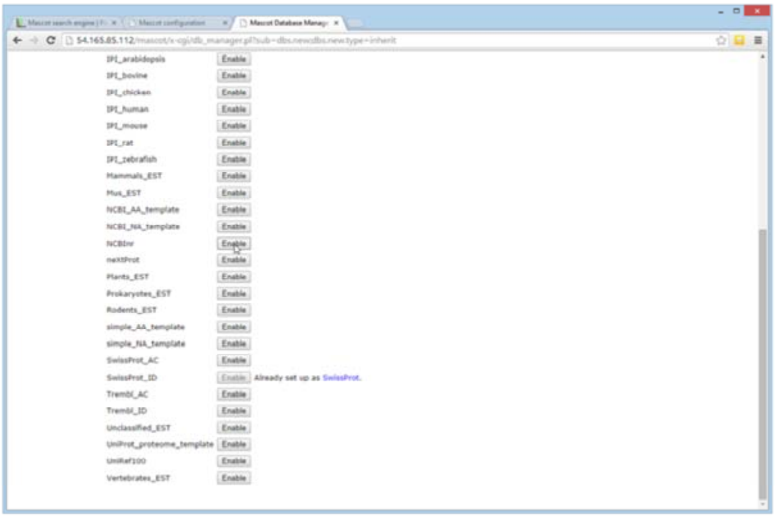
Only one instance of each definition can be enabled at any one time, as database definition names need to be unique. You can copy an existing definition to create more than one instance of a predefined definition.

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	
EST_human	Enable	
EST_mouse	Enable	
EST_others	Enable	
Fungi_EST	Enable	
Human_EST	Enable	Already set up as Human_EST.
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	
NCBI_AA_template	Enable	

Let’s look at the first of these options: enabling a predefined definition, using NCBInr as the example.

Enable Predefined Definition

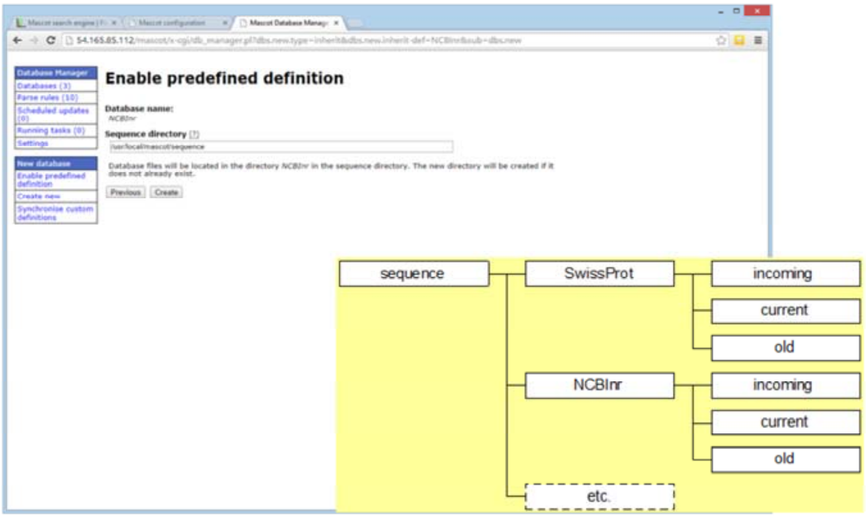


54.165.85.112/mascot/x-cgi/db_manager.pl?sub=dbs.newdbs.new.type=inherit

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
NCBI_AA_template	Enable
NCBI_NA_template	Enable
NCBInr	Enable
neXtProt	Enable
Plants_EST	Enable
Prokaryotes_EST	Enable
Rodents_EST	Enable
simple_AA_template	Enable
simple_NA_template	Enable
SwissProt_AC	Enable
SwissProt_ID	Enable Already set up as SwissProt.
Trembl_AC	Enable
Trembl_ID	Enable
Unclassified_EST	Enable
UniProt_proteome_template	Enable
UniRef100	Enable
Vertebrates_EST	Enable

Scroll down to NCBInr and choose Enable

Enable Predefined Definition



Enable predefined definition

Database name:
NCBInr

Sequence directory [?]
/usr/local/mascot/sequence

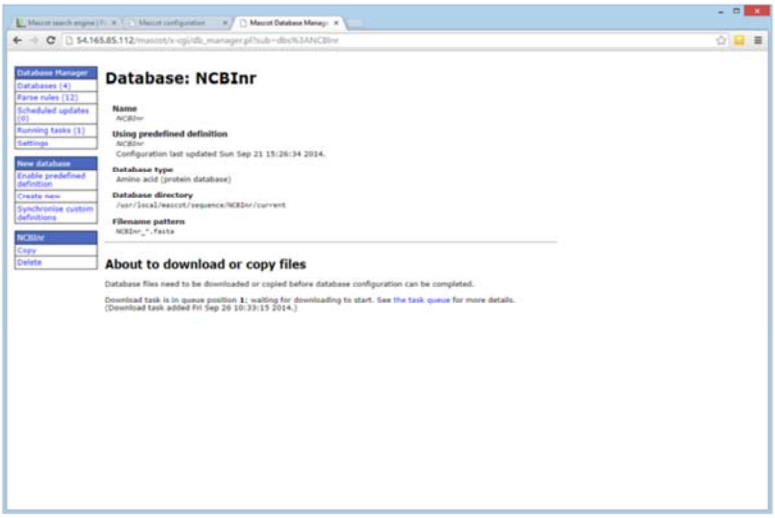
Database files will be located in the directory *NCBInr* in the sequence directory. The new directory will be created if it does not already exist.

MASCOT : *Sequence Database Administration* *© 2007-2014 Matrix Science* MATRIX SCIENCE

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



Database: NCBInr

Name
NCBInr

Using predefined definition
NCBInr
Configuration last updated Sun Sep 21 15:26:34 2014.

Database type
Amino acid (protein database)

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

Filename pattern
`NCBInr_*.fasta`

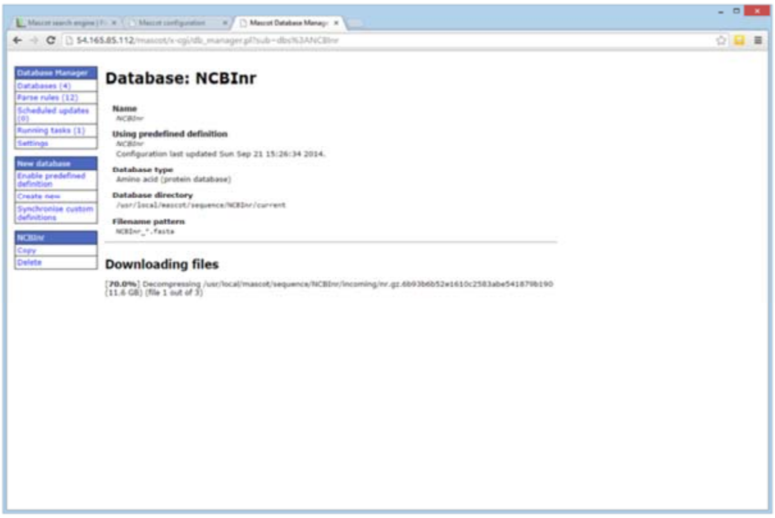
About to download or copy files

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 Fri Sep 26 10:33:15 2014.)

All the files will now be downloaded automatically. For NCBInr, 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



Database: NCBInr

Name
NCBInr

Using predefined definition
NCBInr
Configuration last updated Sun Sep 21 15:26:34 2014.

Database type
Amino acid (protein database)

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

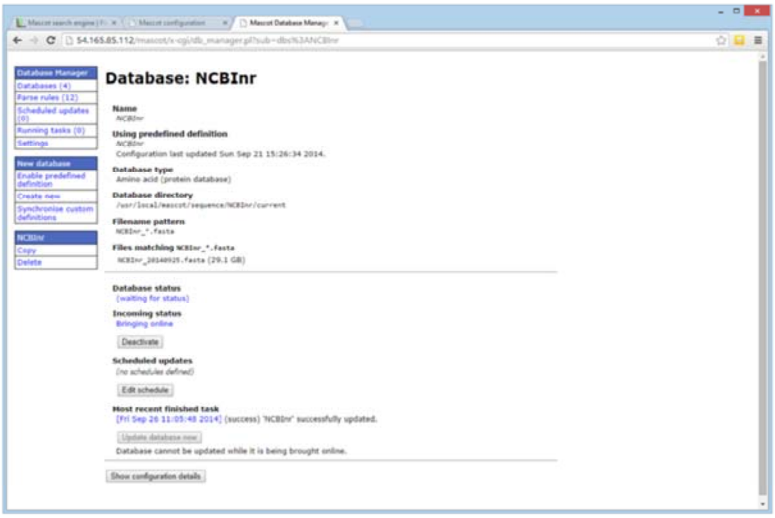
Filename pattern
`NCBInr_*.fasta`

Downloading files

[**70.0%**] Decompressing /usr/local/mascot/sequence/NCBInr/incoming/nr.gz.6b93b6b52e1610c2583abe541879b190 (11.6 GB) (file 1 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



Database: NCBInr

Name
NCBInr

Using predefined definition
NCBInr
Configuration last updated Sun Sep 21 15:26:34 2014.

Database type
Amino acid (protein database)

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

Filename pattern
`NCBInr_*.fasta`

Files matching `NCBInr_*.fasta`
`NCBInr_20140925.fasta` (29.1 GB)

Database status
(waiting for status)

Incoming status
Bringing online

Scheduled updates
(no schedules defined)

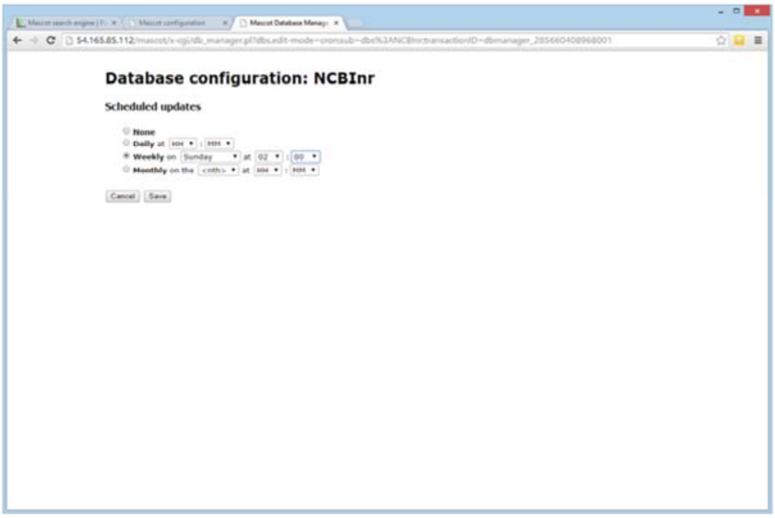
Most recent finished task
[Fri Sep 26 11:05:48 2014] (success) 'NCBInr' successfully updated.

Database cannot be updated while it is being brought online.

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



54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.edit-mode=cronsub=dbs%3ANCBInr;transactionID=dbmanager_285660408968001

Database configuration: NCBInr

Scheduled updates

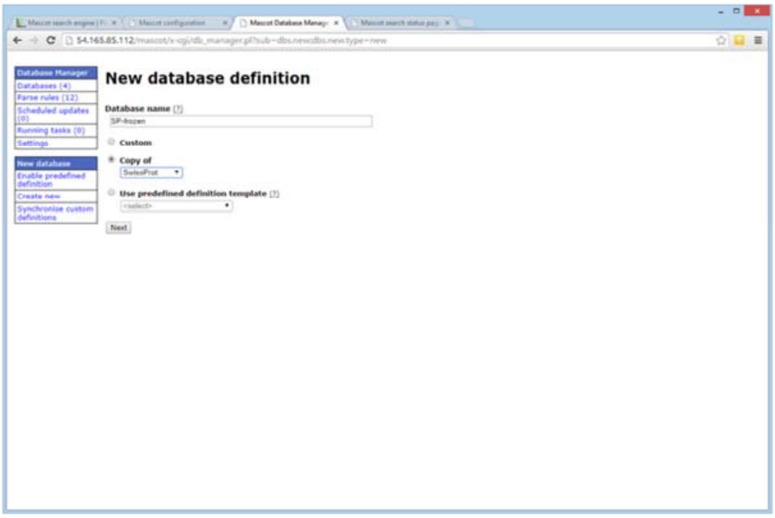
- ○ **None**
- ○ **Daily** at HH ▾ : MM ▾
- ● **Weekly** on Sunday ▾ at 02 ▾ : 00 ▾
- ○ **Monthly** on the <nth> ▾ at HH ▾ : MM ▾

Cancel Save

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



New database definition

Database name [?]
SP-frozen

○ Custom

● Copy of
SwissProt

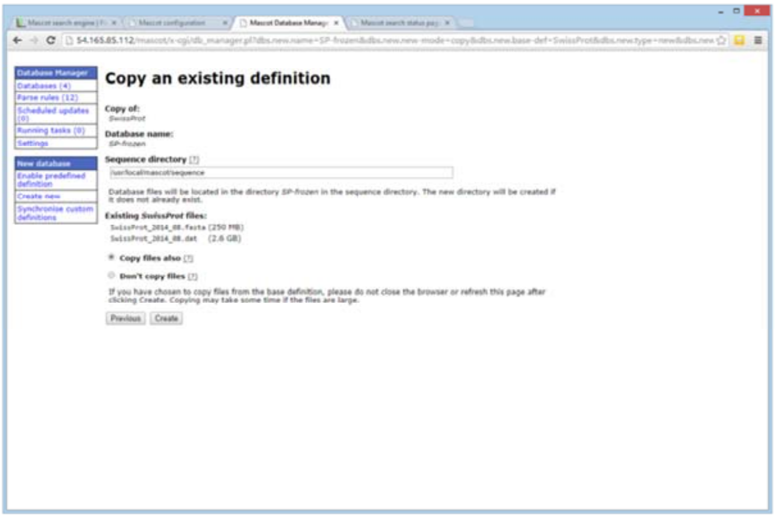
○ Use predefined definition template [?]
<select>

Next

MASCOT : *Sequence Database Administration* © 2007-2014 Matrix Science MATRIX SCIENCE

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



Copy an existing definition

Copy of:
SwissProt

Database name:
SP-frozen

Sequence directory [?]
/usr/local/mascot/sequence

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

Existing *SwissProt* files:
SwissProt_2014_08.fasta (250 MB)
SwissProt_2014_08.dat (2.6 GB)

● **Copy files also** [?]

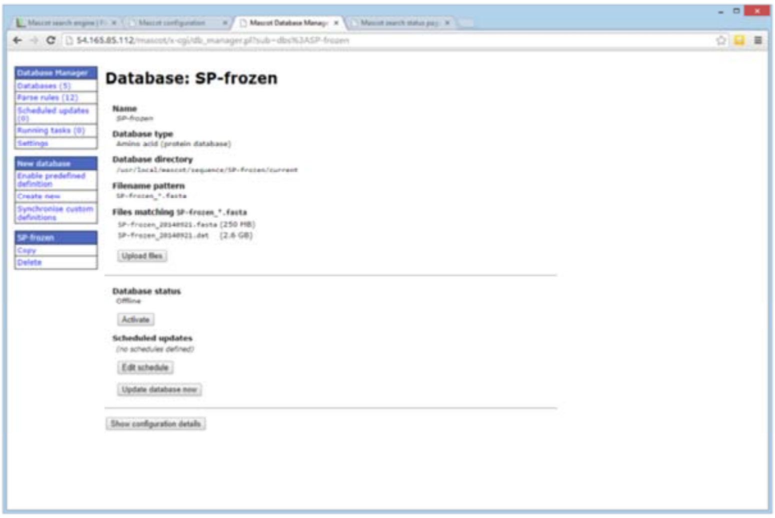
○ **Don't copy files** [?]

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.

MASCOT : *Sequence Database Administration* © 2007-2014 Matrix Science MATRIX SCIENCE

You are given the option to copy the existing files.

Create New - Copy of



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

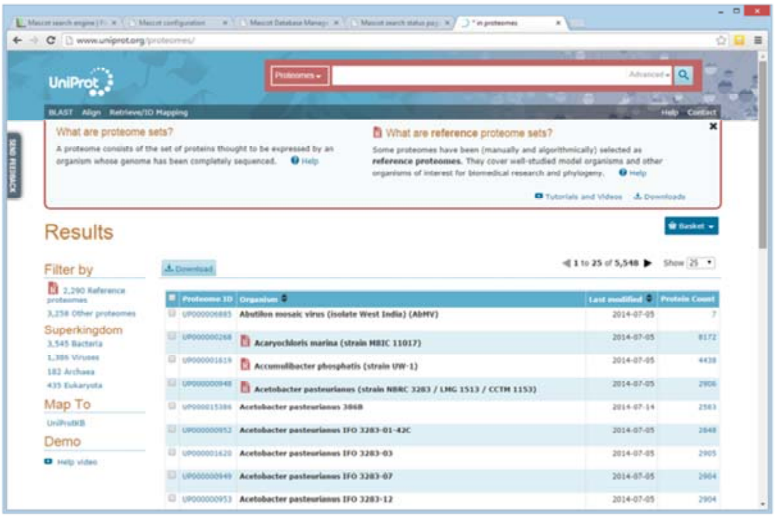
Files matching SP-frozen_*.fasta
SP-frozen_20140921.fasta (250 MB)
SP-frozen_20140921.dat (2.6 GB)

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



www.uniprot.org/proteomes/

What are proteome sets?

A proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced. Help

What are reference proteome sets?

Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny. Help

Results

Filter by
2,290 Reference proteomes
3,258 Other proteomes

Superkingdom
3,545 Bacteria
1,386 Viruses
182 Archaea
435 Eukaryota

Map To
UniProtKB

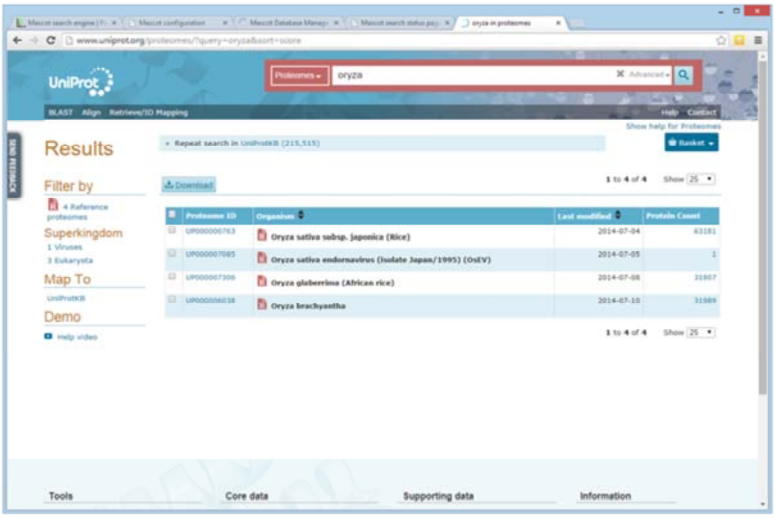
Demo
Help video

1 to 25 of 5,548 ▶ Show 25

Proteome ID	Organism	Last modified	Protein Count
UP000006885	Abutilon mosaic virus (isolate West India) (AbMV)	2014-07-05	7
UP000000268	Acaryochloris marina (strain MBIC 11017)	2014-07-05	8172
UP000001619	Accumulibacter phosphatis (strain UW-1)	2014-07-05	4438
UP000000948	Acetobacter pasteurianus (strain NBRC 3283 / LMG 1513 / CCTM 1153)	2014-07-05	2906
UP000015386	Acetobacter pasteurianus 386B	2014-07-14	2583
UP000000952	Acetobacter pasteurianus IFO 3283-01-42C	2014-07-05	2848
UP000001620	Acetobacter pasteurianus IFO 3283-03	2014-07-05	2905
UP000000949	Acetobacter pasteurianus IFO 3283-07	2014-07-05	2904
UP000000953	Acetobacter pasteurianus IFO 3283-12	2014-07-05	2904

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

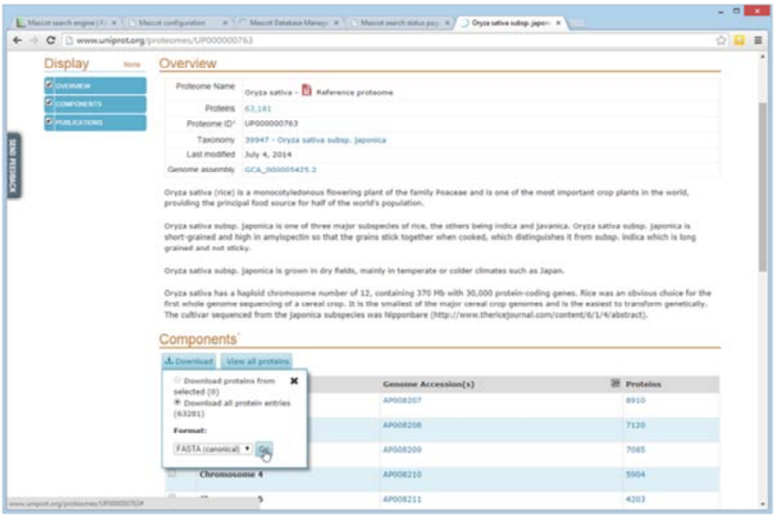
Create New - Predefined as Template



□	Proteome ID	Organism	Last modified	Protein Count
□	UP000000763	Oryza sativa subsp. japonica (Rice)	2014-07-04	63181
□	UP000007085	Oryza sativa endornavirus (Isolate Japan/1995) (OsEV)	2014-07-05	1
□	UP000007306	Oryza glaberrima (African rice)	2014-07-08	31807
□	UP000006038	Oryza brachyantha	2014-07-10	31989

There is a choice of three (one hit is a virus that infects rice). We'll choose Oryza sativa subsp. Japonica.

Create New - Predefined as Template



www.uniprot.org/proteomes/UP000000763

Overview

Proteome Name	Oryza sativa - Reference proteome
Proteins	63,181
Proteome ID	UP000000763
Taxonomy	39947 - Oryza sativa subsp. japonica
Last modified	July 4, 2014
Genome assembly	GCA_000005425.2

Oryza sativa (rice) is a monocotyledonous flowering plant of the family Poaceae and is one of the most important crop plants in the world, providing the principal food source for half of the world's population.

Oryza sativa subsp. japonica is one of three major subspecies of rice, the others being indica and javanica. Oryza sativa subsp. japonica is short-grained and high in amylopectin so that the grains stick together when cooked, which distinguishes it from subsp. indica which is long grained and not sticky.

Oryza sativa subsp. japonica is grown in dry fields, mainly in temperate or colder climates such as Japan.

Oryza sativa has a haploid chromosome number of 12, containing 370 Mb with 30,000 protein-coding genes. Rice was an obvious choice for the first whole genome sequencing of a cereal crop. It is the smallest of the major cereal crop genomes and is the easiest to transform genetically. The cultivar sequenced from the japonica subspecies was Nipponbare (http://www.thericejournal.com/content/6/1/4/abstract).

Components

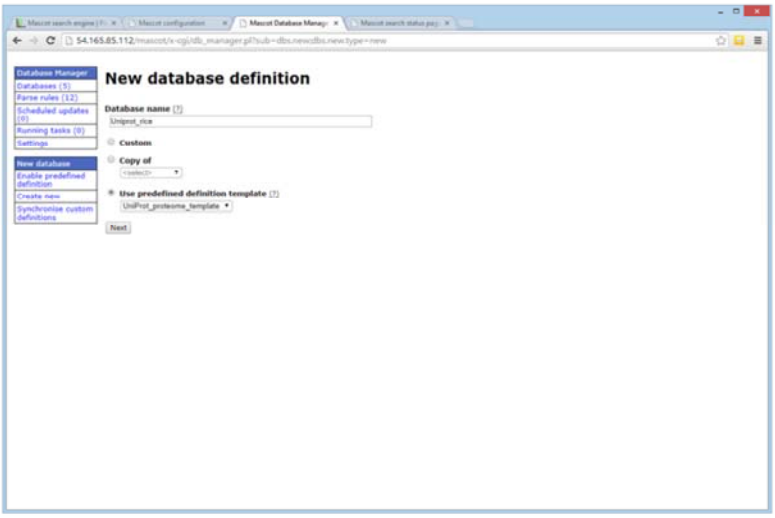
○ Download proteins from selected (0)
● Download all protein entries (63281)

	Genome Accession(s)	Proteins
	AP008207	8910
	AP008208	7120
	AP008209	7065
Chromosome 4	AP008210	5904
5	AP008211	4203

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 taxonomy ID for this strain - 39947

Create New - Predefined as Template



New database definition

Database name [?]

Uniprot_rice

○ Custom

○ Copy of

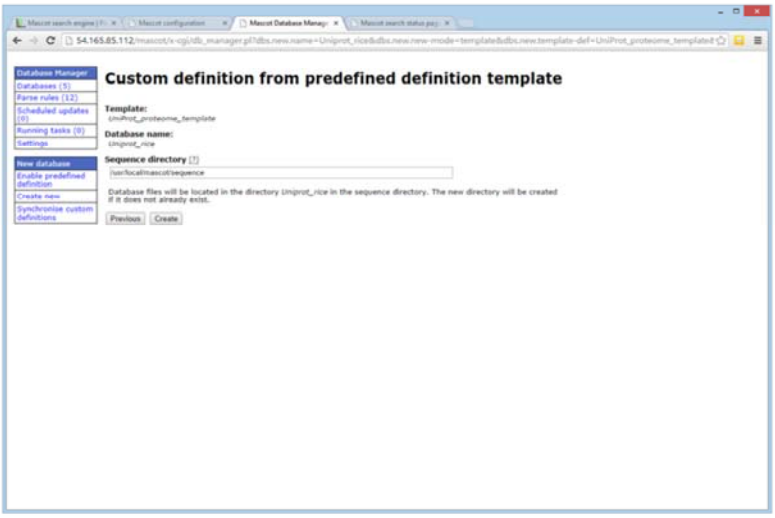
<select>

● Use predefined definition template [?]

UniProt_proteome_template

In Database Manager, choose Create new. Enter a suitable name and select the uniprot proteome predefined definition as a template.

Create New - Predefined as Template



Custom definition from predefined definition template

Template:
UniProt_proteome_template

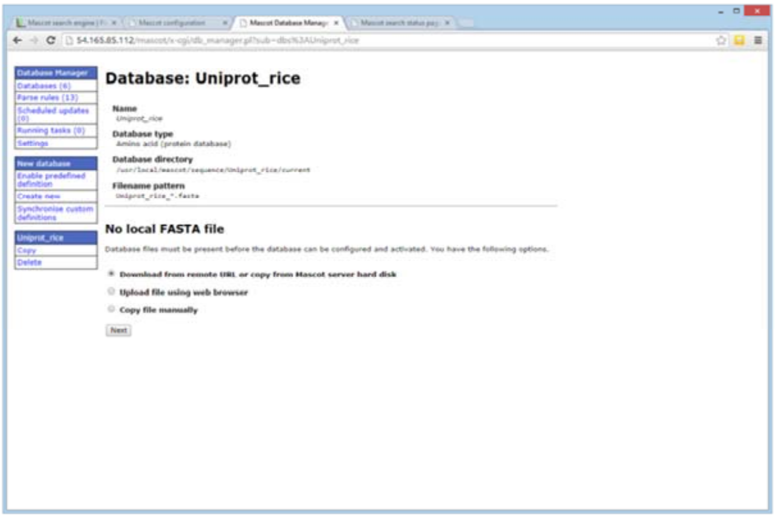
Database name:
Uniprot_rice

Sequence directory [?]
/usr/local/mascot/sequence

Database files will be located in the directory *Uniprot_rice* in the sequence directory. The new directory will be created if it does not already exist.

Choose Create

Create New - Predefined as Template



Database: Uniprot_rice

Name
Uniprot_rice

Database type
Amino acid (protein database)

Database directory
/usr/local/mascot/sequence/Uniprot_rice/current

Filename pattern
Uniprot_rice_*.fasta

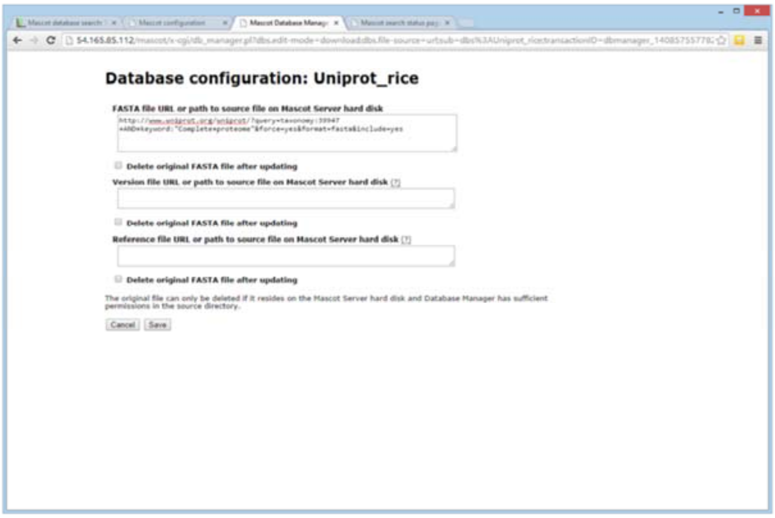
No local FASTA file

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 search
Mascot configuration
Mascot Database Manag
Mascot search status pag
54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.edit-mode=download;dbs.file-source=url;sub=dbs%3AUniprot_rice;transactionID=dbmanager_14085755778

Database configuration: Uniprot_rice

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

```
http://www.uniprot.org/uniprot/?query=taxonomy:39947
+AND+keyword:"Complete+proteome"&force=yes&format=fasta&include=yes
```

□ **Delete original FASTA file after updating**

Version file URL or path to source file on Mascot Server hard disk [?]

□ **Delete original FASTA file after updating**

Reference file URL or path to source file on Mascot Server hard disk [?]

□ **Delete original FASTA file after updating**

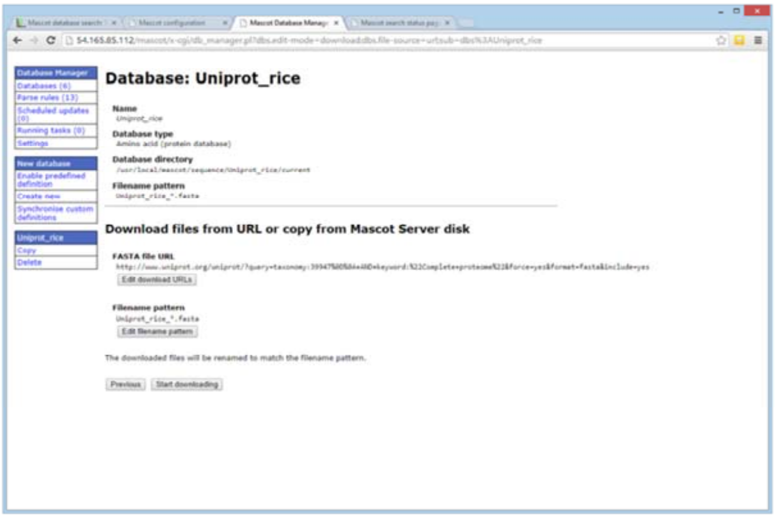
The original file can only be deleted if it resides on the Mascot Server hard disk and Database Manager has sufficient permissions in the source directory.

Cancel Save

MASCOT : *Sequence Database Administration* © 2007-2014 *Matrix Science* MATRIX SCIENCE

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 taxonomy 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



Database: Uniprot_rice

Name
Uniprot_rice

Database type
Amino acid (protein database)

Database directory
/usr/local/mascot/sequence/Uniprot_rice/current

Filename pattern
Uniprot_rice_*.fasta

Download files from URL or copy from Mascot Server disk

FASTA file URL
http://www.uniprot.org/uniprot/?query=taxonomy:39947%20%26+AND+keyword:%22Complete+proteome%22&force=yes&format=fasta&include=yes

Edit download URLs

Filename pattern
Uniprot_rice_*.fasta

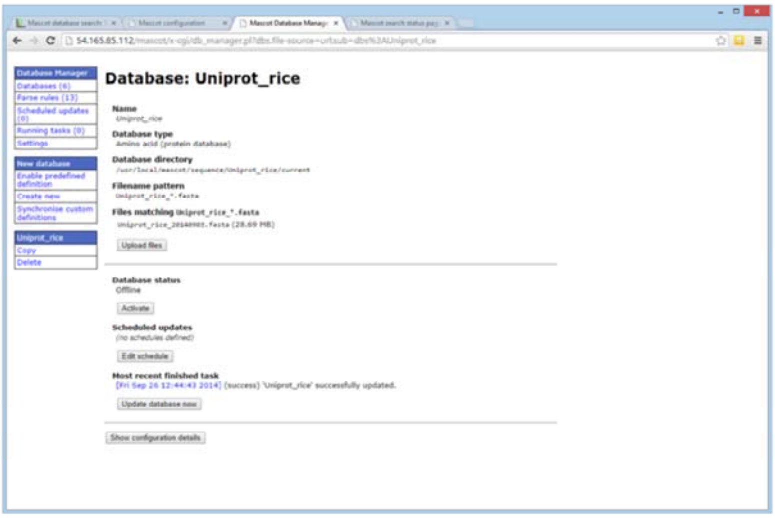
Edit filename pattern

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

Previous | Start downloading

Save, and we're ready to start downloading

Create New - Predefined as Template



54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.file-source=urlsub~dbs%3AUniprot_rice

Database: Uniprot_rice

Name
Uniprot_rice

Database type
Amino acid (protein database)

Database directory
/usr/local/mascot/sequence/Uniprot_rice/current

Filename pattern
Uniprot_rice_*.fasta

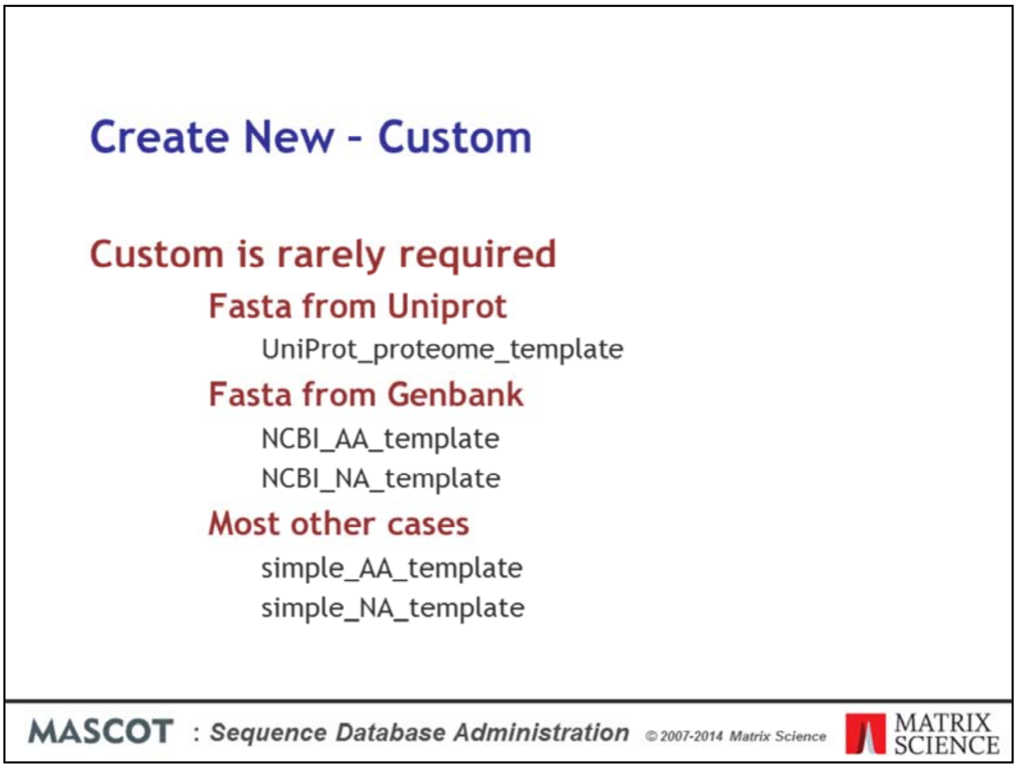
Files matching Uniprot_rice_*.fasta
Uniprot_rice_20140901.fasta (28.69 MB)

Database status
Offline

Scheduled updates
(no schedules defined)

Most recent finished task
[Fri Sep 26 12:44:43 2014] (success) 'Uniprot_rice' 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.



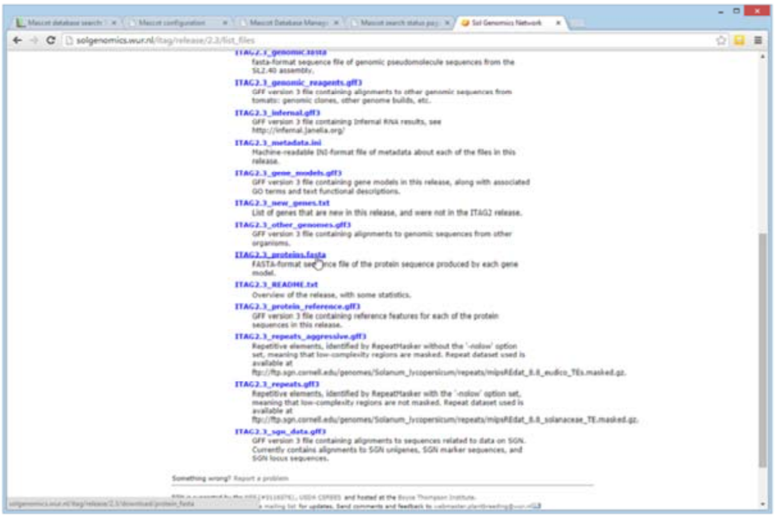
Create New - Custom

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



solgenomics.wur.nl/itag/release/2.3/list_files

ITAG2.3_genomic.fasta
fasta-format sequence file of genomic pseudomolecule sequences from the SL2.40 assembly.

ITAG2.3_genomic_reagents.gff3
GFF version 3 file containing alignments to other genomic sequences from tomato: genomic clones, other genome builds, etc.

ITAG2.3_infernal.gff3
GFF version 3 file containing Infernal RNA results, see http://infernal.janelia.org/

ITAG2.3_metadata.ini
Machine-readable INI-format file of metadata about each of the files in this release.

ITAG2.3_gene_models.gff3
GFF version 3 file containing gene models in this release, along with associated GO terms and text functional descriptions.

ITAG2.3_new_genes.txt
List of genes that are new in this release, and were not in the ITAG2 release.

ITAG2.3_other_genomes.gff3
GFF version 3 file containing alignments to genomic sequences from other organisms.

ITAG2.3_proteins.fasta
FASTA-format sequence file of the protein sequence produced by each gene model.

ITAG2.3_README.txt
Overview of the release, with some statistics.

ITAG2.3_protein_reference.gff3
GFF version 3 file containing reference features for each of the protein sequences in this release.

ITAG2.3_repeats_aggressive.gff3
Repetitive elements, identified by RepeatMasker without the '-nolow' option set, meaning that low-complexity regions are masked. Repeat dataset used is available at
ftp://ftp.sgn.cornell.edu/genomes/Solanum_lycopersicum/repeats/mipsREdat_8.8_eudico_TEs.masked.gz.

ITAG2.3_repeats.gff3
Repetitive elements, identified by RepeatMasker with the '-nolow' option set, meaning that low-complexity regions are not masked. Repeat dataset used is available at
ftp://ftp.sgn.cornell.edu/genomes/Solanum_lycopersicum/repeats/mipsREdat_8.8_solanaceae_TE.masked.gz.

ITAG2.3_sgn_data.gff3
GFF version 3 file containing alignments to sequences related to data on SGN. Currently contains alignments to SGN unigenes, SGN marker sequences, and SGN locus sequences.

Something wrong? Report a problem

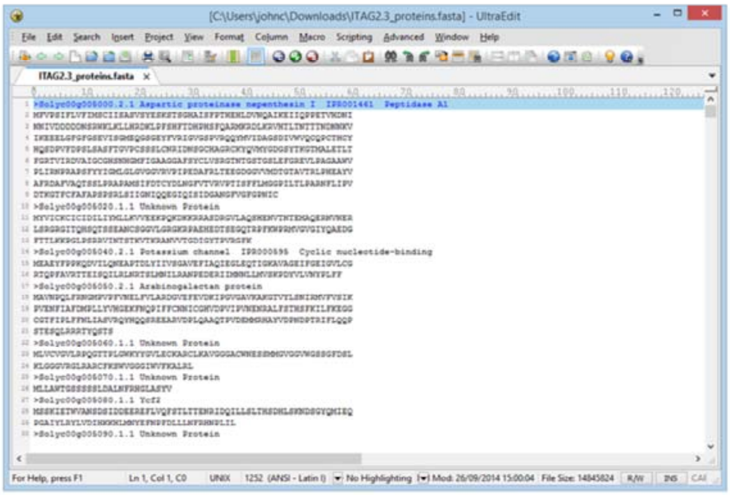
SGN is supported by the NSF (#0116076), USDA CSREES and hosted at the Boyce Thompson Institute.

solgenomics.wur.nl/itag/release/2.3/download/protein_fasta

a mailing list for updates. Send comments and feedback to webmaster.plantbreeding@wur.nl

But, to illustrate, lets configure a protein Fasta for the recently completed tomato genome. Choose a file and download it

Create New - Custom

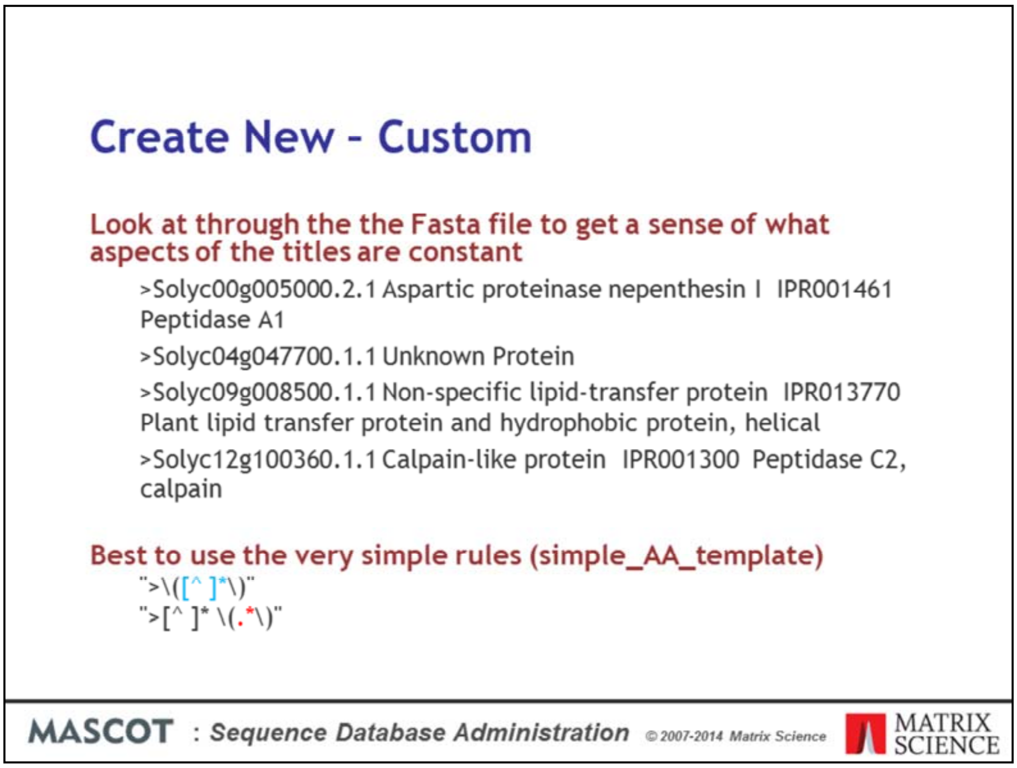


[C:\Users\johnc\Downloads\ITAG2.3_proteins.fasta] - UltraEdit

```
>Solyc00g005000.2.1 Aspartic proteinase nepenthesin I  IPR001461  Peptidase A1
MFVPSIFLVFIMSCIISASVSYEEKSTSGHAISFPTNENLDVNQAIKEIIQPPETVKDNI
NNKIVDDDDDNSRNKLKLLHRDKLPFSNFTDHPHSFQARMKRDLKRVNTLINTTTNDNNKV
IKEEELGFGFGSEVISGMEQGSGEYFVRIGVGSPVRQQYMVIDAGSDIVWVQCQPCTHCY
HQSDPVFDPSLSASFTGVPCSSSLCNRIDNSGCHAGRCKYQVMYGDGSYTKGTMALETLT
FGRTVIRDVAIGCGHSNNGMFIGAAGGAFSYCLVSRGTNTGSTGSLEFGREVLPAGAAWV
PLIRNPRAPSFYYIGMLGLGVGGVRVPIPEDAFRLTEEGDGGVVMDTGTAVTRLPHEAYV
AFRDAFVAQTSSLPRAPAMSIFDTCYDLNGFVTVRVPTISFFLMGGPILTLPARNFLIPV
DTKGTFCFAFAPSPSRLSIIGNIQQEGIQISIDGANGFVGFGPNIC
>Solyc00g005020.1.1 Unknown Protein
MYVICKCICIDILIYMLLMVVEEKPQKDKKRRASDRGVLAQSHEMVTNTEMAQERMVNER
LSRGRGITQRSQTSSEANCSGGVLGRGKRPAEHEDTSEGQTRPFKNPRMVGVGIYQAEDG
FTTLKKPGLPSRRVINTSTKVTKRANVVTGDIGYTPVRGFK
>Solyc00g005040.2.1 Potassium channel  IPR000595  Cyclic nucleotide-binding
MEAKYFPPKQDVILQNEAPTDLYIIVSGAVEFIAQIEGLEQTIGKAVAGEIFGEIGVLCG
RTQPFAVRTTEISQILRLNRTSLMNILRANPEDERIIMNNLLMVSKPDYVLVNYPLFF
>Solyc00g005050.2.1 Arabinogalactan protein
MAVNPQLFRNGMPVPFVNELFVLARDGVEFEVDKIPGVGAVKAKSTVTLSNIRMVFVSIK
PVENFIAFDMPLLYVNGEKFNQPIFFCNNICGMVDPVIPVNENRALFSTMSFKILFKEGG
CGTFIPLFFMLIASVRQYNQQSREEARVDPLQAAQTPVDEDGSHAYVDPNDPTRIFLQQP
STESQLRRRTVQSTS
>Solyc00g005060.1.1 Unknown Protein
MLVCVGVLRPQGTTPLGWKYYGVLECKARCLKAVGGGACWNESSMMGVGGVWGSSGFDSL
KLGGGVRGLRARCFKSWVGGGIWVFKALRL
>Solyc00g005070.1.1 Unknown Protein
MLLANTGSSSSSLDALNFRNGLASYV
>Solyc00g005080.1.1 Ycf2
MSSKISTWVANSDSIDDEEREFLVQFSTLTTENRIDQILLSLTHSDHLSNNDSGYQMIEQ
PGAIYLRYLVDIHKKNLMNYEFNPFDLLLNFSHNSLIL
>Solyc00g005090.1.1 Unknown Protein
```

For Help, press F1 | Ln 1, Col 1, C0 | UNIX | 1252 (ANSI - Latin I) | No Highlighting | Mod: 26/09/2014 15:00:04 | File Size: 14845824 | R/W | INS | CAP

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

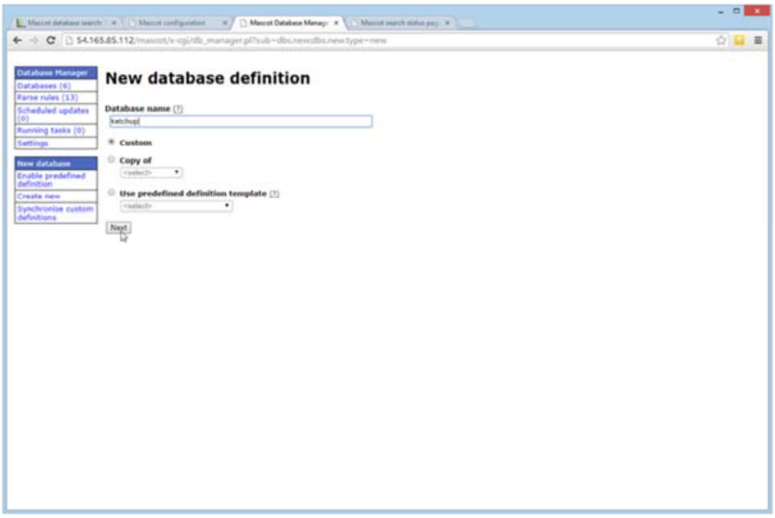
```
>Solyc00g005000.2.1 Aspartic proteinase nepenthesin I  IPR001461
Peptidase A1
>Solyc04g047700.1.1 Unknown Protein
>Solyc09g008500.1.1 Non-specific lipid-transfer protein  IPR013770
Plant lipid transfer protein and hydrophobic protein, helical
>Solyc12g100360.1.1 Calpain-like protein  IPR001300  Peptidase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



http://54.165.85.112/mascot/x-cgi/db_manager.pl?sub=dbs.newdbs.new.type=new

New database definition

Database name [?]

katchup

● Custom

○ Copy of

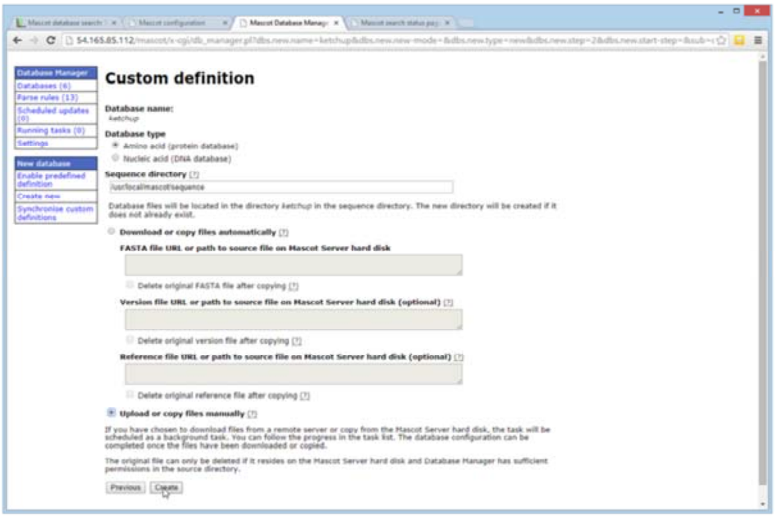
<select>

○ Use predefined definition template [?]

<select>

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

Create New - Custom



Custom definition

Database name:
ketchup

Database type

● Amino acid (protein database)
○ Nucleic acid (DNA database)

Sequence directory [?]
/usr/local/mascot/sequence

Database files will be located in the directory *ketchup* in the sequence 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 [?]

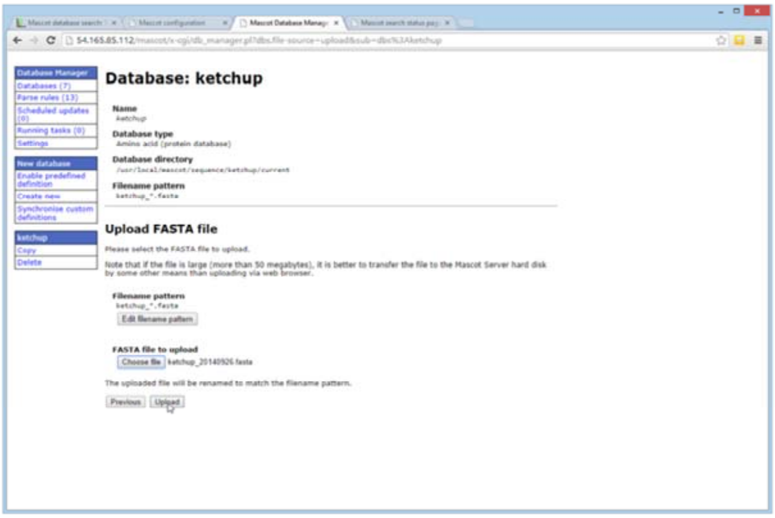
● **Upload or copy files manually** [?]

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. The database 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.

This is an Amino Acid database and we already have the Fasta file. Set these options and choose Create

Create New - Custom



Database: ketchup

Name
ketchup

Database type
Amino acid (protein database)

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

Filename pattern
ketchup_*.fasta

Upload FASTA file

Please select the FASTA file to upload.

Note that if the file is large (more than 50 megabytes), it is better to transfer the file to the Mascot Server hard disk by some other means than uploading via web browser.

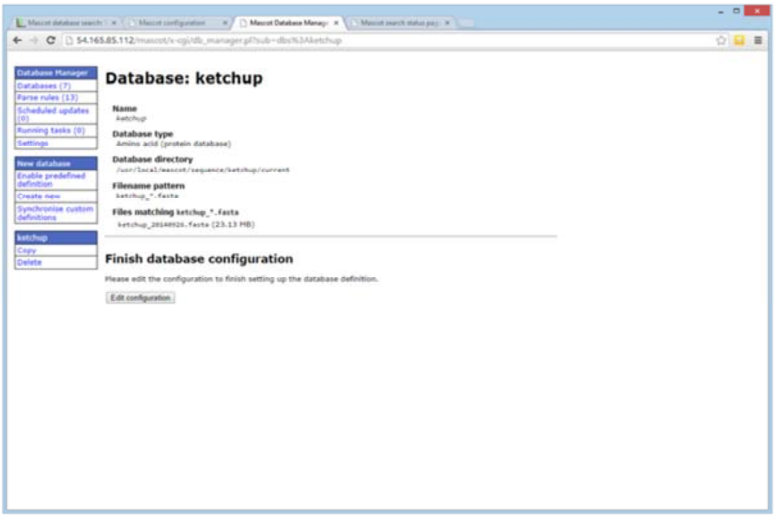
Filename pattern
ketchup_*.fasta

FASTA file to upload
Choose file ketchup_20140926.fasta

The uploaded file will be renamed to match the filename pattern.

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

Create New - Custom



Database: ketchup

Name
ketchup

Database type
Amino acid (protein database)

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

Filename pattern
ketchup_*.fasta

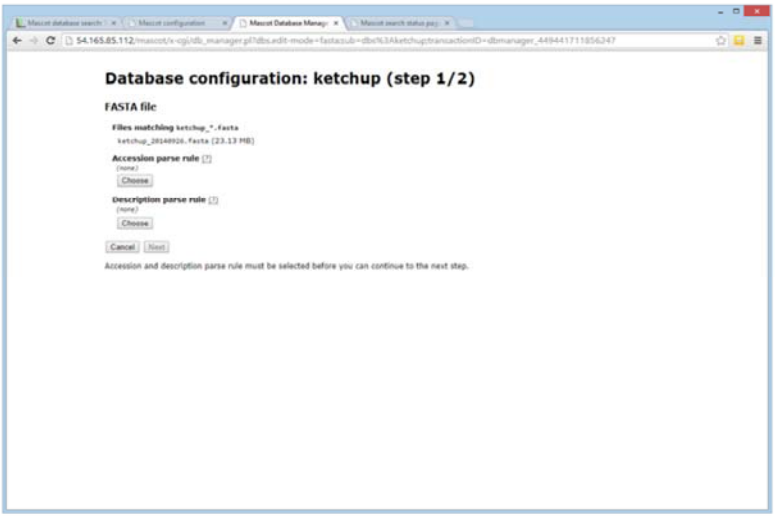
Files matching ketchup_*.fasta
ketchup_20140926.fasta (23.13 MB)

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 search
Mascot configuration
Mascot Database Manager
Mascot search status page
54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.edit-mode=fastasub=dbs%3Aketchup;transactionID=dbmanager_449441711856247

Database configuration: ketchup (step 1/2)

FASTA file

Files matching `ketchup_*.fasta`

`ketchup_20140926.fasta` (23.13 MB)

Accession parse rule [?]
(none)
Choose

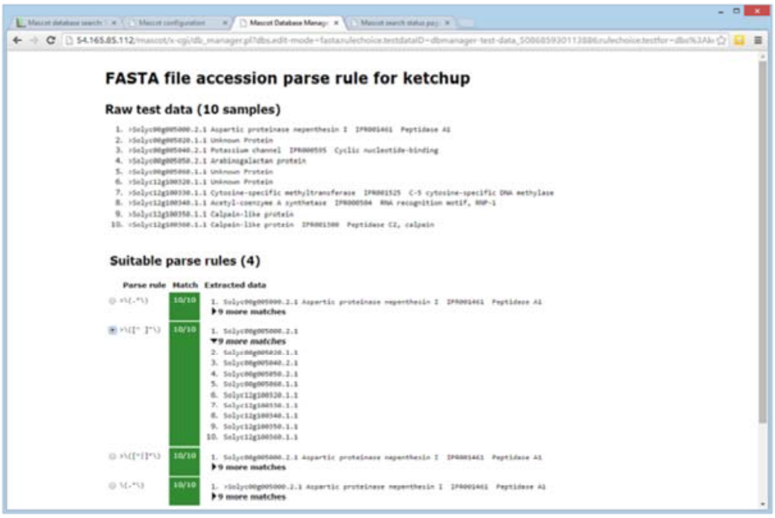
Description parse rule [?]
(none)
Choose

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 search
Mascot configuration
Mascot Database Manage
Mascot search status pag
54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.edit-mode=fasta.rulechoice.testdataID=dbmanager-test-data_50668593011388&rulechoice.testfor=dbs%3Ak

FASTA file accession parse rule for ketchup

Raw test data (10 samples)

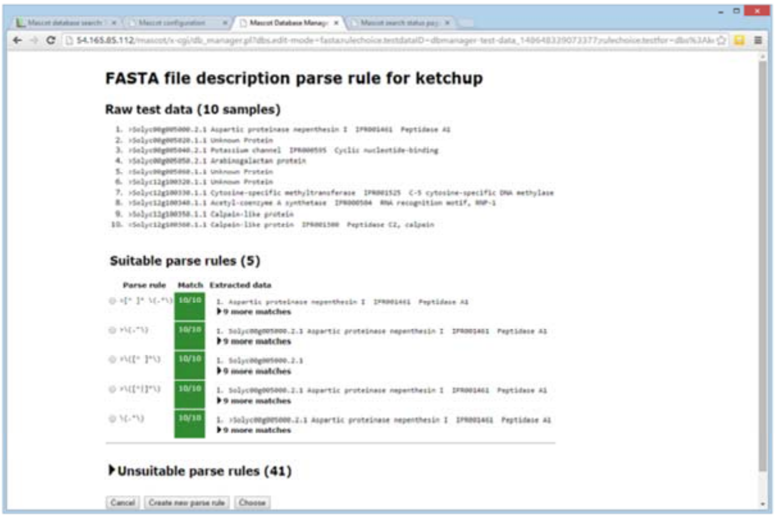
```
 1. >Solyc00g005000.2.1 Aspartic proteinase nepenthesin I  IPR001461  Peptidase A1
 2. >Solyc00g005020.1.1 Unknown Protein
 3. >Solyc00g005040.2.1 Potassium channel  IPR000595  Cyclic nucleotide-binding
 4. >Solyc00g005050.2.1 Arabinogalactan protein
 5. >Solyc00g005060.1.1 Unknown Protein
 6. >Solyc12g100320.1.1 Unknown Protein
 7. >Solyc12g100330.1.1 Cytosine-specific methyltransferase  IPR001525  C-5 cytosine-specific DNA methylase
 8. >Solyc12g100340.1.1 Acetyl-coenzyme A synthetase  IPR000504  RNA recognition motif, RNP-1
 9. >Solyc12g100350.1.1 Calpain-like protein
10. >Solyc12g100360.1.1 Calpain-like protein  IPR001300  Peptidase C2, calpain
```

Suitable parse rules (4)

Parse rule	Match	Extracted data
○ >\(.*\)	10/10	1. Solyc00g005000.2.1 Aspartic proteinase nepenthesin I IPR001461 Peptidase A1 ▶9 more matches
● >\([^]*\)	10/10	1. Solyc00g005000.2.1 ▼9 more matches 2. Solyc00g005020.1.1 3. Solyc00g005040.2.1 4. Solyc00g005050.2.1 5. Solyc00g005060.1.1 6. Solyc12g100320.1.1 7. Solyc12g100330.1.1 8. Solyc12g100340.1.1 9. Solyc12g100350.1.1 10. Solyc12g100360.1.1
○ >\([^|]*\)	10/10	1. Solyc00g005000.2.1 Aspartic proteinase nepenthesin I IPR001461 Peptidase A1 ▶9 more matches
○ \(.*\)	10/10	1. >Solyc00g005000.2.1 Aspartic proteinase nepenthesin I IPR001461 Peptidase A1 ▶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 the most suitable

Create New - Custom



Mascot database search
Mascot configuration
Mascot Database Manag
Mascot search status pa
54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.edit-mode=fasta.rulechoice.testdataID=dbmanager-test-data_1486483390733778rulechoice.testfor=dbs%3Ak

FASTA file description parse rule for ketchup

Raw test data (10 samples)

```
 1. >Solyc00g005000.2.1 Aspartic proteinase nepenthesin I  IPR001461  Peptidase A1
 2. >Solyc00g005020.1.1 Unknown Protein
 3. >Solyc00g005040.2.1 Potassium channel  IPR000595  Cyclic nucleotide-binding
 4. >Solyc00g005050.2.1 Arabinogalactan protein
 5. >Solyc00g005060.1.1 Unknown Protein
 6. >Solyc12g100320.1.1 Unknown Protein
 7. >Solyc12g100330.1.1 Cytosine-specific methyltransferase  IPR001525  C-5 cytosine-specific DNA methylase
 8. >Solyc12g100340.1.1 Acetyl-coenzyme A synthetase  IPR000504  RNA recognition motif, RNP-1
 9. >Solyc12g100350.1.1 Calpain-like protein
10. >Solyc12g100360.1.1 Calpain-like protein  IPR001300  Peptidase C2, calpain
```

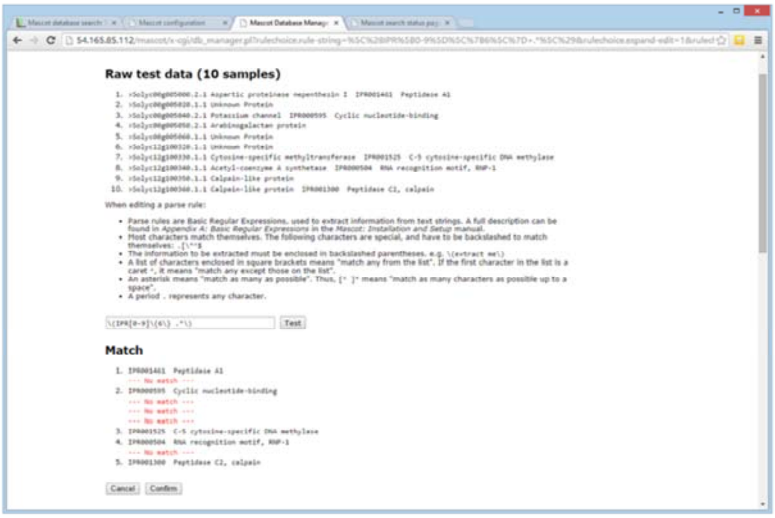
Suitable parse rules (5)

Parse rule	Match	Extracted data
○ >[^]* \(.*\)	10/10	1. Aspartic proteinase nepenthesin I IPR001461 Peptidase A1 ▶9 more matches
○ >\(.*\)	10/10	1. Solyc00g005000.2.1 Aspartic proteinase nepenthesin I IPR001461 Peptidase A1 ▶9 more matches
○ >\([^]*\)	10/10	1. Solyc00g005000.2.1 ▶9 more matches
○ >\([^|]*\)	10/10	1. Solyc00g005000.2.1 Aspartic proteinase nepenthesin I IPR001461 Peptidase A1 ▶9 more matches
○ \(.*\)	10/10	1. >Solyc00g005000.2.1 Aspartic proteinase nepenthesin I IPR001461 Peptidase A1 ▶9 more matches

▶Unsuitable parse rules (41)

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 search
Mascot configuration
Mascot Database Manag
Mascot search status pag
54.165.85.112/mascot/x-cgi/db_manager.pl?rulechoice.rule-string=%5C%28IPR%5B0-9%5D%5C%7B6%5C%7D+.*%5C%29&rulechoice.expand-edit=1&ruleId

Raw test data (10 samples)

```
1. >Solyc00g005000.2.1 Aspartic proteinase nepenthesin I  IPR001461  Peptidase A1
2. >Solyc00g005020.1.1 Unknown Protein
3. >Solyc00g005040.2.1 Potassium channel  IPR000595  Cyclic nucleotide-binding
4. >Solyc00g005050.2.1 Arabinogalactan protein
5. >Solyc00g005060.1.1 Unknown Protein
6. >Solyc12g100320.1.1 Unknown Protein
7. >Solyc12g100330.1.1 Cytosine-specific methyltransferase  IPR001525  C-5 cytosine-specific DNA methylase
8. >Solyc12g100340.1.1 Acetyl-coenzyme A synthetase  IPR000504  RNA recognition motif, RNP-1
9. >Solyc12g100350.1.1 Calpain-like protein
10. >Solyc12g100360.1.1 Calpain-like protein  IPR001300  Peptidase C2, calpain
```

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.

`\(IPR[0-9]\{6\} .*\)` Test

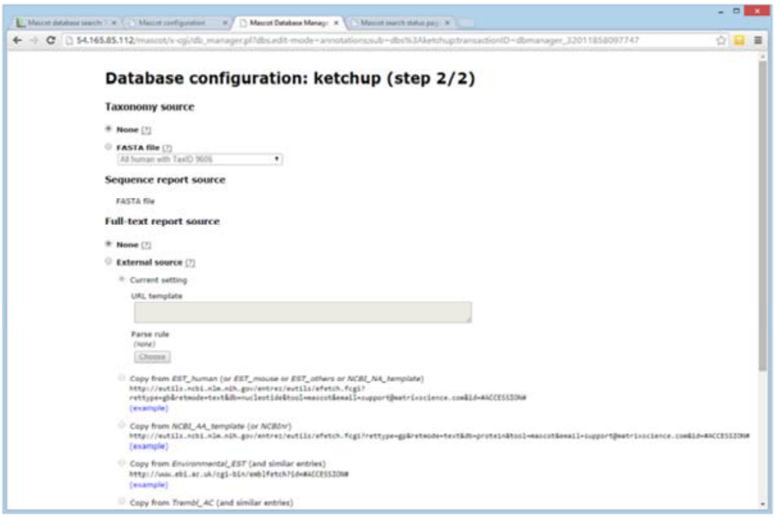
Match

```
1. IPR001461  Peptidase A1
   --- No match ---
2. IPR000595  Cyclic nucleotide-binding
   --- No match ---
   --- No match ---
   --- No match ---
3. IPR001525  C-5 cytosine-specific DNA methylase
4. IPR000504  RNA recognition motif, RNP-1
   --- No match ---
5. IPR001300  Peptidase C2, calpain
```

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 appears to work. It locks on to the text IPR followed by 6 digits, then takes everything from that point to the end of the line. This is actually not a good rule, because not all titles include an Interpro reference, 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 search
Mascot configuration
Mascot Database Manag
Mascot search status pag
54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.edit-mode=annotations;sub=dbs%3Aketchup;transactionID=dbmanager_32011858097747

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)

Choose

○ Copy from *EST_human* (or *EST_mouse* or *EST_others* or *NCBI_NA_template*)
http://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 *NCBI_AA_template* (or *NCBInr*)
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?rettype=gp&retmode=text&db=protein&tool=mascot&email=support@matrixscience.com&id=#ACCESSION#
[example]

○ Copy from *Environmental_EST* (and similar entries)
http://www.ebi.ac.uk/cgi-bin/emblfetch?id=#ACCESSION#
[example]

○ Copy from *Trembl_AC* (and similar entries)

The next step deals with taxonomy and annotations.

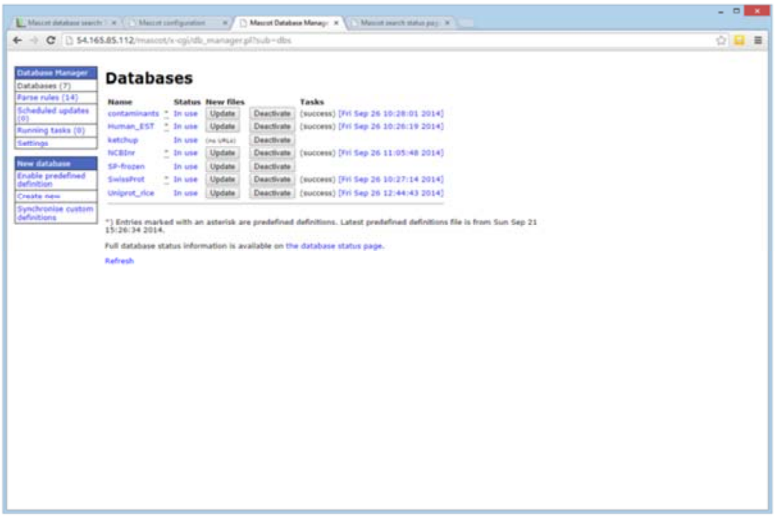
If this was a comprehensive database, containing entries from many different organisms, we might want to select a rule for determining the taxonomy of each entry. There is a drop down list of rules for the major public databases. Most of these require additional files, which can be downloaded automatically 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

Name		Status	New files		Tasks
contaminants	*	In use	Update	Deactivate	(success) [Fri Sep 26 10:28:01 2014]
Human_EST	*	In use	Update	Deactivate	(success) [Fri Sep 26 10:26:19 2014]
ketchup		In use	(no URLs)	Deactivate	
NCBInr	*	In use	Update	Deactivate	(success) [Fri Sep 26 11:05:48 2014]
SP-frozen		In use	Update	Deactivate	
SwissProt	*	In use	Update	Deactivate	(success) [Fri Sep 26 10:27:14 2014]
Uniprot_rice		In use	Update	Deactivate	(success) [Fri Sep 26 12:44:43 2014]

*) Entries marked with an asterisk are predefined definitions. Latest predefined definitions file is from Sun Sep 21 15:26:34 2014.

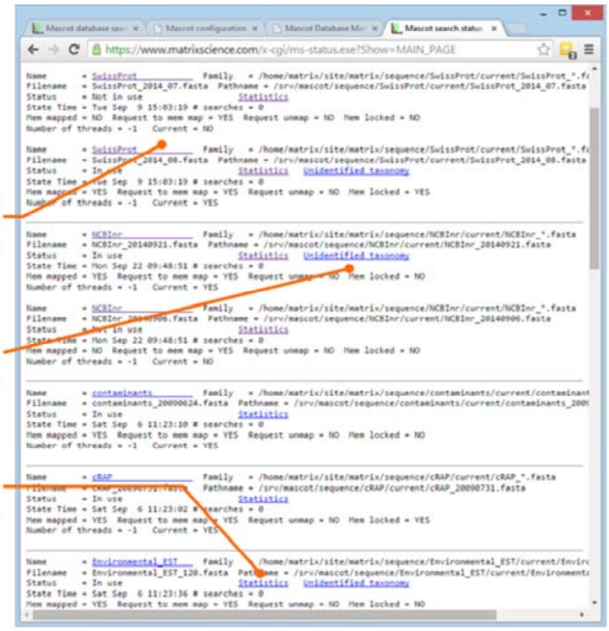
Full database status information is available on the database status page.

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



```
Name       = SwissProt              Family   = /home/matrix/site/matrix/sequence/SwissProt/current/SwissProt_*.fa
Filename   = SwissProt_2014_07.fasta  Pathname = /srv/mascot/sequence/SwissProt/current/SwissProt_2014_07.fasta
Status     = Not in use                      Statistics
State Time = Tue Sep  9 15:03:19 # searches = 0
Mem mapped = NO  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = NO

Name       = SwissProt              Family   = /home/matrix/site/matrix/sequence/SwissProt/current/SwissProt_*.fa
Filename   = SwissProt_2014_08.fasta  Pathname = /srv/mascot/sequence/SwissProt/current/SwissProt_2014_08.fasta
Status     = In use                          Statistics  Unidentified taxonomy
State Time = Tue Sep  9 15:03:19 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = YES
Number of threads = -1   Current = YES

Name       = NCBInr                 Family   = /home/matrix/site/matrix/sequence/NCBInr/current/NCBInr_*.fasta
Filename   = NCBInr_20140921.fasta  Pathname = /srv/mascot/sequence/NCBInr/current/NCBInr_20140921.fasta
Status     = In use                          Statistics  Unidentified taxonomy
State Time = Mon Sep 22 09:48:51 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = YES

Name       = NCBInr                 Family   = /home/matrix/site/matrix/sequence/NCBInr/current/NCBInr_*.fasta
Filename   = NCBInr_20140906.fasta  Pathname = /srv/mascot/sequence/NCBInr/current/NCBInr_20140906.fasta
Status     = Not in use                      Statistics
State Time = Mon Sep 22 09:48:51 # searches = 0
Mem mapped = NO  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = NO

Name       = contaminants           Family   = /home/matrix/site/matrix/sequence/contaminants/current/contaminant
Filename   = contaminants_20090624.fasta  Pathname = /srv/mascot/sequence/contaminants/current/contaminants_200
Status     = In use                          Statistics
State Time = Sat Sep  6 11:23:10 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
Number of threads = -1   Current = YES

Name       = cRAP                   Family   = /home/matrix/site/matrix/sequence/cRAP/current/cRAP_*.fasta
Filename   = cRAP_20090731.fasta    Pathname = /srv/mascot/sequence/cRAP/current/cRAP_20090731.fasta
Status     = In use                          Statistics
State Time = Sat Sep  6 11:23:02 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = YES
Number of threads = -1   Current = YES

Name       = Environmental_EST      Family   = /home/matrix/site/matrix/sequence/Environmental_EST/current/Envir
Filename   = Environmental_EST_120.fasta  Pathname = /srv/mascot/sequence/Environmental_EST/current/Environment
Status     = In use                          Statistics  Unidentified taxonomy
State Time = Sat Sep  6 11:23:36 # searches = 0
Mem mapped = YES  Request to mem map = YES  Request unmap = NO  Mem locked = NO
```

"Old" & "New"

Unidentified taxonomy

Statistics

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

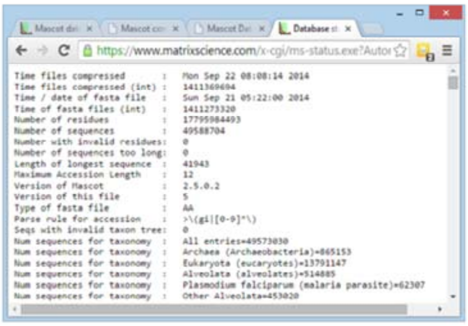
Initially, there will be a single information block for each database on this page. When a database is updated, a second information block is added. One is for the new or incoming database, the other is for the old or outgoing. If all is well, one of the pair will have the status of "In use", and the 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.

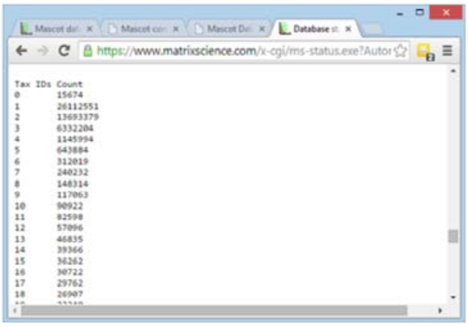
The database statistics are very useful for diagnosing problems and checking up on the health of a database

Database Statistics

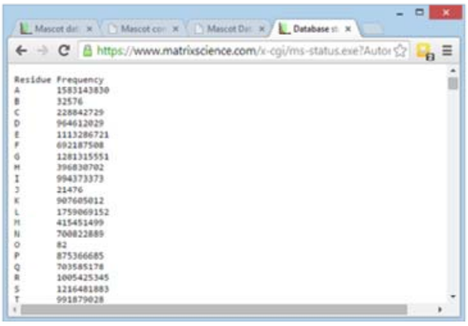
- Is the number of entries correct?
- Any invalid codes?
- Any entries “too long”?
- Is an AA database all ACGT?
- If using taxonomy, is the success rate > 99%?



```
Time files compressed      :   Mon Sep 22 08:08:14 2014
Time files compressed (int) :  1411369694
Time / date of fasta file   :  Sun Sep 21 05:22:00 2014
Time of fasta files (int)   :  1411273320
Number of residues          :  17795984493
Number of sequences         :  49588704
Number with invalid residues:  0
Number of sequences too long:  0
Length of longest sequence  :  41943
Maximum Accession Length    :  12
Version of Mascot           :  2.5.0.2
Version of this file        :  5
Type of fasta file          :  AA
Parse rule for accession    :  >\(gi|[0-9]*\)
Seqs with invalid taxon tree:  0
Num sequences for taxonomy  :  All entries=49573030
Num sequences for taxonomy  :  Archaea (Archaeobacteria)=865153
Num sequences for taxonomy  :  Eukaryota (eucaryotes)=13791147
Num sequences for taxonomy  :  Alveolata (alveolates)=514885
Num sequences for taxonomy  :  Plasmodium falciparum (malaria parasite)=62307
Num sequences for taxonomy  :  Other Alveolata=453020
```



```
Tax IDs Count
0       15674
1       26112551
2       13693379
3       6332204
4       1145994
5       643884
6       312019
7       240232
8       148314
9       117063
10      90922
11      82598
12      57096
13      46835
14      39366
15      36262
16      30722
17      29762
18      26907
```



```
Residue Frequency
A       1583143830
B       32576
C       228842729
D       964612029
E       1113286721
F       692187508
G       1281315551
H       396830702
I       994373373
J       21476
K       907605012
L       1759069152
M       415451499
N       700822889
O       82
P       875366685
Q       703585178
R       1005425345
S       1216481883
T       991879028
```

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

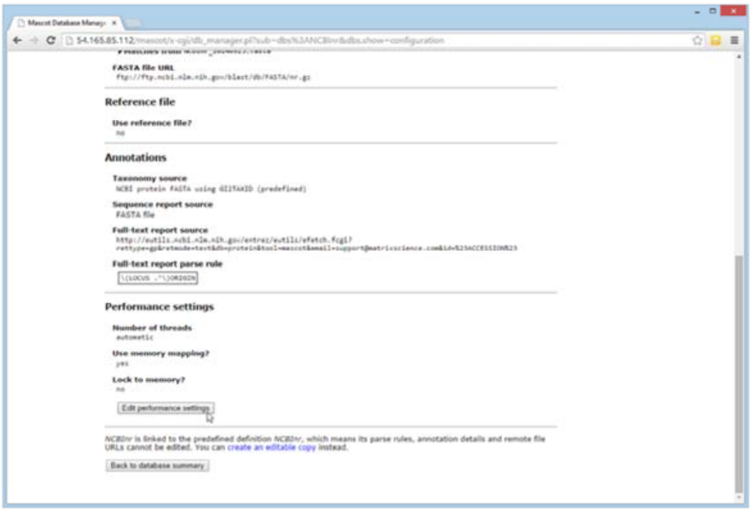
Are there any invalid characters in the sequences? If there are, this should definitely be investigated

Mascot has a parameter, MaxSequenceLen, to set the length of the longest sequence. The default is 50,000. 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 Manag
54.165.85.112/mascot/x-cgi/db_manager.pl?sub=dbs%3ANCBInr&dbs.show=configuration

FASTA file URL
ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

Reference file

Use reference file?
no

Annotations

Taxonomy source
NCBI protein FASTA using GI2TAXID (predefined)

Sequence report source
FASTA file

Full-text report source
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?
rettype=gp&retmode=text&db=protein&tool=mascot&email=support@matrixscience.com&id=%23ACCESSION%23

Full-text report parse rule
\(LOCUS .*\)ORIGIN

Performance settings

Number of threads
automatic

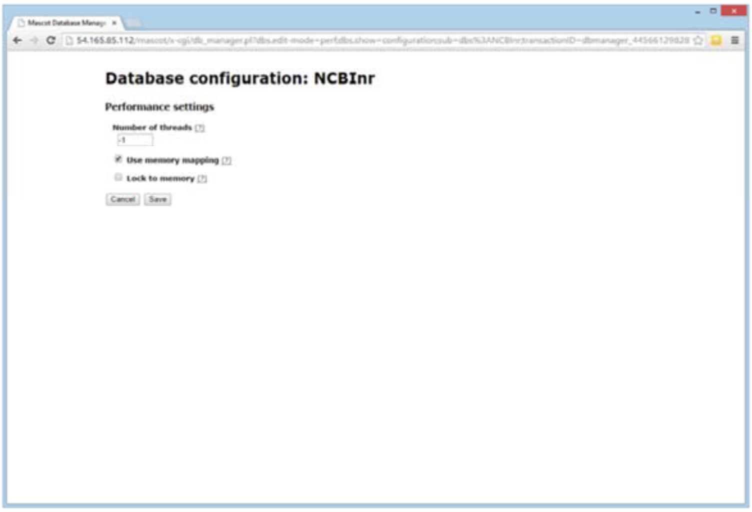
Use memory mapping?
yes

Lock to memory?
no

NCBInr is linked to the predefined definition NCBInr, which means its parse rules, annotation details and remote file URLs cannot be edited. You can create an editable copy instead.

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

Configuration - Performance


Mascot Database Manag
54.165.85.112/mascot/x-cgi/db_manager.pl?dbs.edit-mode=perf;dbs.show=configuration;sub=dbs%3ANCBInr;transactionID=dbmanager_44566129828

Database configuration: NCBInr

Performance settings

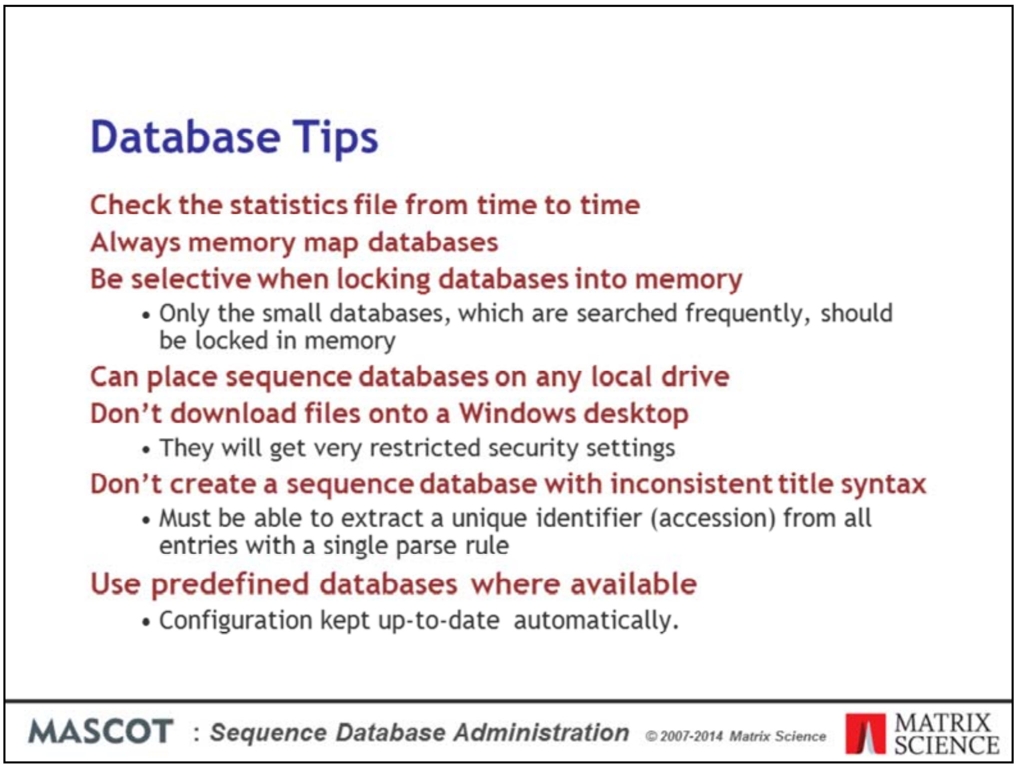
Number of threads [?]
-1

☑ 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.



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.

MASCOT : *Sequence Database Administration* © 2007-2014 Matrix Science MATRIX SCIENCE

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.