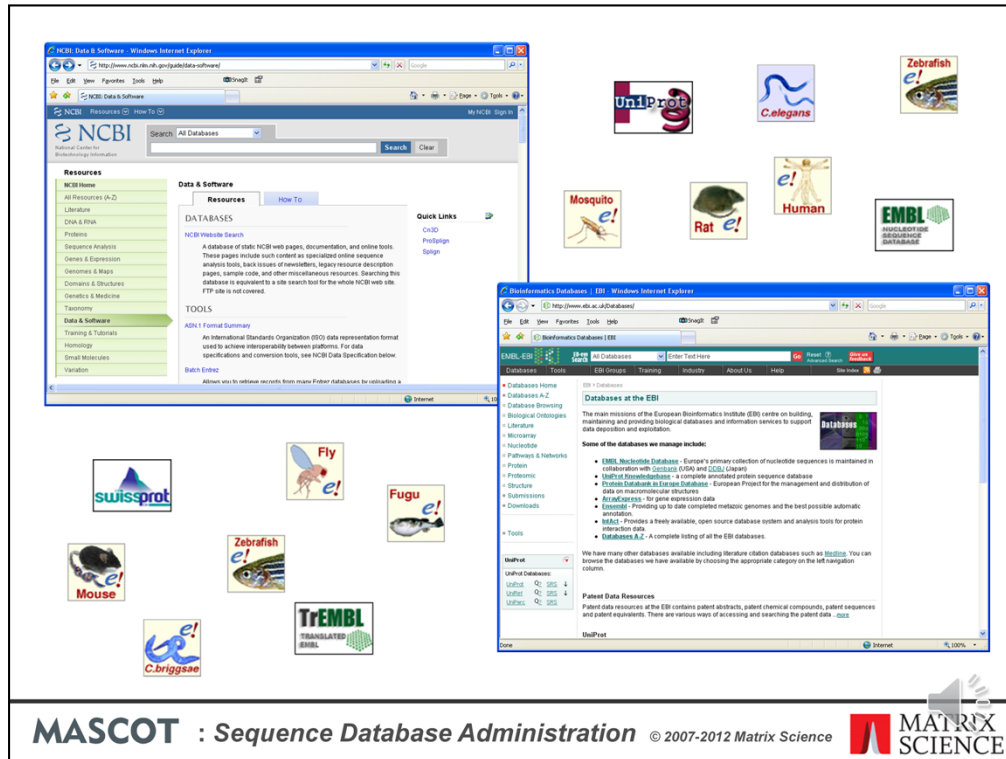


Sequence Database Administration

MASCOT





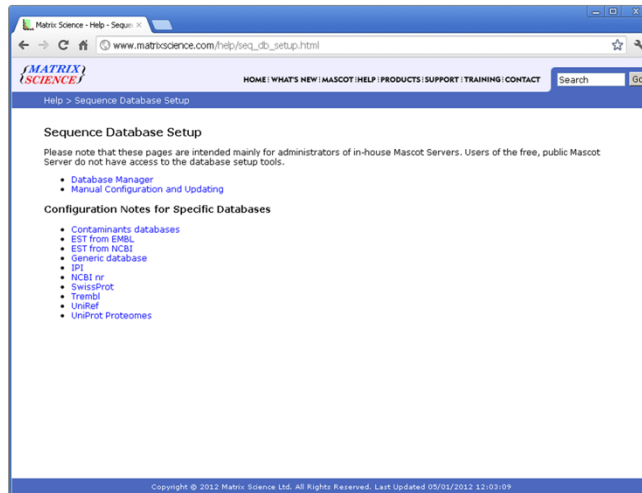
When you install Mascot, it includes a copy of the Swiss-Prot protein database. However, it is almost certain that you and your colleagues will want to search other databases as well. There are very many to choose from, and Mascot allows you to have as many databases as you wish on-line for searching 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 : *Sequence Database Administration* © 2007-2012 Matrix Science



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 The help pages in your in-house copy of Mascot are similar, but become progressively out-of-date.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

Must have local Fasta file

- (Mascot streams through the database during each search)

Mascot can search any database available in Fasta format

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

Other files are optional

- Taxonomy indexes
- Full text annotations.

MASCOT : Sequence Database Administration © 2007-2012 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* © 2007-2012 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* © 2007-2012 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.

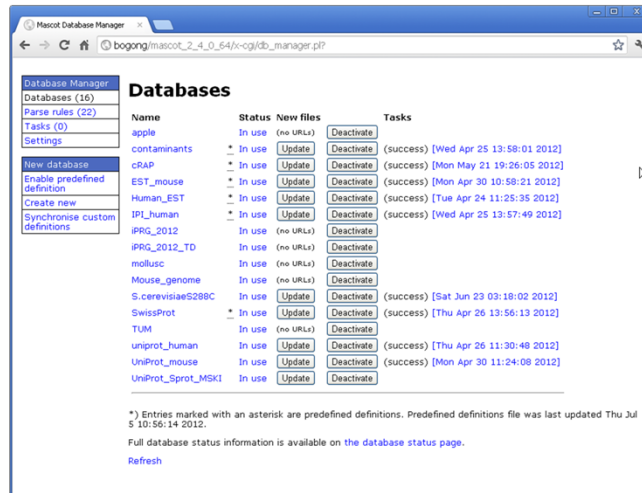
Here, for example, we have a title line from the IPI human database. Let's say that we want to use IPI00043251 as the unique accession string and everything after the first space should be treated as the description.

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.

Mascot 2.4: Database Manager



MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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:

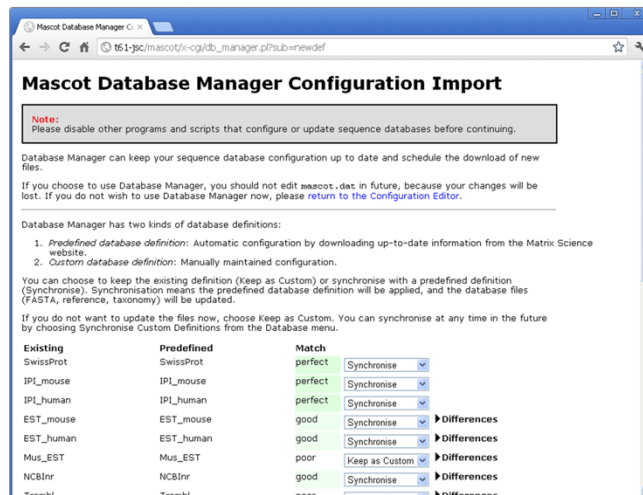
A Predefined Database Definition is one in which the configuration information is kept up-to-date on the Matrix Science web site, and downloaded as required by Database Manager. You don't need to know file URLs or worry about parse rules, etc. for a Predefined Database.

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 : Sequence Database Administration © 2007-2012 Matrix Science

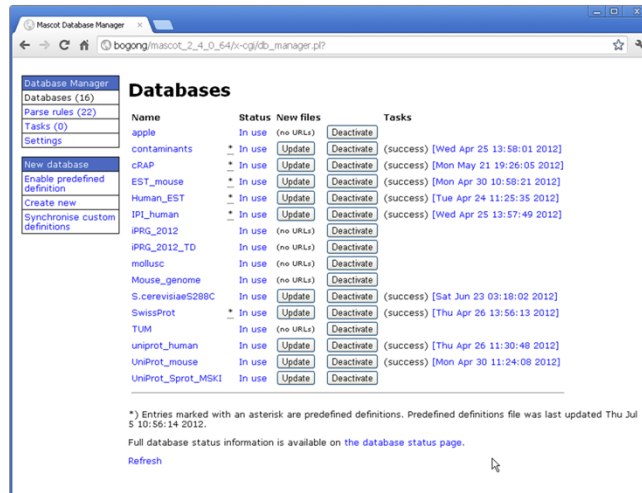


Database Manager must be allowed exclusive control of database configuration. Editing mascot.dat outside of Database Manager will just cause confusion because Database Manager re-writes mascot.dat whenever a configuration changes. If you prefer to configure sequence sequence databases manually, by editing 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* unless the match is *perfect*. This is because synchronisation of any definition where the match is not perfect requires the database files to be updated.

Initialisation



MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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.

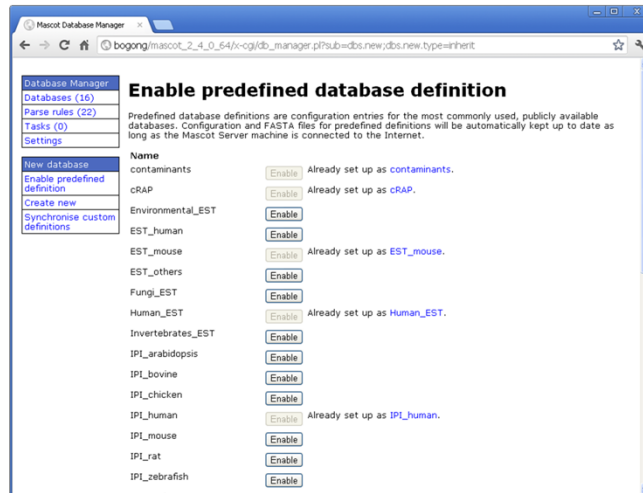
Create New; Use predefined definition template

- 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

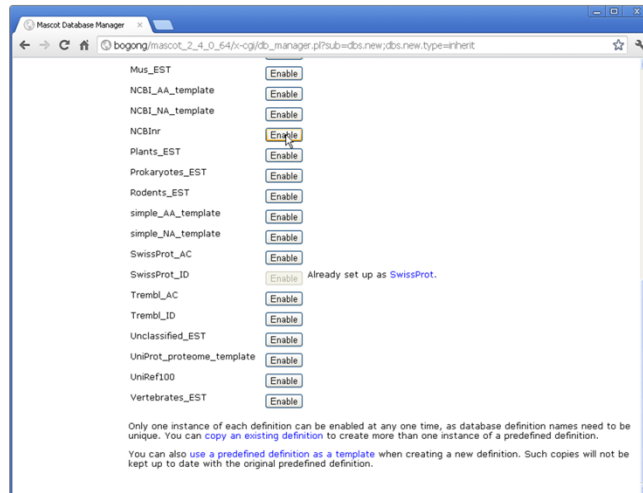


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

Enable Predefined Definition

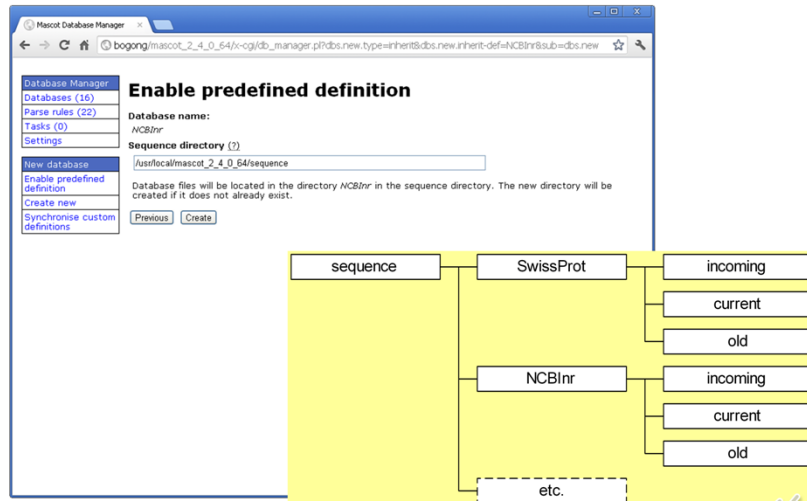


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



Scroll down to NCBIInr and choose Enable

Enable Predefined Definition



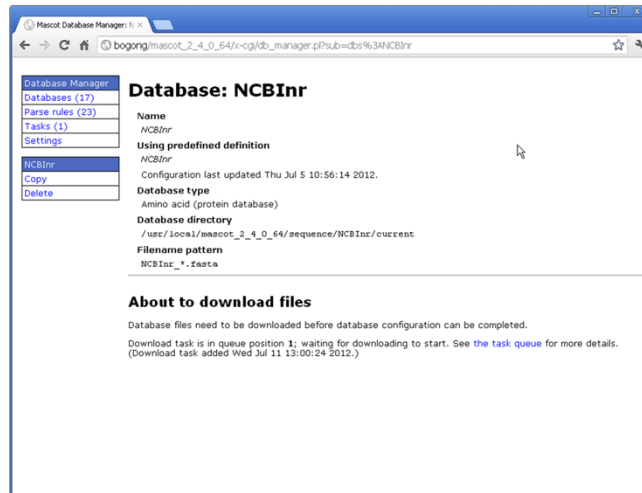
MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



The default location for the local copies of the sequence database files is specified in Database Manager settings. You can also change it 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

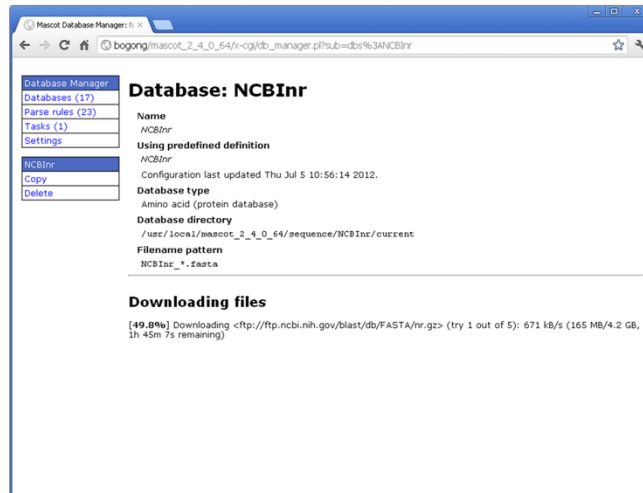


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

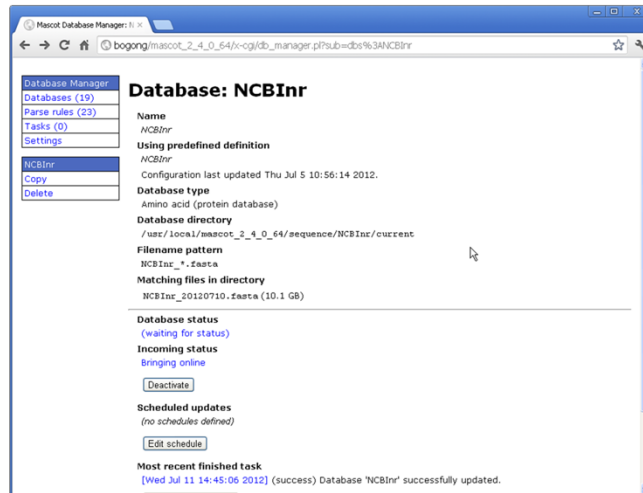


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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 : Sequence Database Administration © 2007-2012 Matrix Science



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

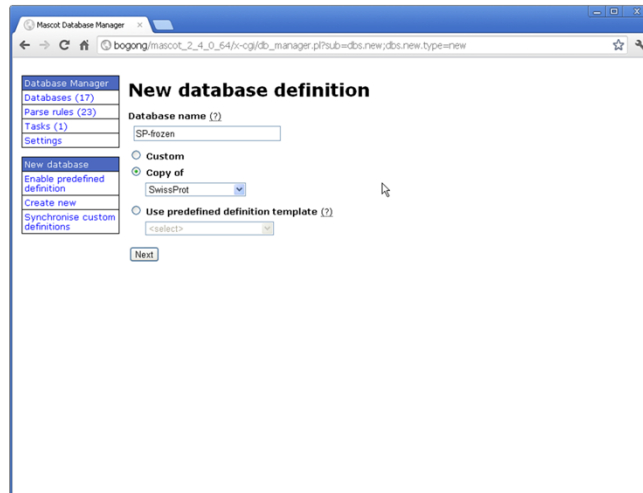
Enable Predefined Definition

The screenshot shows a web browser window with the address bar displaying a URL related to the Mascot Database Manager. The main content area is titled 'Database configuration: NCBIInr'. Under the 'Scheduled updates' section, there are four radio button options: 'None', 'Daily at', 'Weekly on', and 'Monthly on the'. The 'Weekly on' option is selected, and it shows 'Sunday' as the day and '02:00' as the time. Below these options are 'Cancel' and 'Save' buttons. The 'Save' button is highlighted with a mouse cursor.

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



The screenshot shows a web browser window titled 'Mascot Database Manager'. The address bar shows a URL: 'bogong/mascot_2_4_0_64/-cgi/db_manager.pl?sub=db.new;db.new.type=new'. On the left is a sidebar menu with options: 'Database Manager', 'Databases (17)', 'Parse rules (23)', 'Tasks (1)', 'Settings', 'New database', 'Enable predefined definition', 'Create new', 'Synchronise custom definitions'. The main content area is titled 'New database definition'. It contains a 'Database name (?)' text field with 'SP-frozen' entered. Below this are three radio buttons: 'Custom', 'Copy of' (which is selected), and 'Use predefined definition template (?)'. Under 'Copy of' is a dropdown menu showing 'SwissProt'. Under 'Use predefined definition template (?)' is a dropdown menu showing '<select>'. At the bottom of the form is a 'Next' button.

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

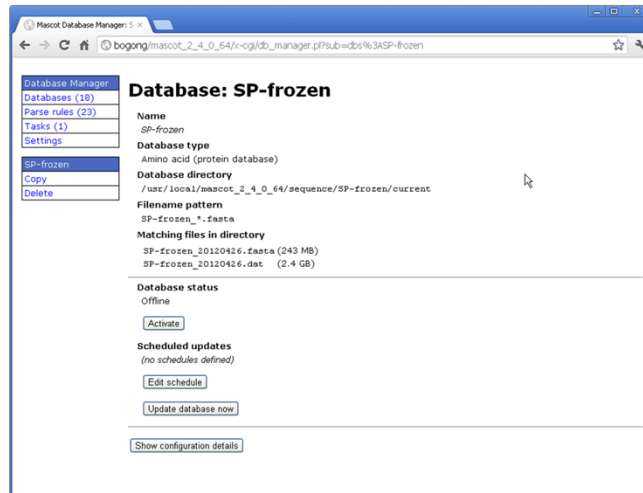
The screenshot shows a web browser window titled 'Mascot Database Manager'. The address bar shows a URL with parameters for creating a new database. The left sidebar contains a menu with options: 'Database Manager', 'Databases (17)', 'Parse rules (23)', 'Tasks (1)', 'Settings', 'New database', 'Enable predefined definition', 'Create new', and 'Synchronise custom definitions'. The main content area is titled 'Copy an existing definition'. It includes a 'Copy of:' dropdown set to 'SwissProt', a 'Database name:' field set to 'SP-frozen', and a 'Sequence directory (?)' text box containing '/usr/local/mascot_2_4_0_64/sequence'. Below this, a note states: '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.' A section titled 'Existing SwissProt files:' lists 'SwissProt_2012_04.fasta (243 MB)' and 'SwissProt_2012_04.dat (2.4 GB)'. There are two radio button options: 'Copy files also (?)' (which is selected) and 'Don't copy files (?)'. A warning message follows: '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.' At the bottom are 'Previous' and 'Create' buttons.

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



You are given the option to copy the existing files.

Create New - Copy of

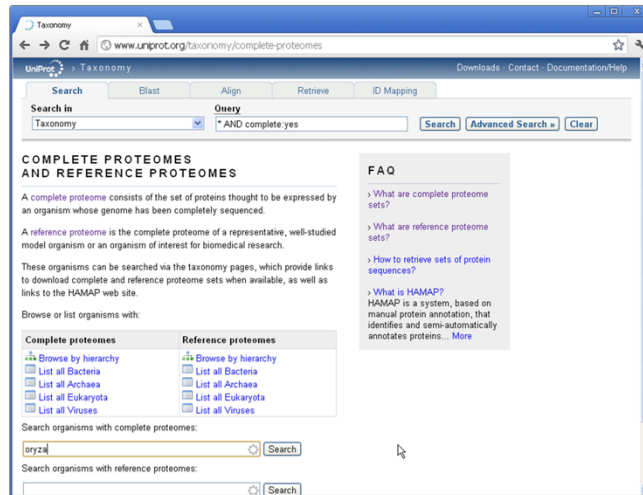


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

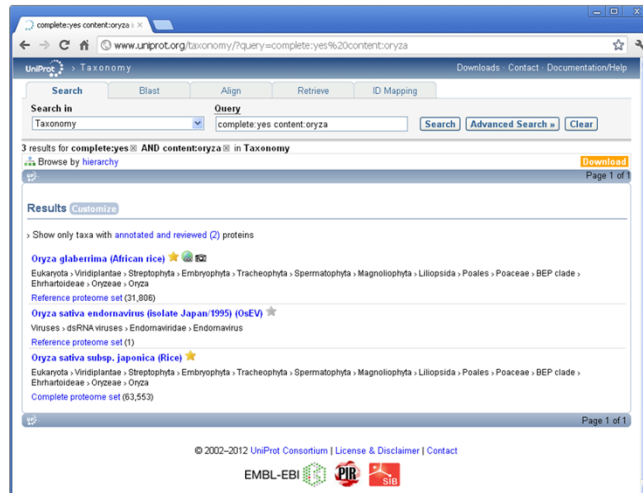


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

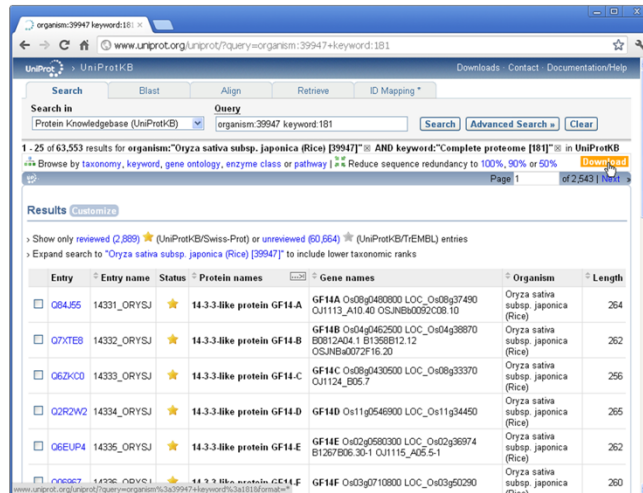


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



There is a choice of two (the middle hit is a virus that infects rice). We'll choose *Oryza sativa* subsp. *japonica*.

Create New - Predefined as Template



MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



You could download the file manually. If so, click on the yellow button and choose Canonical and isoform sequence data in FASTA format.

A better option is to make a note of the URL because this will allow us to configure automatic updating of the database files. The URL displayed in the browser address bar is <http://www.uniprot.org/uniprot/?query=organism:39947+keyword:181>

Create New - Predefined as Template

URL in browser address bar

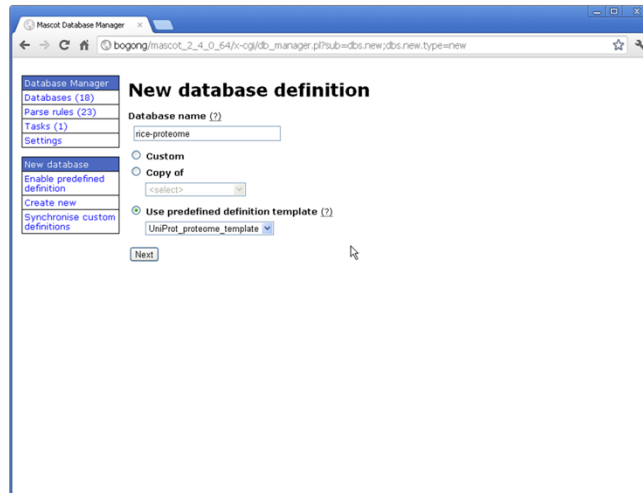
```
http://www.uniprot.org/uniprot/?query=organism:39947+keyword:181
```

URL to force Fasta file download

```
http://www.uniprot.org/uniprot/?query=organism:39947+keyword:181&force=yes&format=fasta&include=yes
```

Looking through the Uniprot help will show that you can convert this to a URL that returns the appropriate Fasta file by adding &force=yes&format=fasta&include=yes to the end

Create New - Predefined as Template

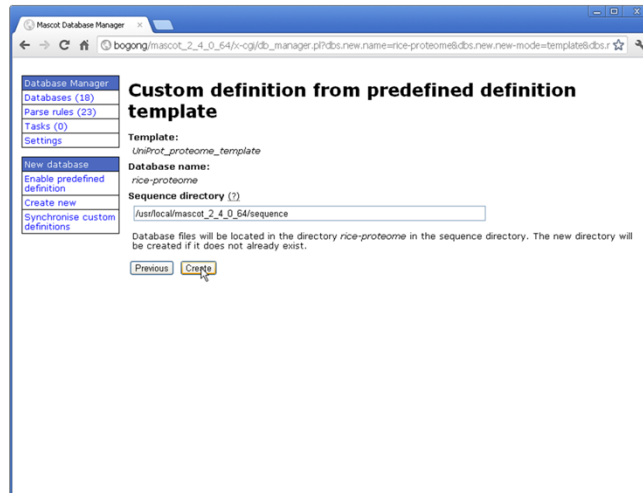


The screenshot shows the 'New database definition' form in the Mascot Database Manager. The form is titled 'New database definition' and has a left sidebar with navigation links: 'Database Manager', 'Databases (18)', 'Parse rules (23)', 'Tasks (1)', 'Settings', 'New database', 'Enable predefined definition', 'Create new', 'Synchronise custom definitions', and 'Settings'. The main form area contains the following fields and options:

- Database name (?)**: A text input field containing 'rice-proteome'.
- Custom**: A radio button that is selected.
- Copy of**: A dropdown menu showing '<select>'.
- Use predefined definition template (?)**: A radio button that is selected.
- UniProt_proteome_template**: A dropdown menu showing 'UniProt_proteome_template'.
- Next**: A button at the bottom of the form.

So, the steps to configure the Uniprot complete proteome of rice are, in Database Manager, Create new. We enter a suitable name and select the uniprot proteome predefined definition as a template.

Create New - Predefined as Template



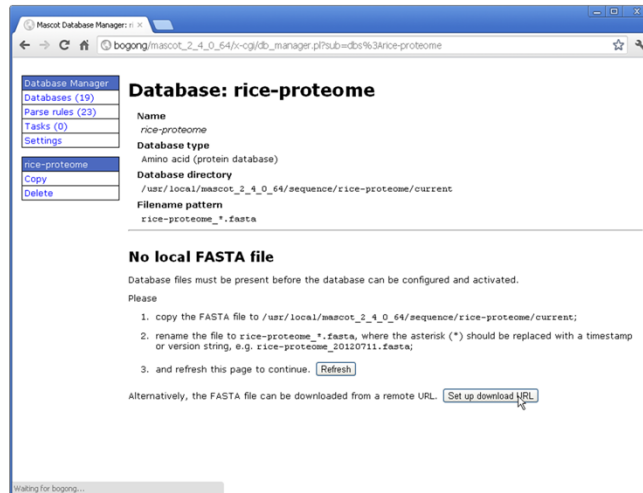
The screenshot shows the Mascot Database Manager web interface. The browser address bar displays the URL: `http://bogong/mascot_2_4_0_64/cgi/db_manager.pl?db.new.name=rice-proteome&db.new.new-mode=template&db.new.new-template=Uniprot_proteome_template`. The left sidebar contains a menu with options: Database Manager, Databases (18), Parse rules (23), Tasks (0), Settings, New database, Enable predefined definition, Create new, and Synchronise custom definitions. The main content area is titled "Custom definition from predefined definition template". It shows the following fields: Template: `Uniprot_proteome_template`, Database name: `rice-proteome`, and Sequence directory: `/usr/local/mascot_2_4_0_64/sequence`. Below these fields, a message states: "Database files will be located in the directory `rice-proteome` in the sequence directory. The new directory will be created if it does not already exist." At the bottom of the form are two buttons: "Previous" and "Create".

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



Set the local directory and choose Create

Create New - Predefined as Template

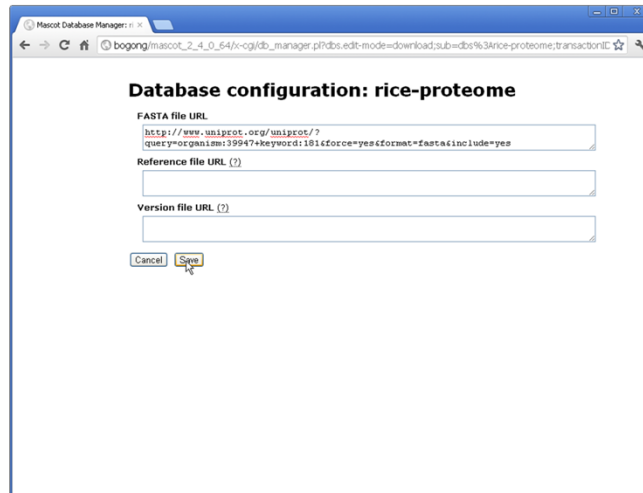


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



If we had chosen to download the Fasta manually, we would follow the instructions to copy the file to the target directory and rename it. Since we want to schedule automatic file updates, we choose instead to set up a download URL

Create New - Predefined as Template



The screenshot shows a web browser window titled 'Mascot Database Manager'. The address bar shows a URL starting with 'bogong/mascot_2_4_0_64/'. The main content area is titled 'Database configuration: rice-proteome'. It contains three text input fields: 'FASTA file URL' (pre-filled with a Uniprot query), 'Reference file URL (?)', and 'Version file URL (?)'. At the bottom of the form are 'Cancel' and 'Save' buttons.

Database configuration: rice-proteome

FASTA file URL
<http://www.uniprot.org/uniprot/?query=organism:39947+keyword:181&force=yes&format=fasta&include=yes>

Reference file URL (?)

Version file URL (?)

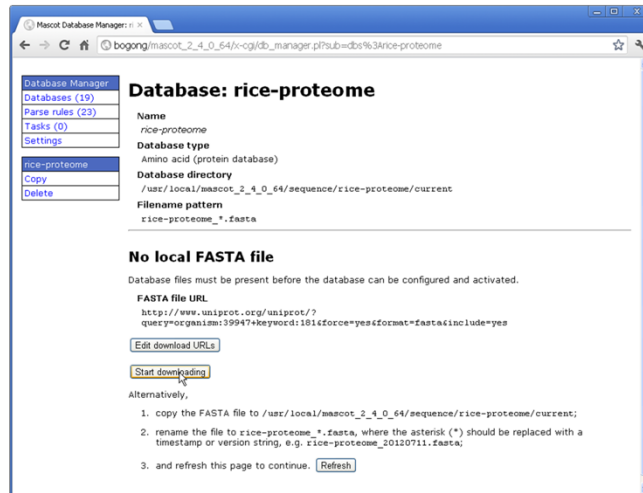
Cancel Save

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



We enter the URL for the Fasta file. There is no reference file to download ... we'll use a hyperlink to Uniprot to get annotation text for the Mascot result reports. There is no version file either, so each update will be identified using an ISO datestamp

Create New - Predefined as Template

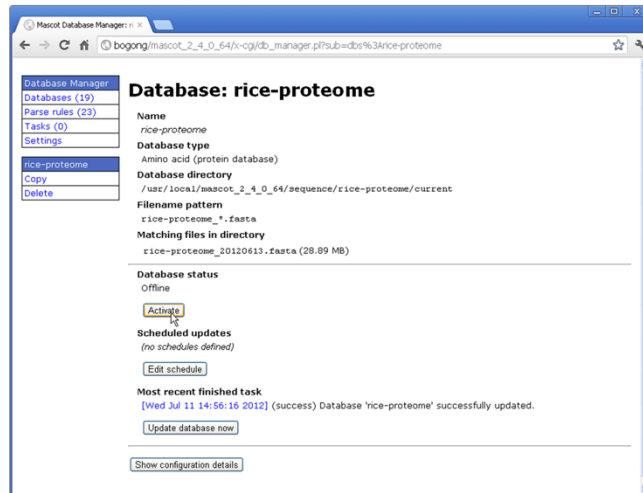


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



Ready to go

Create New - Predefined as Template



MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

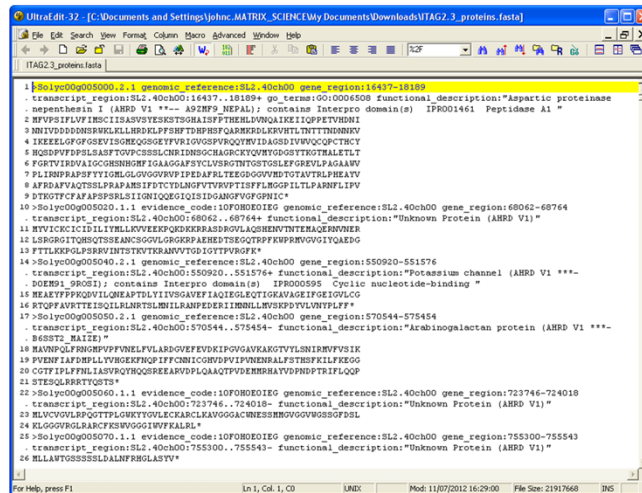


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



But, to illustrate, let's configure the protein Fasta for the recently completed tomato genome

Create New - Custom



```
1 >Sslyc00g050000.2.1 genomic_reference:SL2.40ch00 gene_region:16437-18189
- transcript_region:SL2.40ch00:16437..18189+ go_terms:GO:0006500 functional_description:"Aspartic proteinase
- pepstatin 1 (AHRD V1 ***- AP2IF9_NEPAL); contains Interpro domain(s) IP001461 Peptidase A1 "
2 RTVPSITFLVIEICISASVYTESKTSQALSFTHSLFVQALIEIIGPPVYVQHI
3 MNIVDDSDNSKWLKLLHRDLKLPFSHTDPSHSFQARMEKDLKRVHTLTNTTNNKIV
4 IIEELGFGFQSEVISQHQGGSETFVRLGVGSPVPQQTWIDAGSDIVPVQCPQCTRCY
5 RQSPPTFDELASLTQPCSSSLCHSINDSCCHAGRCXYQVTDGSGTGTALTLT
6 FGRTVIPDVAIGCHSNHMF IGAAGGAFSTCLVSRQNTNOSTOSLEFQREVLPAQAAMV
7 FLINHPAPSPFTYIOHGLGVGVVVP IPEDAFPLTEEGGGVVDTOTAVTLPHAEIV
8 AFDAFPAQTSLPAPAPANSIFQCYVLNGVTVVPTTISFLMGGFILTLPANFLIPV
9 DTRQTFCAFAPSPSRLSIGNIQEQIGISIDGANGVFGGPRIC"
10 >Sslyc00g050020.1.1 evidence_code:10F0R0E0IEG genomic_reference:SL2.40ch00 gene_region:68062-68764
- transcript_region:SL2.40ch00:68062..68764+ functional_description:"Unknown Protein (AHRD V1)"
11 RTVICKICIDILITLRLKVVEKPKQKXKRRASBROVLQSHENVNTNHAQERVRNER
12 LSHGRDITQHSQTSREANCQGVLGKRPAREHETSGQTRFKFPRVVOGVIQAKDG
13 FTLKXLPQSHRVINTKTKTANVVDGLOTTVPVQPS"
14 >Sslyc00g050040.2.1 genomic_reference:SL2.40ch00 gene_region:550920-551576
- transcript_region:SL2.40ch00:550920..551576+ functional_description:"Potassium channel (AHRD V1 ***-
- DQMD1_9R0G1); contains Interpro domain(s) IP000595 Cyclic nucleotide-binding "
15 REATYFPKQVILQEAPELTIIYVSGAFVYIAGIEGLEQIGKAVAGEIFQEIOLVLC
16 RTQFFAVRTTEISQILRLNLSLMLILRANFEDERIIIMNLLWSEPDVLYNPLFF"
17 >Sslyc00g050050.2.1 genomic_reference:SL2.40ch00 gene_region:570544-575454
- transcript_region:SL2.40ch00:570544..575454+ functional_description:"Arabinogalactan protein (AHRD V1 ***-
- B6SST2_MAIZI)"
18 RANPQLFPHQRPVFPWELVLABSGVEFVDEKIPQVAVKACTVTLNIDRVFVSIK
19 PVNFIIFRNLVYGEKTFQPIFFCNICQGVDPVIPNENRALFSTREKILFKKGG
20 COTFIFLFLIASVBOYHQSRREARVPLQAAGTVPVEMMSBAYVDPHDPTRILOQP
21 STEQLRRRTYQTS"
22 >Sslyc00g050060.1.1 evidence_code:10F0R0E0IEG genomic_reference:SL2.40ch00 gene_region:723746-724018
- transcript_region:SL2.40ch00:723746..724018- functional_description:"Unknown Protein (AHRD V1)"
23 RLVCVQVLPQQTITPLQNTYVLEKARCLKAVGGGACWESRHRVGVGVSSGFSLSL
24 ELGGVGLRANCTENVGQGIWFKLSL"
25 >Sslyc00g050070.1.1 evidence_code:10F0R0E0IEG genomic_reference:SL2.40ch00 gene_region:755100-755543
- transcript_region:SL2.40ch00:755100..755543- functional_description:"Unknown Protein (AHRD V1)"
26 MLLAVTQSSSSLDALNPHGLASTV"
```

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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 line. If you don't have a suitable text editor, you can use more at a command prompt.

Create New - Custom

Look at the Fasta file to choose a parse rule

```
>Solyc00g005000.2.1 genomic_reference:SL2.40ch00
gene_region:16437-18189
transcript_region:SL2.40ch00:16437..18189+
go_terms:GO:0006508 functional_description:"Aspartic
proteinase nepenthesin I (AHRD V1 **--
A9ZMF9_NEPAL); contains Interpro domain(s)
IPR001461 Peptidase A1"
```

Could use the very simple rules (simple_AA_template)

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

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



As is often the case, a simple rule that takes everything between the ">" symbol and the first space as the accession will work. Everything after the first space could 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.

To illustrate the custom definition route, let's imagine we want to take the description from the text inside the double quotes. This will require us to create a new parse rule inside Database Manager

Create New - Custom

The screenshot shows a web browser window titled 'Mascot Database Manager'. The address bar shows the URL: `http://bogong/mascot_2_4_0_64/cgi/db_manager.pl?sub=db.new;db.new.type=new`. On the left is a sidebar menu with options: 'Database Manager', 'Databases (19)', 'Parse rules (23)', 'Tasks (0)', 'Settings', 'New database', 'Enable predefined definition', 'Create new', and 'Synchronise custom definitions'. The main content area is titled 'New database definition'. It contains a 'Database name (?)' field with the text 'ITAG23' entered. Below this are two radio button options: 'Custom' (which is selected) and 'Copy of', followed by a '<select>' dropdown. Below these is another radio button option: 'Use predefined definition template (?)', also followed by a '<select>' dropdown. At the bottom of the form is a 'Next' button.

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

Create New - Custom

The screenshot shows the 'Mascot Database Manager' web interface. On the left is a sidebar with navigation links: 'Databases (19)', 'Parse rules (23)', 'Tasks (0)', 'Settings', 'New database', 'Enable predefined definition', 'Create new', and 'Synchronise custom definitions'. The 'Create new' link is highlighted. The main content area is titled 'Custom definition' and contains the following fields and options:

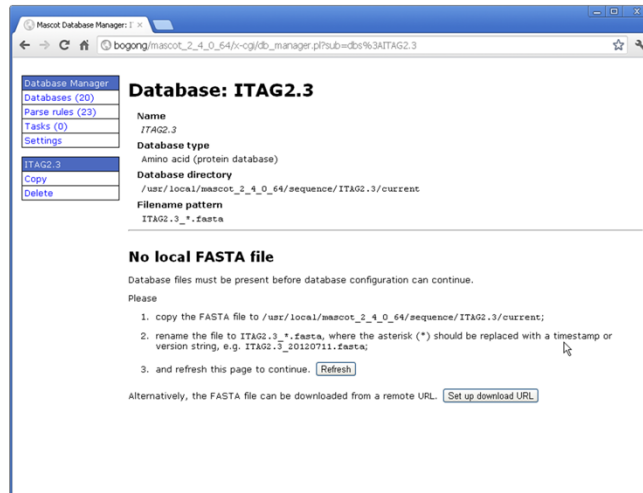
- Database name:** A text field containing 'ITAG2.3'.
- Database type:** Two radio button options: 'Amino acid (protein database)' (which is selected) and 'Nucleic acid (DNA database)'. A note below states: '(Note that database type cannot be changed later.)'
- Sequence directory (?)**: A text field containing '/usr/local/mascot_2.4.0_64/sequence'.
- A paragraph of text: 'Database files will be located in the directory ITAG2.3 in the sequence directory. The new directory will be created if it does not already exist.'
- Download database files (?)**: An unselected radio button option.
- FASTA file URL**: A text input field.
- Version file URL (optional) (?)**: A text input field.
- Reference file URL (optional) (?)**: A text input field.
- I will copy the files to the database directory (?)**: A selected radio button option.
- A paragraph of text: 'If you have chosen to download files from a remote server, the downloads 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.'
- At the bottom are two buttons: 'Previous' and 'Create'.

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

Create New - Custom

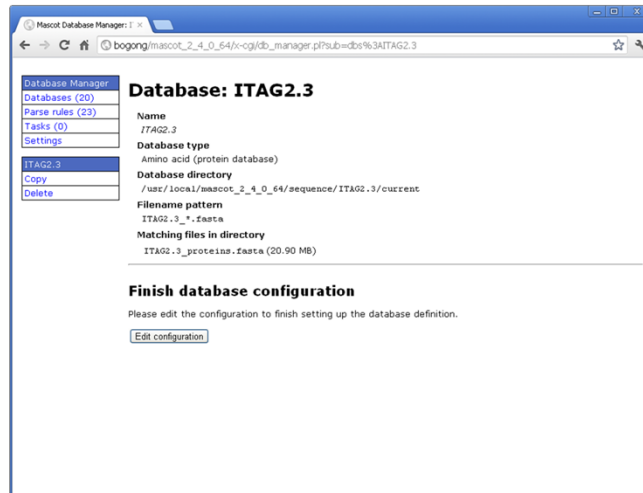


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



We follow the instructions to copy the Fasta file to the target directory and rename it

Create New - Custom

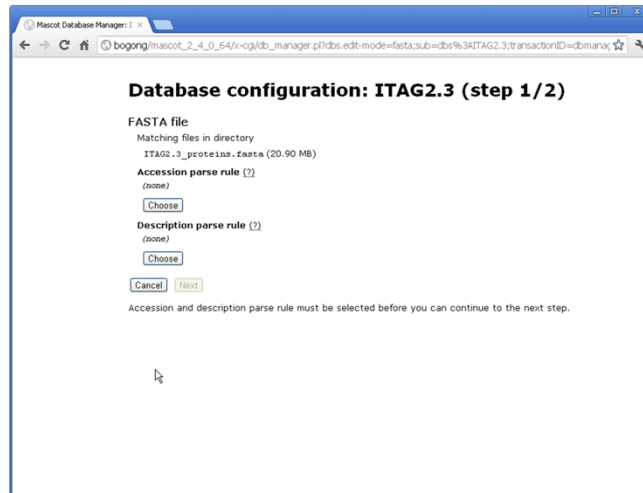


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



Once the file is in place and the name matches the wild card pattern, it will be recognized, and we can proceed to edit the configuration. If Database Manager stays sat on the previous page, 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



Database configuration: ITAG2.3 (step 1/2)

FASTA file
Matching files in directory
ITAG2_3_proteins.fasta (20.90 MB)

Accession parse rule (?)
(none)

Description parse rule (?)
(none)

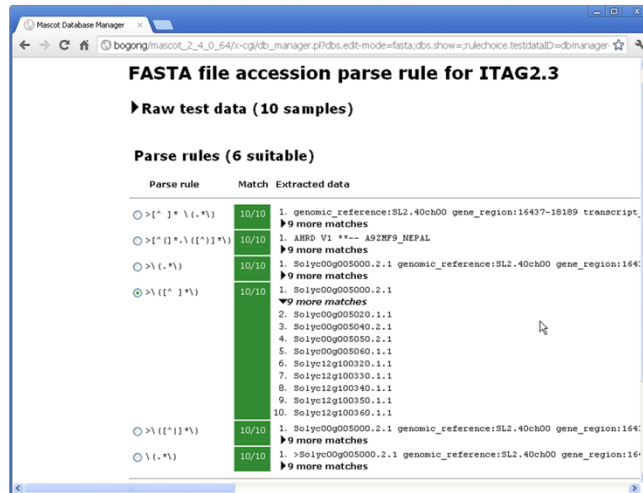
Accession and description parse rule must be selected before you can continue to the next step.

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

Create New - Custom

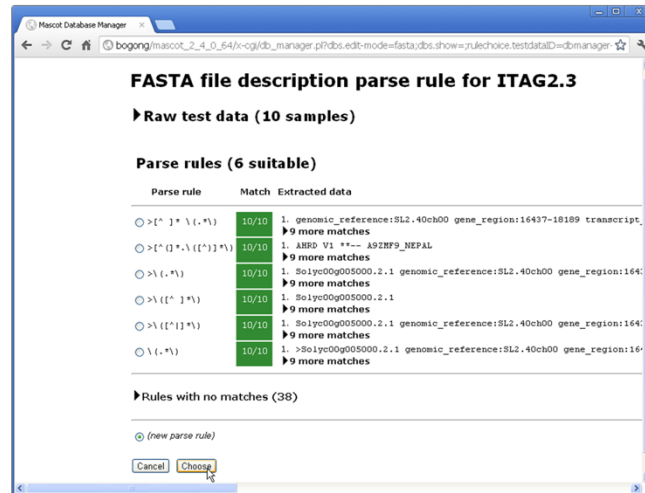


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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. Six 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

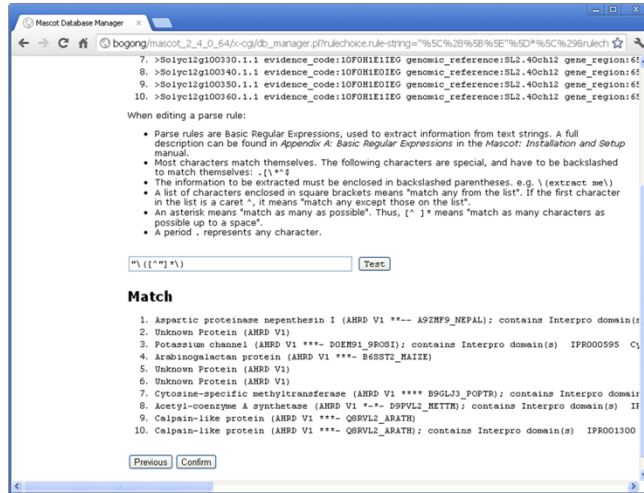


MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



When it comes to the description parse rule, the first one listed would work OK, but as a challenge we want to devise a new rule to extract the text inside the double quotes, so we select 'new parse rule'

Create New - Custom



MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



There is a brief reminder of how basic regular expressions work, and you can use trial and error; pressing test until you get something useful. This screen shot shows a rule that appears to work. It looks for the first instance of double quotes then takes all the characters after that that are not double quotes. We confirm our choices and move on to the remainder of the configuration

Create New - Custom

Database configuration: ITAG2.3 (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_A4_template)

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?rettype=gb&retmode=text&db=nucleotide&tool=mascot&email=support@matrixscience.com&id=#ACI> (example)

☐ Copy from NCBI_A4_template (or NCBIhr)

<http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?>

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



The next step deals with taxonomy and annotations.

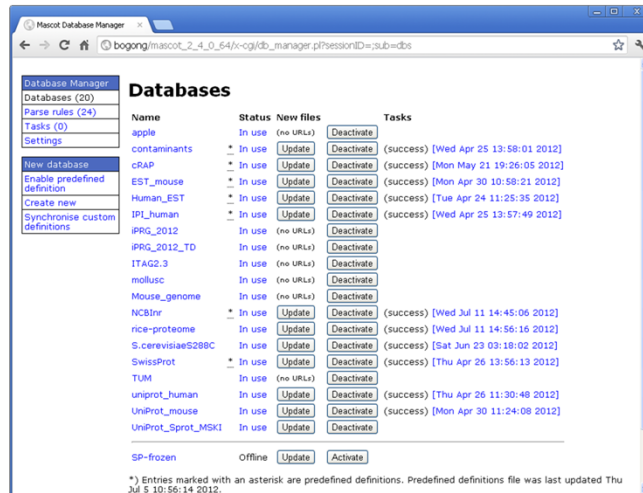
If this was a comprehensive database, containing entries from many different organisms, we might want to select a rule for determining the taxonomy of each entry. There is a drop down list of rules for the major public databases. Most of these require additional files, which can be downloaded automatically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



MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



All being well, a short while later, our new database will show as 'In Use'. You'll notice that, for ITAG2.3, there is no option to update 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

"Old" & "New"

Compression warnings

Unidentified taxonomy

Statistics

Name	Family	Filename	Pathname	Status	State Time	Mem mapped	Request to mem map	Request unmap	Mem locked	Number of threads	Current
NCBIInc	/home/matrix/site/sequence/NCBIInc/current/NCBIInc	NCBIInc_20041113.fasta	/home/matrix/site/sequence/NCBIInc/current/NCBIInc	Not in use	Sun Nov 21 04:39:43	NO	YES	NO	NO	1	NO
NCBIInc	/home/matrix/site/sequence/NCBIInc/current/NCBIInc	NCBIInc_20041117.fasta	/home/matrix/site/sequence/NCBIInc/current/NCBIInc	In use	Sun Nov 21 04:39:46	YES	YES	NO	YES	1	YES
EST_human	/home/matrix/site/sequence/EST_human/current/EST_	EST_human_20041113.fasta	/home/matrix/site/sequence/EST_human/current/EST_	Not in use	Sun Nov 21 09:24:39	NO	YES	NO	NO	1	NO
EST_human	/home/matrix/site/sequence/EST_human/current/EST_	EST_human_20041117.fasta	/home/matrix/site/sequence/EST_human/current/EST_	In use	Sun Nov 21 09:24:40	YES	YES	NO	NO	1	YES
EST_mouse	/home/matrix/site/sequence/EST_mouse/current/EST_	EST_mouse_20041113.fasta	/home/matrix/site/sequence/EST_mouse/current/EST_	Not in use	Sun Nov 21 10:49:52	NO	YES	NO	NO	1	NO
EST_mouse	/home/matrix/site/sequence/EST_mouse/current/EST_	EST_mouse_20041117.fasta	/home/matrix/site/sequence/EST_mouse/current/EST_	In use	Sun Nov 21 10:49:54	YES	YES	NO	NO	1	YES

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science

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 it will be necessary to follow links to the error log or compression warning log to see what has gone wrong.

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%?

Database statistics - Microsoft Internet Explorer

Time files compressed : Sun Nov 21 04:22:43 2004
 Time files compressed (int) : 1101010963
 Time / date of fasta file : Wed Nov 17 08:27:00 2004
 Time of fasta files (int) : 1100680020
 Number of residues : 736969646
 Number of sequences : 2171599
 Number with invalid residues: 0
 Number of sequences too long: 0
 Length of longest sequence : 37777
 Version of Mascot : 2.0.03
 Version of this file : 2
 Maximum accession length : 20
 Seqs with invalid taxon tree: 14
 Num sequences for taxonomy : All entries=2168522
 Num sequences for taxonomy : Archaea (ArchaeaBacteria)=62823
 Num sequences for taxonomy : Eukaryota (eucaryotes)=1017421
 Num sequences for taxonomy : Alveolata (alveolates)=30561
 Num sequences for taxonomy : Plasmodium falciparum (malaria parasite)=9177
 Num sequences for taxonomy : Other Alveolata=21392
 Num sequences for taxonomy : Metazoa (Animals)=642609

Database statistics - Microsoft Internet Explorer

Tax ID	Count
0	3416
1	1177306
2	631043
3	196425
4	78910
5	30514
6	19451
7	11704
8	5020
9	5003
10	2745
11	2010
12	1295
13	1208

Database statistics - Microsoft Internet Explorer

Length	Count
6	692
7	828
8	857
9	976
10	1195
11	1016
12	1076
13	1208

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science **MATRIX SCIENCE**

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 0 taxonomy identifiers represents less than 1% of the database

Configuration - Performance

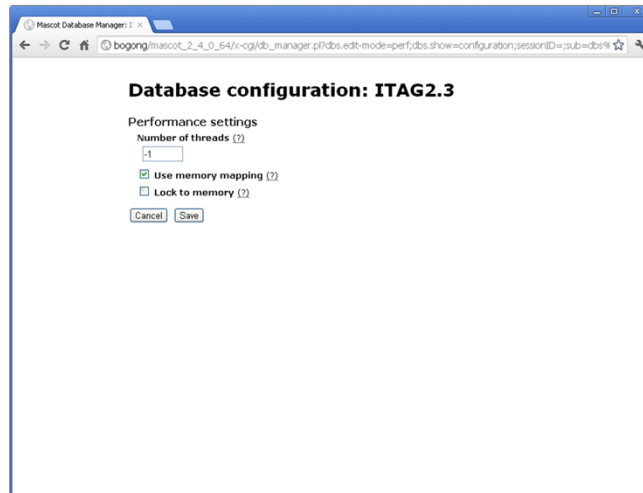
[illegible]

MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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

Configuration - Performance



MASCOT : Sequence Database Administration © 2007-2012 Matrix Science



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* © 2007-2012 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.