



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

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

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

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



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 The help pages in your in-house copy of Mascot are similar, but become progressively out-of-date.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 ORE's mRNA etc. 	
• PSI Extended FASTA Format used by the neXtProt database	
 Must have local file (Mascot streams through the database during each search) 	
Other files are optional • Taxonomy indexes • Full text annotations.	
ASCOT : Sequence Database Administration © 2007-2022 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.

Spectral Libraries Requirements	
 Mascot can also search Spectral libraries Libraries of peptide spectra Can include spectra of peptides with modification And missed or non-specific cleavages 	
 NIST MSP format ISB SpectraST format X! Hunter Annotated Spectrum Library (ASL) Automatically converted to MSP format prior to searching 	
 Must have local files (Mascot streams through the database during each search) Have to assign a Protein reference database Required for protein inference 	
ASCOT : Sequence Database Administration © 2007-2022 Matrix Science	MATRIX

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

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

There are three supported Spectral library formats:

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

The second one is Institute for Systems Biology SpectraST format.

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

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

When configuring a spectral library you need to specify a Protein reference database that is used for protein inference. Swissprot is commonly used along with a taxonomy filter but individual proteome databases can also be used. Spectral libraries can be download from the internet or created from Mascot search results. In this talk we will only cover spectral libraries downloaded from the internet. Libraries created from search results are covered in the Spectral Library talk.

FAST	A Format	
	<pre>>Title text SEQUENCESEQUENCESEQUENCESEQUENCESEQUENCE SEQUENCESEQUENCESEQUENCESEQUENCESEQUENCESEQUENCE SEQUENCESEQUENCESEQUENCE >Next title NEXTSEQUENCE</pre>	
	<pre>>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 </pre>	
MASCOT	: Sequence Database Administration © 2007-2022 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	
<pre>>sp P14384 CBPM_HUMAN Carboxypeptidase M OS=Homo sapiens OX=9606 GN=CPM PE=1 SV=2</pre>	
Accession from Fasta title: "> \([^]*\)"	
Description from Fasta title: ">[^]* \(.*\)"	
MASCOT : Sequence Database Administration © 2007-2022 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 UniProt Proteome human database. Let's say that we want to use P14384 as the unique accession string and everything after the first space should be treated as the description.

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

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

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



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 Concept	:S
Predefined Database Def	finition
 Configuration infor Science web site, a 	rmation for the most popular public databases is kept up-to-date on the Matrix and downloaded as required by Database Manager
Custom Database Definit	ion
 If you want to sear you want to config Definition 	ch a database that is not included in the list of Predefined Database Definitions, or if ure one of these databases in some non-standard way, you create a Custom Database
Synchronisation	
 If a custom definit definition by being 	ion is very similar to a predefined definition, it can be converted into a predefined g synchronised
Update Schedule	
• A schedule can be	created to update all the files associated with a database automatically.
SCOT : Sequenc	CE Database Administration © 2007-2022 Matrix Science

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.

isation	`				
Isatio	•				
← → C @	ngine Protein × Mescot Database Manager Cen × +	manager.pl?sub=newdef		0 \$	- ⊔ ×
Mascot Da	tabase Manager Configur	ation Import			î
Note: Please disable other	programs and scripts that configure or update sequen	ce databases before continuin	g.		
Database Manager ca files.	n keep your sequence database configuration up to dat	e and schedule the download	of new		
If you choose to use I lost. If you do not wis	Database Manager, you should not edit mascot.dat in f h to use Database Manager now, please return to the (uture, because your changes Configuration Editor.	will be		
Database Manager ha	s two kinds of database definitions:				
 Predefined data Custom databas 	base definition: Automatic configuration by downloading ie definition: Manually maintained configuration.	up-to-date information from	the Matrix Science website.		
You can choose to ke (Synchronise). Synchr (FASTA, reference, ta	ap the existing definition (Keep as Custom) or synchron onisation means the predefined database definition will konomy) will be updated.	ise with a predefined definition be applied, and the database	files		
If you do not want to by choosing Synchron	update the files now, choose Keep as Custom. You car se Custom Definitions from the Database menu.	synchronise at any time in t	ne future		
Database definition NIST_Human_HCD	Action Keep as Custom Synchronise with NIST_Human_HCD Synchronise with NIST_Human_HCD_TRAQ_1 Synchronise with NIST_Huma_HCD_TRAQ_2 Synchronise with NIST_Huma_HCD_TRAQ_3 Synchronise with NIST_Huma_HCD_Semitryp	Match quality perfect good Differences good Differences good Differences good Differences			
PRIDE_Contaminant	 Keep as Custom Synchronise with PRIDE_Contaminants 	perfect			
SwissProt	 Keep as Custom Synchronise with SwissProt_ID Synchronise with SwissProt_AC 	perfect poor Differences			
- contaminants	O Keen as Oistom				~,

Database Manager must be allowed <u>exclusive control</u> 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.



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.



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.

Fnabl	e Predefined Definition
	L Macca Datasa Mangar × + - □ ×
	C 🛆 O la boahost 8097/mascot 2,6,02/e-og/db_manager p?hub=dbu.new.dbf.new.dbf.dbs.new.dbf.dbss.new.dbf.db
	Subject Apple
MASCOT	: Sequence Database Administration © 2007-2022 Matrix Science

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

Scroll down to neXtProt and click Enable

Enabl	e Predefined Definition Uncertainty to * * * * * * * * * * * * * * * * * *
	Exable predefined definition Database (n) Database name: Northold updates Database name: North
	Under your of the sequence SwissProt incoming Sequence SwissProt old
	etc.
MASCOT	: Sequence Database Administration © 2007-2022 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



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

Enabl	e Predefined Definition ✓
	Paramate (10) Paramate (10) <td< th=""></td<>
MASCOT	: Sequence Database Administration © 2007-2022 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
5 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	absolution Andrew State (100) bar Material Own Control (100) bar Material
MASCOT	Sequence Database Administration © 2007-2022 Matrix Science 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 from here or via the Scheduled Updates side menu



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.



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



You are given the option to copy the existing files.

Create	New - Copy of
Create	<image/>
MASCOT	: Sequence Database Administration © 2007-2022 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		Dradafia		Tomolata		
Create	New -	Predefin	ed as	remplate		_
	2 * in proteomes	× +			- = >	
	< → C @	A https://www.uniprot.org/proteor	nes/			<u>^</u>
	UniProt	Proteomes -			Advanced - Q Search	
	BLAST Align Retrieve/ID ma	oping Peptide search SPAROL			Helo Contact	1
	Proteomes 2	021 04 results				1
	What are proteomes?			What are reference proteomes?	×	
	A proteome is the set of pro proteomes for species with	teins thought to be expressed by an organism completely sequenced genomes.	. UniProt provides	Some proteomes have been (manually and algorithm cover well-studied model organisms and other organ phylogeny. O Help	nically) selected as reference proteomes . They nisms of interest for biomedical research and	
				Proteomes help vide	o Other tutorials and videos Downloads	
	Filter by	🛓 Download 🛛 🗶 Columns			I to 25 of 381,828 ▶ Show 25 ∨	
	20,382 Reference proteomes	Show only non-redundant proteomes? Show only not-excluded proteomes?			×	
	361,446 Other proteomes	Destaams ID Oeaspies	Orazoiem ID. Brotein o	BUSCO	CPD Consensus representation (RefCon)	
	270,855 Bacteria		1403320 4947	Single Duplicated Fragmented Missing ©	Close to full	
	104,078 Viruses	Alteromonadaceae bacterium 2753L.S.0a.02		n:820 C:92.7% (S:91.8% D:0.9%) F:0.4% M:7%	standard (high	
	3,589 Archaea	UP000295807 R Anseongella	496056 3575	alteromonadales_odb10	Close to full	
	3,306 Eukaryota	ginsenosidimutans (Stra DSM 21100)	n:	n:1068 C:86.5% (S:86.1% D:0.4%) F:0.3% M:13.2%	(low value)	
	UniProtKB UniParc (for redundant proteomes)	UP000583740 R Cardinalis cardinalis (Northern cardinal)	98964 13821	n:10844 C:73.3% (5:73.1% D:0.2%) F:5.9% M:20.9% passerformes_odb10	Standard full	
	Demo	UP000503504 Regomovirus-	1911908 1		Standard full	
	Help video	associated alphasatellite sp.	271257 2		Chandrard full	
		E 09000203973 K	2/123/ 2		Standard fui	*
MASCOT : S	Sequence D	atabase Admir	nistration	© 2007-2022 Matrix Science		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

reate New	- Predefir	ned	as Temp	late		
O oryza is proteomes	× +		I		- 0	×
 ↔ c a 	O A https://www.uniprot.org/proteom	es/?query=Oryza&si	rt=score		☆ ± >>	±*
UniProt	Proteomes - onza	1225			Advanced - Q Search	Ê
	maneira Bantida casesh SBAROJ				Page Hole Control	
Proteomes	2021 04 results				nep contac	
What are proteomes	?		What are reference proteon	nes?		×
A proteome is the set of proteomes for species wi	proteins thought to be expressed by an organism. th completely sequenced genomes. 😗 Help	UniProt provides	Some proteomes have been (manually cover well-studied model organisms a phylogeny.	y and algorithmically) select and other organisms of intere	ed as reference proteomes. They ist for biomedical research and	
			Protec	omes help video 🛛 Other I	autorials and videos 🕹 Downloads	
Filter by	🛓 Download 🛛 🗶 Columns				1 to 14 of 14 Show 25	×
R 11 Reference proteome	Repeat search in UniProtKB (850,323)				3	¢
3 Other proteomes Superkingdom	Proteome ID Organism	Organism ID Pro	tein count 🗣 📕 BUSCO	d Missing 🗘 🗘	Genome representation (RefSeq)	e.
1 Viruses 13 Eukaryota	UP000059680 Doyza sativa subsp. japonica (Rice) (Strain: cv. Ninoonbare)	39947 48	n:4896 C:84.4% (S:78.1% D:6.3%) F:5% I	Outlier (high M:10.6% value)	full	
Map to UniProtes	UP000007085 Oryza sativa endornavirus (isolate	654933 1	pomes_odb10	Standard	full	
Demo	Japan/1995) (OsEV) UP000006591 Oryza nivara (Indian wild rice) (Oryza sativa f.	4536 46	176 n:4896 C:95.8% (S:69.2% D:26.6%) F:0.9	Outlier (high value)	full	
	uP000479710 Oryza meyeriana var. granulata (Strain: cv. Menghai)	110450 44	poales_odb10 256 n:4896 C:87.2% (S:68.9% D:18.2%) F:3.5	Outlier (high value)	full	
	UP000007013 🔃 Oryza sativa subsp.	39946 37	poales_odb10 183 n:4896	Outlier (high	full	v
T · Sequenc	o Databaso Adm	inistra				
Sequenc	e Database Adm	mistra	CONTRACTOR CONTRA	ix Science		

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

	_					
Create Ne	ew - Prec	lefined	as Templ	ate		
Ocra strius schen jane	wka/E-X +				- n x	
	Q A https://www.uniprot.org/org	teomes/UP000059680		日 80% ☆		
					R.	^
BLAST Align Retrieve/ID map	pping Peptide search SPARQL	aioa (Rico)		The second section in the second s	Help Contact	1
Froteomes - C						
2 Overview	None Overview					
Components	Proteins ¹ 48,900					
 Publications 	Gene count ¹ 43,672 - Download one pr	otein sequence per gene (FASTA)				
Map to	Taxonomy 39947 - Oryza sativa subs	p. japonica				
UniProtKB (48.900)	Strain cv. Nipponbare					
Reviewed (4.110)	Last modified May 24, 2021 Decome essembly 0C4 0014333335.1 from B	ALTEMPS Full				
Unreviewed (44,790)	and annotation ¹					
54MBL	Busto ¹ C:84.4%[S:78.1%,D:6.3%],F:5%,M:10.6%,n:4896 poales_odb10				
	Completeness* Outlier (high value)					
	Oryza sativa (rice) is a monocotyledonous flowe	ring plant of the family Poaceae and is one of i	the most important crop plants in the world, providing the	e principal food source for half of the	world's population.	
	when cooked, which distinguishes it from subsp.	indica which is long grained and not sticky.	ning president corporations and appropriate in an or granter		a granta acce together	
	Oryza sativa subsp. japonica is grown in dry fiel	ds. mainly in temperate or colder dimates such	h as Japan.			
	Oryza sativa has a haploid chromosome numbe smallest of the major cereal crop genomes and	r of 12, containing 370 Mb with about 36,000 p is the easiest to transform genetically. The cult	protein-coding genes. Rice was an obvious choice for the tivar sequenced from the japonica subspecies was Nippor	tirst whole genome sequencing of a abare (mm.thericejournal.com/conter	ereal crop. It is the 4/6/1/4/abstract).	
	Components'					
	& Download View all proteins					
	×	Genome Accession(s)	Component representation		Proteins	
	selected components	AP014957			6678	
	Download protein entries from all 14 components	AP014958			5345	
	Formati (FASTA (canonical)	AP014959			5766	
	Compressed Uncompressed					
	Preview first 10 * Go	AP014960			4387	
	Chromosome 5	AP014961			3915	
MASCOT : Sec	quence Databas	e Administra	tion © 2007-2022 Matrix	Science		MATRIX SCIENCE

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

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



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



Choose Create



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



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



Save, and we're ready to start downloading

Creat	<section-header><complex-block></complex-block></section-header>	
MASCOT	: Sequence Database Administration © 2007-2022 Matrix Science	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.



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



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



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



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.



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

Creations Control of the second seco	<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	- C × 4 » #	
MASCOT	: Sequence Database Administration © 2007-2022 M	latrix Science	MATRIX SCIENCE

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

Create New - Custom	
📙 Marad Balana Manya Ito X 📕 Manad adalasa asartu Seya X 📕 Manad Balana Hanya X + - 🗆 🗡	
Subdates (i) Subdate Subdates Subdates	
MASCOT : Sequence Database Administration © 2007-2022 Matrix Science	MATRIX SCIENCE

Set these options and choose Create



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



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.



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

_		
Create New	/ - Custom	
Mascot Data	asse Manager NC × La Mascat database search: Seq × La Mascat Database Manager × + ×	
\leftarrow \rightarrow C a	🛇 👌 mascotserver26/mascot/s-cgi/db_manager.pl?dbs.edit-mode=fastazulechoice.testdatalD=dbmanager-test-data_3051678245536098:: 🏠 🙁 😫	^
	Accession parse rule for ketchup	
	Raw test data (10 samples)	
	1. bólyvölyölyölyöl (1. 11. Komeshon leusia-ripper protein (AMBD '73.11 +* tr(20094)[20196]_IUND) 2. bólyvölyölyölyöl, il Bukonsynäh =<-boxtkyintarfartar-tak-leusiaten (ROST) (1. 400 '73.11 +* tr(20095]_ARATE) 3. bólyvölyölyölyöl, il Bukonsyn protein 4. bólyvölyölyölyöl, il Bukonsyn protein 5. bólyvilgiolyöl, il reintarfartar-SIME enthyliranaferas-IL enthyliranafe	
	Suitable parse rules (6)	
	Parse rule Match Extracted data	
	<pre>O >\(.*\) 1010 1. 5.01y000g1603020.1. Homesbox leucine-zipper protein (AIRD V3.11 +-* tr(QRES43[QRES43_IIRVI)</pre>	
	○>\([^,]*\) 10/10 1. Solpeop(6226).11 9 more matches	
	(*) 14/15 14/15 9	
L	O >\((*) *\) 10/10 1. Soly=00pi60246.1.1 Homeobox leucine-zipper protein (ABRO V3.11 *-* tr bh mean-matcher	v.
MASCOT : Seque	nce Database Administration © 2007-2022 Matrix Science	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. Four parse rules give matches to all ten. We need to study the matches and choose the one that pulls out a suitable accession string. I think its fairly obvious that the one we've selected is suitable



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'



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

Create New	- Custom
← → ⊂ @	🔿 🛔 mascotserve26/manager.p?/dbs.addi-mode+annotations.sub+dbs%3AtetdruptransacionID+dbmanager,?1556483967 🏠 🕹 🔅 📫
	Database configuration: ketchup (step 2/2)
	Taxonomy source
	© None (?)
	O Pasta Ref ()
	Sequence report source
	FASTA Rie
	Full-text report source
	(e) None (?)
	O External source ()
	© Current setting
	A
	Parte Inte
	COUNNEL Copy from Environmental_EST (and similar entries) http://www.ebi.ac.ukuTcols/ddetan/emblfecch1g=HACCESICOM
	Copy from TrembLAC (and similar entries) http://www.unitgrot.erg/Misprot/#ACCESSION.tax (samanda)
	Copy from EST_human (or EST_mouse or EST_others) https://woilik.nih.gov/entrel/exils/fetch.fogiTrettype+gbiretmode=testidb=mucleolideitool=maarooisemail=support@matrixscience.com (example) (example)
<u>ـ</u> ــــــــــ	Copy from MCDDr (or MCDDred) https://weilik.nbb.nbh.nbh.mbh.gov/entres/eutla/efecth.fogl?rettype=gpcretmode=text4db=groteintcol=masecotemal=mupport@matrixecience.com 1====================================
MASCOT : Sequen	ce Database Administration © 2007-2022 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 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'.

Creat	te New - Custom	
	📕 Marcet Ditabase Manager I X 📕 Marcet ditabase search Sey X 👢 Marcet Ditabase Manager I X 📕 Marcet Ditabase Manager X 📕 Marcet Ditabase Manager X + X	
	← → C △ O △ maccotaever26/mascot/v.cg/ido_manager.pl? ☆ ☆ ☆	
	Nucleus Databases (a) Description Synta sink (12) Name Note (?) Type (?) Salas Synta sink (12) Name Note (?) Type (?) Salas Lates task Synta sink (12) OLA Note (?) Type (?) Salas Lates task Synta sink (12) OLA In task Centern files Description Synta sink (12) Note (?) Type (?) Salas Lates task Contraments In task Centern files Description Description Extrage Note (?) Note (?) Description Description Description PMC_contramenters PMC_contramenters No task Centern files Description Update succeeded (view log)	
	Cardial propried/metry SP-Total (setuple) Cardial (setuple) All (watery) Cardial (setuple) Description (setuple) Description (setuple) <thdescription (setuple)<="" th=""> <thdescription (set<="" th=""><th></th></thdescription></thdescription>	
	Cincide production database_202111127083150.uml) and Fin Nev 12 01:31:50 2021 (spectral Maranes: Maranes: Maranes: 202111127083150.uml). Cincide production Gatabase status is available on the database status page. Cincide production Filterian Spectral Marane Filterian	
	L A A A A A A A A A A A A A A A A A A A	
MASCOT	: Sequence Database Administration © 2007-2022 Matrix Science	MATRIX SCIENCE

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 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



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%

Configu	ration - Performance		
•	🖥 Marcel Database Manager 1: X 📄 Marcel database search Sec. X 📜 Marcel Database Manager 1: X 📄 Marcel tearch status page 🗴 👘 Ma	scot Database Manager × + ×	
	← → C @ O A mascotserver26/mascot/x-cgi/db_managerpl?sub=dbs%3Aketchup&idbs.show=configuration	☆ 또 » ≝	
	Use reference file?		
	Edit file settings	_	
	Version file		
	version mercae or parti (none) Edit file sations		
	(managed and the second	_	
	Taxonomy source		
	(Some) Sequence report source		
	Full-text report source None		
	Edit annotation settings		
	Performance settings	_	
	Number of threads automatic		
	Use memory mapping?		
	Lock to memory?		
	con heuromonice security	-	
	[Boox to database summary]		v,
MASCOT :	Sequence Database Administration © 2007-2022 Matrix	(Science	A MATRIX SCIENCE

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



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.



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



Save the directory location as per normal



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



And Mascot Server will download and uncompress the spectral library then activate it.

Adding a new Spectral Library definition and creating a Spectral Library from search results is covered in the Spectral Library talk.

Datal	pase 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 © 2007-2022 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.