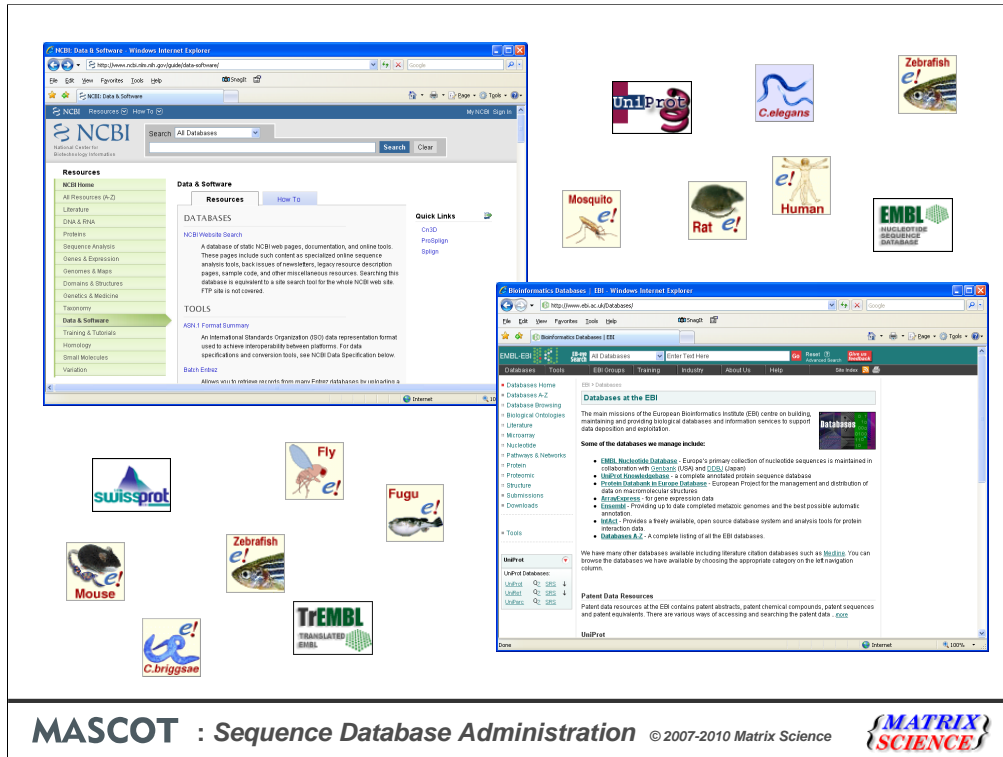


# Sequence Database Administration

MASCOT

*{MATRIX}*  
*{SCIENCE}*



When you install Mascot, it includes a copy of the Swiss-Prot protein database. However, it is almost certain that you and your colleagues will want to search other databases as well. There are very many to choose from, and Mascot allows you to have as many databases as you wish on-line for searching at any one time. (In Mascot 2.2 and earlier, there was a limit of 64 active databases)

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

## Sequence Database Requirements

### Mascot can search any database available in Fasta format

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

### Must have local Fasta file

- (Mascot streams through the database during each search)

### Other files are optional

- Taxonomy indexes
- Full text annotations.

**MASCOT** : Sequence Database Administration © 2007-2010 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 translate the sequences. Mascot performs a 6 frame translation during each search. Nucleic acid databases come in several flavours. The contents 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 search by taxonomy, this will require some additional files, which vary from database to database

Some databases, such as Swiss-Prot, 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* © 2007-2010 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* © 2007-2010 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.

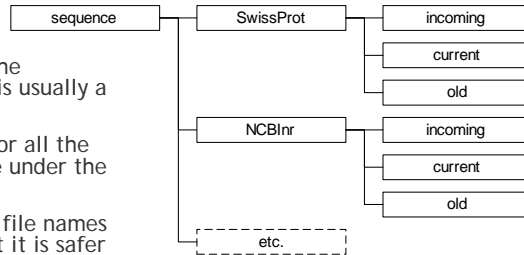
## Adding a New Database

### 1. Choose a name for the Database

- Short, descriptive, *case sensitive*

### 2. Create a local directory structure

- Giving the database and the directory the same name is usually a good idea
- There is no requirement for all the database directories to be under the sequence directory
- Under Windows, path and file names are not case sensitive, but it is safer to treat them as if they were
- Mascot does not support Windows UNC paths.
- Under Unix, links provide great flexibility



**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



We'll summarise the general procedure for adding a new database, then illustrate this with a couple of examples.

The first step is to choose a name for the database. This is the name that will appear in the drop-down list in the search form, so we don't want to write an essay. Note that database names are case sensitive.

The second step is to create a local directory structure on the Mascot server. The recommended arrangement is to have a dedicated directory for each database. Within 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.

Giving the database and the directory the same name is usually a good idea, but is not a requirement. Also, there is no requirement for all the database directories to be placed in the mascot/sequence directory. You can place the files on whichever local drive has most space.

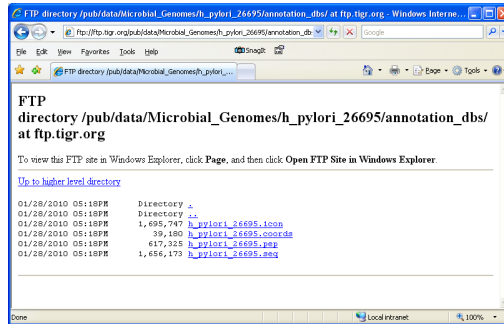
Under Windows, path and file names are not case sensitive, but it is safer to treat them as if they were. Note that Mascot does not support Windows UNC paths

Under Unix, links provide great flexibility. If the Fasta file is actually a link, then Mascot will create the compressed files in the directory containing the link, not in the target directory containing the Fasta file. If you want the compressed files to be on a remote drive, you can do this by making a link at the directory level. However, ensure that the network bandwidth is sufficient, and that the operating system supports memory mapping of NFS mounted files.

## Adding a New Database

### 3. Download the files

- Fasta file is required.
- A "reference" file, containing annotation text and cross-reference information, is optional
- Taxonomy index files may be required



**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



Step 3 is to download at least one release of the database manually, so as to verify the filenames and URLs.

Some databases come with a "reference" file, containing annotation text and cross-reference information in addition to the sequence. An example would be the Swiss-Prot Dat file. Mascot may be able to use the reference file to get more accurate taxonomy information. It can also display the full text for an protein hit in the Protein View report. Even when a full text file is not available for download, Mascot may be able to retrieve equivalent text from a remote HTTP server, such as NCBI's Entrez or an SRS server.

If database entries contain taxonomy information, Mascot can use this as a filter during a search. Many of the most popular databases, such as Swiss-Prot and NCBI nr, include taxonomy. To determine taxonomy accurately, Mascot requires database specific supporting files. Details of these can be found in the help pages for the individual databases. Note that these supporting files have to be copied to the taxonomy directory, not to the sequence database directory. Also, some files need to be unpacked (using tar) as well as uncompressed.

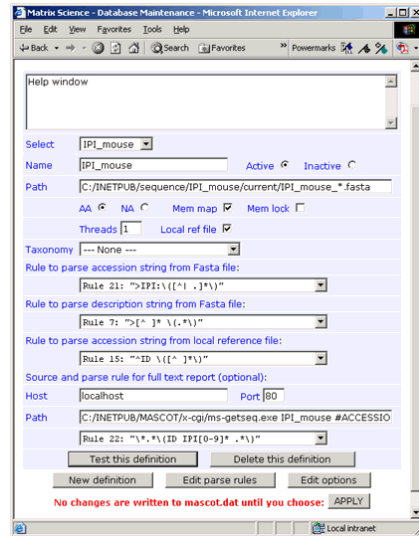
## Adding a New Database

### 4. Configure the database

- You can edit mascot.dat directly
- Easier and safer to use the Configuration Editor:
  - Less likely to make errors
  - Automatically creates a backup
  - Allows the configuration to be tested

### 5. Bring the database on-line

- Mascot
  - Compresses the Fasta file
  - Creates taxonomy indexes, if required
  - Runs a test search
  - Memory maps the files



**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



Step 4 is configure the database. The recommended way to do this is the Database Maintenance module of the Configuration Editor, accessible from a hyperlink on the Mascot home page. If you prefer, you can edit the configuration file, mascot.dat, in a text editor. However, be careful when editing configuration files and always make a backup. A small syntax error can stop the system working.

Step 5 is to bring the Database on-line. Once you 'Apply' a new definition, Mascot Monitor will look to see if there is a Fasta file that matches the specified path. If so, it will begin to compress the Fasta file, so as to minimise the memory requirements. If taxonomy has been defined for the database, Monitor will also create a taxonomy index.

Once this is complete, the new database is tested by running a standard search. If this succeeds, the new database is mapped into memory and becomes available for general use.

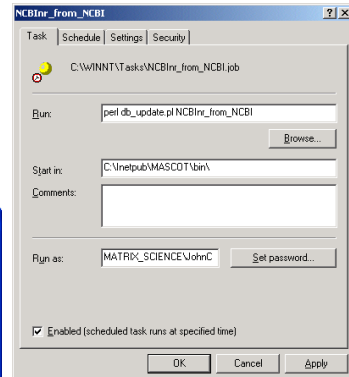


## Adding a New Database

### 6. Configure Automatic Updating

- New entries may be added frequently
- Mascot can update a database in background
- Automate using the Database Update script.

```
matrix@pubwww2: /home/matrix
matrix@pubwww2:~$
matrix@pubwww2:~$ crontab -l
# Send output to root
# All Mascot databases are updated on Sundays
00 4 * * * Sun $HOME/site/bin/db_update.pl NCBIInr_from_NCBI > /dev/null 2>&1
00 8 * * * Sun $HOME/site/bin/db_update.pl EST_human_from_NCBI > /dev/null 2>&1
00 10 * * * Sun $HOME/site/bin/db_update.pl EST_mouse_from_NCBI > /dev/null 2>&1
##00 12 * * * Sun $HOME/site/bin/db_update.pl EST_others_from_NCBI > /dev/null
00 17 * * * Sun $HOME/site/bin/db_update.pl Spot_varsplc_from_EBI > /dev/null 2>&1
00 20 * * * Sun $HOME/site/bin/db_update.pl MSDB_from_EBI > /dev/null 2>&1
matrix@pubwww2:~$
```



**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science 

Step 6 is optional; to configure automatic updating. Downloading and processing database updates by hand is tedious. Once the general procedure has been verified, it can be automated using the Database Update script. Under Unix, you would use Cron to schedule when this script should run. Under Windows, you might use Windows Scheduled Tasks.

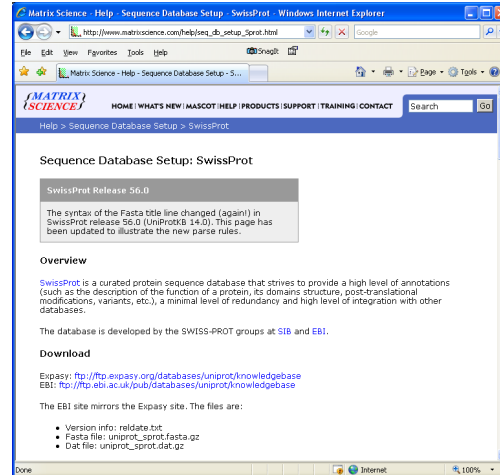
Databases can then be updated as often as you wish, with no disruption for Mascot users. Whenever Monitor sees a new Fasta file that matches the defined path, the new database is compressed and tested. If errors are detected in the new database, the database exchange process is abandoned.

Assuming the test is successful, all new searches are performed against the new database, while searches that are in progress against the old database are allowed to continue. Once the final search against the old database is complete, it is unmapped from memory and the files moved to the "old" directory. The new database is then memory mapped and the system becomes ready for the next update cycle.

## Example: SwissProt

### URL's and file names change constantly

- Up to date information about the major public databases can be found on the Matrix Science web site



**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



Our first example is Swiss-Prot, a high quality, well annotated protein databases. URL's and file names change constantly. For Swiss-Prot, and other major databases, the latest information can be found on the Matrix Science web site. Look under Help; Sequence Database Setup.

## Example: SwissProt

### Primary FTP sites are

- Expasy: <ftp://ftp.expasy.org/databases/uniprot/knowledgebase>
- EBI: <ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase>

### Files are:

- Version info: [reldate.txt](#)
- Fasta file: [uniprot\\_sprot.fasta.gz](#)
- Dat file: [uniprot\\_sprot.dat.gz](#)

### If you want to filter entries by taxonomy, you will also need

- <ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz>
- <ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/docs/speclist.txt>

You can download Swiss-Prot files from either Expasy or EBI. The EBI site mirrors the Expasy site. There are three files: version information, a Fasta file, and a full annotation text Dat file.

Taxonomy for Swiss-Prot is pre-defined in `mascot.dat`. Even if you download the Swiss-Prot Dat file, choose "Swiss-prot FASTA", because this is more accurate. Note that the taxonomy files go into the taxonomy directory, not into the sequence database directory. Also, some files need to be unpacked (using `tar`) as well as uncompressed.

## Example: SwissProt

### Parse rules

- Parse rules are basic regular expressions
- A typical SWISS-PROT Fasta title line is:  
`>sp|Q4U9M9|104K_THEAN 104 kDa microneme-rhoptry antigen precursor`
- You can use either the ID (104K\_THEAN) or the AC (Q4U9M9) as the identifier
- Many people prefer the ID because it is descriptive.  
ID from Fasta title: `">..|[^\|]*|\([^\ ]*\)"`  
AC from Fasta title: `">..|\([^\|]*\)"`  
Description from Fasta title: `">[^\ ]* \(.*\)"`
- The corresponding line in the Dat file is:  
ID 104K\_THEAN Reviewed; 893 AA.  
ID from Ref file: `"^ID \([^\ ]*\)"`

As mentioned earlier, we need to parse a unique identifier from each entry.

You can choose either the ID (104K\_THEAN) or the AC (Q4U9M9) as the identifier. Many people prefer the ID because it is descriptive.

A regular expression that skips over the characters up to and including the accession and the following pipe symbol then takes everything up to the first space will extract the ID from the Fasta title line. The description is then becomes everything after the first space.

For the Dat file, a different rule is needed because the syntax is different. In this case, we are looking for a line that starts with ID, (the circumflex represents the start of the line), followed by 3 spaces. Then we take everything that is not a space.

You'll find that all these parse rules are pre-defined in mascot.dat, it is just a case of selecting them.

## Example: SwissProt

**Wild card in file name**

**Always memory map**

**Local copy of DAT file**

**Taxonomy from Fasta**

**Always test**

**MASCOT : Sequence Database Administration** © 2007-2010 Matrix Science **MATRIX SCIENCE**

In fact, SwissProt is easy because the complete configuration is pre-defined in mascot.dat. It just isn't enabled.

Some points to watch:

The wild card is important. First because it masks the time-stamp or version number. Second, because it allows the database to be updated without interrupting ongoing searches. Even if you don't want to use a time-stamp or version number, you must still include a wild card. Note that the wild card goes in the name, not in the file extension. If you use a wild card extension, Mascot won't be able to distinguish the Fasta file from the Reference file, with interesting results

All databases should be memory mapped, because this makes access much faster. But, unless you have lots of RAM and a 64-bit OS, only the smaller databases, which are searched regularly, should be locked in memory. If you try to lock a database in memory and there isn't enough room, the operation fails, and everything is OK. The real problem is when there is just enough RAM to lock the database, but very little left over for Mascot searches and other applications. Searches will then be very slow, the disk will thrash, and eventually the system is likely to crash or hang.

Checking the "local ref file" box indicates that you have downloaded a local copy of the Dat file. The utility will try to catch conflicts, such as looking for a full text report in a local dat file even though this checkbox is clear.

We used to recommend using the organism line in the Dat file to determine taxonomy, but now we find we can get better accuracy from the ID in the Fasta file, so if you have an old definition, you might want to update it with this change

Always test a new definition before applying the changes to mascot.dat.

## Example: SwissProt



Mascot Database Maintenance

Testing Database Definition SwissProt

Testing entries at beginning and end of  
C:/sequence/SwissProt/current/SwissProt\_57.14.fasta:

Accession	Description
002R_IIV3	Uncharacterized protein 002R OS=Invertebrate iridescent virus 3 GN=IV3-002R PE=4 SV=1
003L_IIV3	Uncharacterized protein 003L OS=Invertebrate iridescent virus 3 GN=IV3-003L PE=4 SV=1
005L_IIV3	Uncharacterized protein 005L OS=Invertebrate iridescent virus 3 GN=IV3-005L PE=4 SV=1
006L_IIV6	Putative KIAA domain-containing protein 006L OS=Invertebrate iridescent virus 6 GN=IIV6-006L PE=3 SV=1
007R_IIV3	Uncharacterized protein 007R OS=Invertebrate iridescent virus 3 GN=IV3-007R PE=4 SV=1
Z_SABVB	RING finger protein Z OS=Sabia virus (isolate Human/Brazil/SPH114202/1990) GN=Z PE=3 SV=1
Z_SHEEP	Putative uncharacterized protein Z OS=Ovis aries PE=4 SV=1
Z_TACV	RING finger protein Z OS=Tacaribe virus GN=Z PE=1 SV=3
Z_TAMVU	RING finger protein Z OS=Tamiami virus (isolate Rat/United States/W 10777/1964) GN=Z PE=3 SV=1
Z_WWAVU	RING finger protein Z OS=Whitewater arrowy virus (isolate Rat/United States/WV 9310135/1995) GN=Z PE=3 SV=1

Local reference file tests OK  
(C:/sequence/SwissProt/current/SwissProt\_57.14.dat)

[Return to database definitions](#)

**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



The test checks the parse rules against the first five and last five entries in the database. It will also pick up problems with paths, illegal characters in names, etc. If the database checks out, this is a good sign. But, it isn't a guarantee of success. It is possible that the parse rules succeeded on the tested entries, but will fail somewhere else in the file. It is possible that duplicate accessions may be discovered. The taxonomy files may be missing, etc., etc.

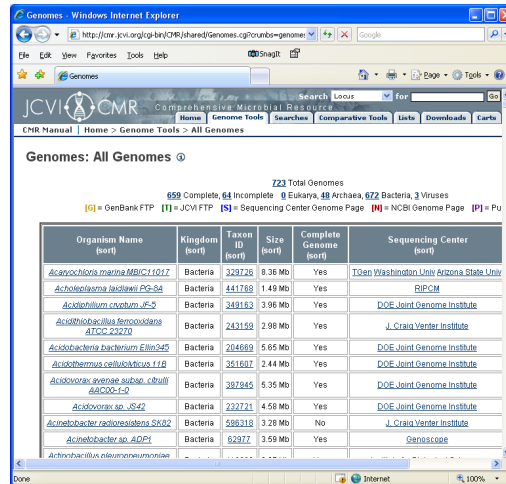
## Example: simple database

### Most databases use a very simple configuration

- All entries have the same taxonomy
- No full text reference file

### For example

- cmr.jcvi.org
- Microbial genomes
- *Helicobacter pylori*



The screenshot shows a web browser displaying the JCVI CMR Genomes database. The page title is "Genomes: All Genomes". It shows a list of 223 total genomes, with 659 complete and 64 incomplete. The list includes columns for Organism Name, Kingdom, Taxon ID, Size, Complete Genome, and Sequencing Center. The following table represents the data shown in the screenshot:

Organism Name (sort)	Kingdom (sort)	Taxon ID (sort)	Size (sort)	Complete Genome (sort)	Sequencing Center (sort)
<a href="#">Acetivibrio maehii M8C11017</a>	Bacteria	229726	8.36 Mb	Yes	DOE Washington Univ. Arizona State Univ.
<a href="#">Acholeplasma laidlawii PG-54</a>	Bacteria	441760	1.48 Mb	Yes	BIPCM
<a href="#">Acetobacterium caryum AT-5</a>	Bacteria	249163	3.96 Mb	Yes	DOE Joint Genome Institute
<a href="#">Acetivibrio maehii fensholtensis ATCC 23270</a>	Bacteria	243159	2.98 Mb	Yes	J. Craig Venter Institute
<a href="#">Acetobacterium baumannii ATCC 29245</a>	Bacteria	204689	5.65 Mb	Yes	DOE Joint Genome Institute
<a href="#">Acetivibrio maehii cellulosus ATCC 29245</a>	Bacteria	351607	2.44 Mb	Yes	DOE Joint Genome Institute
<a href="#">Acetivibrio maehii cellulosus ATCC 29245</a>	Bacteria	387945	5.35 Mb	Yes	DOE Joint Genome Institute
<a href="#">Acetivibrio maehii cellulosus ATCC 29245</a>	Bacteria	232721	4.58 Mb	Yes	DOE Joint Genome Institute
<a href="#">Acetivibrio maehii cellulosus ATCC 29245</a>	Bacteria	589318	3.28 Mb	No	J. Craig Venter Institute
<a href="#">Acetivibrio maehii cellulosus ATCC 29245</a>	Bacteria	62377	3.58 Mb	Yes	Genoscope

**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



For our second example, let's look at the very common case of a sequence database where all the entries have the same taxonomy and there is no full text reference file. JCVI, The J. Craig Venter Institute, has a list of completed microbial genomes.

## Example: simple database

### Look at the Fasta file to choose a parse rule

```
>HP0001 hypothetical protein {Helicobacter pylori 26695}  
MATRTQARGAVVELLYAFESGNEEIKKIASSMLEEKKIKNNQLAFALSLFNGVLEKINEI  
DALIEPHLKDWDFFKRLGSM EKAILRLGAYEIGFTPTQNP I I I N E C I E L G K L Y A E P N T P K F  
LNAILDSL SKKLTQKPLN  
>HP0002 riboflavin synthase beta chain (ribE) {Helicobacter pylori  
26695}  
MQIIEGKLQLQGNERVAITSRFNHIITDRLQEGAMDCFKRHGGDELLDIVLVPGAYEL  
PFILDKLLESEKYDGVCLGAIIRGGTPHFDYVSAEATKGI AHAMLYSMPVSVFGLTTD  
NIEQAIERAGSKAGNKGFEAMSTLIELLSLCQTLKG  
>HP0003 3-deoxy-d-manno-octulosonic acid 8-phosphate synthetase  
(kdsA) {Helicobacter pylori 26695}  
MKTSKTKTPKSVLIAGPCVIESLENLRSIATKLQPLANNERLDFYFKASF DKANRTSLES  
YRGPGLKGL EMLQTIKEEFGYKILTDVHESYQASVAKVADILQIPAFLCRQTDLIVEV
```

### Can usually use rules 4 and 5

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

**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



To decide on a suitable parse rule, you need to examine the title lines. If the database file is a large one, it may not be a good idea to open it in a standard word processor or text editor. Most platforms support a command line utility called *more* that can be used to browse a file of any size. In the case of `h_pylori_26695.pep`, the first few lines look like this

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 can be treated as the description. These rules are pre-defined in `mascot.dat` as rules 4 and 5.



## Example: simple database

### Parse Rule Tips

- If a rule looks like it should work, and doesn't, it may be because a space is actually a tab.  
Character class of all printing characters is [!~]
- Don't make a parse rule more precise than it needs to be
- Several parse rules are pre-defined in mascot.dat. Experiment with these before writing a new one.
- If you need to edit a large sequence database file under Windows, you will need an editor that can edit the file without reading it all into memory, e.g. UltraEdit.

If a rule looks like it should work, and doesn't, it may be because the space is actually a tab. If this is the case, then you can use a character class that includes or excludes all the printing characters

Don't make a parse rule more precise than it needs to be. A rule which is too picky is more likely to fail. The goal is simply to get a unique identifier from each entry

Several parse rules are pre-defined in mascot.dat. Experiment with these before writing a new one. If you have to write a new one, remember that these are Basic Regular Expressions, as used in grep, not Extended Regular Expressions, as used in Perl.

If you need to edit a large sequence database file under Windows, you will need an editor that can edit the file without reading it all into memory. One such editor is UltraEdit - <http://www.ultraedit.com/>

## Example: simple database

Consistent, descriptive,  
case-sensitive name

Choose AA or NA

No taxonomy

No local reference file

Always test

The screenshot shows the 'Mascot Database Maintenance: Edit Database Definitions' window. The configuration is as follows:

- Name:** h\_pylori
- Path:** C:/inetpub/MASCOT/sequence/h\_pylori/current/h\_pylori\*.fasta
- AA/NA:** AA (selected)
- Local ref file:**  (unchecked)
- Taxonomy:** --- None ---
- Rule to parse accession string from Fasta file:** Rule 4: %>[ ]\*
- Rule to parse description string from Fasta file:** Rule 5: %>[ ]\* \.(.\*)
- Rule to parse accession string from local reference file:** --- no local reference file ---
- Source and parse rule for full text report (optional):** --- no full text report ---

Buttons at the bottom include: Test this definition, Delete this definition, New definition, Edit parse rules, Edit options, and APPLY. A red message at the bottom states: 'No changes are written to mascot.dat until you choose: APPLY'.

**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



Put a bit of thought into a name that will be unambiguous to users, yet isn't too long.

If you chose one of the nucleic acid files, rather than the protein file, you would need to select NA using this radio button

Because this database is sequences from a single organism, no taxonomy information is required

The local reference file checkbox is clear in this case.

Test, test, test ...

If the database tests OK, choose Apply to save the new configuration in mascot.dat, and follow the link to Database Status

## Database Status

"Old" & "New"

Compression warnings

Unidentified taxonomy

Statistics

```
Mascot search status page - Microsoft Internet Explorer
New mapped = YES Request to mem map = YES Request unmap = NO Mem locked = YES
Number of threads = 1 Current = YES

Name = NCBInr_20041113_ Family = /home/matrix/site/sequence/NCBInr/current/NCBInr_
Filename = NCBInr_20041113.fasta Pathname = /home/matrix/site/sequence/NCBInr/current/NCBInr_
Status = Not in use Statistics
State Time = Sun Nov 21 04:39:43 # searches = 0
New mapped = NO Request to mem map = YES Request unmap = NO Mem locked = NO
Number of threads = 1 Current = NO

Name = NCBInr_20041117_ Family = /home/matrix/site/sequence/NCBInr/current/NCBInr_
Filename = NCBInr_20041117.fasta Pathname = /home/matrix/site/sequence/NCBInr/current/NCBInr_
Status = In use Statistics Compression warnings Unidentified taxon
State Time = Sun Nov 21 04:39:46 # searches = 0
New mapped = YES Request to mem map = YES Request unmap = NO Mem locked = YES
Number of threads = 1 Current = YES

Name = EST_human_20041113_ Family = /home/matrix/site/sequence/EST_human/current/EST_
Filename = EST_human_20041113.fasta Pathname = /home/matrix/site/sequence/EST_human/current/
Status = Not in use Statistics
State Time = Sun Nov 21 09:24:39 # searches = 0
New mapped = NO Request to mem map = YES Request unmap = NO Mem locked = NO
Number of threads = 1 Current = NO

Name = EST_human_20041117_ Family = /home/matrix/site/sequence/EST_human/current/EST_
Filename = EST_human_20041117.fasta Pathname = /home/matrix/site/sequence/EST_human/current/
Status = In use Statistics
State Time = Sun Nov 21 09:24:40 # searches = 0
New mapped = YES Request to mem map = YES Request unmap = NO Mem locked = NO
Number of threads = 1 Current = YES

Name = EST_mouse_20041113_ Family = /home/matrix/site/sequence/EST_mouse/current/EST_
Filename = EST_mouse_20041113.fasta Pathname = /home/matrix/site/sequence/EST_mouse/current/
Status = Not in use Statistics
State Time = Sun Nov 21 10:49:52 # searches = 0
New mapped = NO Request to mem map = YES Request unmap = NO Mem locked = NO
Number of threads = 1 Current = NO

Name = EST_mouse_20041117_ Family = /home/matrix/site/sequence/EST_mouse/current/EST_
Filename = EST_mouse_20041117.fasta Pathname = /home/matrix/site/sequence/EST_mouse/current/
Status = In use Statistics
State Time = Sun Nov 21 10:49:54 # searches = 0
New mapped = YES Request to mem map = YES Request unmap = NO Mem locked = NO
Number of threads = 1 Current = YES
```

**MASCOT** : Sequence Database Administration © 2007-2010 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.

Taxonomy is rarely 100% accurate. Usually, there will be a small number of failures.

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%?

The top screenshot displays the following statistics:

- Time files compressed (int): Sun Nov 21 04:22:43 2004
- Time files compressed (int): 110110943
- Time / date of fasta file: Wed Nov 17 08:27:00 2004
- Time of fasta files (int): 1100680020
- Number of residues: 73898844
- Number of sequences: 2371939
- 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 (archaeobacteria)=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

The bottom-left screenshot shows a table of sequence lengths and their counts:

Length	Count
0	3436
1	1177306
2	631043
3	189425
4	78910
5	30514
6	19481
7	11704
8	5020
9	5003
10	2745
11	2010
12	1296
13	1000

The bottom-right screenshot shows a table of sequence lengths and their counts, along with a list of sequence lengths and their counts:

Length	Count
6	692
7	828
8	857
9	976
10	1195
11	1014
12	1076
13	1208

**MASCOT** : Sequence Database Administration © 2007-2010 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

## Database Update Script

### Enables database updating to be automated using Unix Cron or Windows Scheduled Tasks

- downloading a fasta file plus optional associated files such as a full text reference file or release notes
- optionally downloading one or more taxonomy indexes
- handling variable filenames via wild cards
- uncompressing, unpacking, renaming and moving the files
- time or version stamping
- downloading a file only if a new one is available; resuming an interrupted download
- passive FTP through a firewall; HTTP proxy server authentication.

**MASCOT** : *Sequence Database Administration* © 2007-2010 Matrix Science

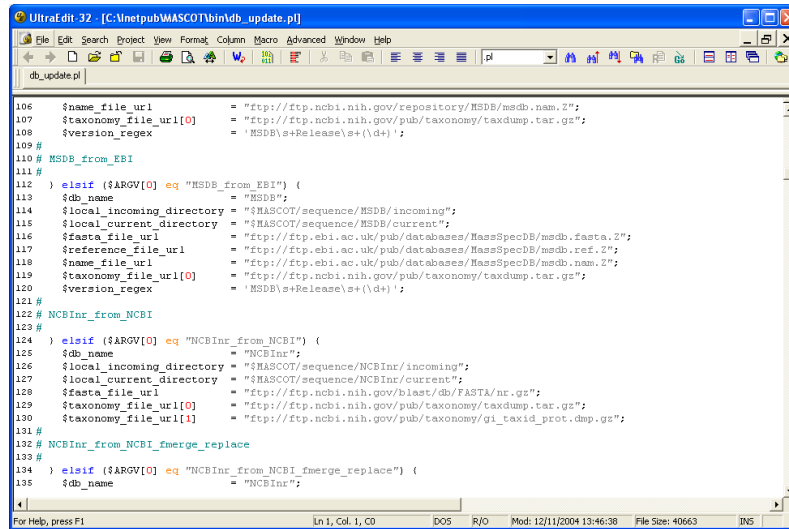


Once a database has been configured, tested, and brought on-line, you'll probably want to automate the downloading of updates.

There is some complexity to this, and the functionality of the Mascot database update script includes:

- downloading a fasta file plus optional associated files such as a full text reference file or release notes
- optionally downloading one or more taxonomy indexes
- handling variable filenames via wild cards
- uncompressing, unpacking, renaming and moving the files
- time or version stamping
- downloading a file only if a new one is available; resuming an interrupted download
- passive FTP through a firewall; HTTP proxy server authentication

## Database Update Script



```
106 $name_file_url      = "ftp://ftp.ncbi.nih.gov/repository/MSDB/msdb.nam.2";
107 $taxonomy_file_url[0] = "ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz";
108 $version_regex      = 'MSDB\s+Release\s+(\d+)';
109 #
110 # MSDB_from_EBI
111 #
112 ) elsif ($ARGV[0] eq "MSDB_from_EBI") {
113     $db_name          = "MSDB";
114     $local_incoming_directory = "$MASCOT/sequence/MSDB/incoming";
115     $local_current_directory = "$MASCOT/sequence/MSDB/current";
116     $fasta_file_url   = "ftp://ftp.ebi.ac.uk/pub/databases/MassSpecDB/msdb.fasta.2";
117     $reference_file_url = "ftp://ftp.ebi.ac.uk/pub/databases/MassSpecDB/msdb.ref.2";
118     $name_file_url    = "ftp://ftp.ebi.ac.uk/pub/databases/MassSpecDB/msdb.nam.2";
119     $taxonomy_file_url[0] = "ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz";
120     $version_regex    = 'MSDB\s+Release\s+(\d+)';
121 #
122 # NCBInr_from_NCBI
123 #
124 ) elsif ($ARGV[0] eq "NCBInr_from_NCBI") {
125     $db_name          = "NCBInr";
126     $local_incoming_directory = "$MASCOT/sequence/NCBInr/incoming";
127     $local_current_directory = "$MASCOT/sequence/NCBInr/current";
128     $fasta_file_url   = "ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz";
129     $taxonomy_file_url[0] = "ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz";
130     $taxonomy_file_url[1] = "ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_prot.dmp.gz";
131 #
132 # NCBInr_from_NCBI_fmerge_replace
133 #
134 ) elsif ($ARGV[0] eq "NCBInr_from_NCBI_fmerge_replace") {
135     $db_name          = "NCBInr";
```

**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



The database update script is a Perl script, which is a text file called db\_update.pl  
You can open it in any text editor.

## Database Update Script

### Installation

- Requires Gnu utilities: wget, tar, gzip
- (Windows - <http://gnuwin32.sourceforge.net/>)
- Must be installed into a directory on the system search path

### Configuration

- May need to change
  - \$MASCOT, the path to the Mascot directories
  - Proxy server information
  - Database download URL's

### Adding additional databases

- Copy an existing database definition block and modify it

### Usage

- `db_update.pl NCBI`

**MASCOT** : Sequence Database Administration © 2007-2010 Matrix Science



`db_update.pl` is located in the Mascot bin directory. The following utilities are also required: `gzip`, `tar`, `wget`.

These are likely to be present on any Unix system. Windows ports of all three utilities can be downloaded from SourceForge. All three utilities should be installed into a directory on the system search path, so they can be executed from any directory, without having to provide path information.

If you have Mascot or sequence database on non-default paths, you'll need to modify certain paths in `db_update.pl`. Note that some definitions are specified independently for Unix and Windows, to minimise the need for editing. You only need to change the definitions for the platform you are using.

Detailed configuration information can be found in the Mascot Setup & Installation manual.

Several common database update definition blocks are pre-configured, and you may not need to add or change anything before using the script. A particular definition block is chosen by means of a keyword argument when the script is executed.

If the name or location of a download file changes, you will need to update the corresponding definition block. If you want to add a new database, the easiest way is to make a copy of a similar looking definition block and then modify it.

# Database Update Script

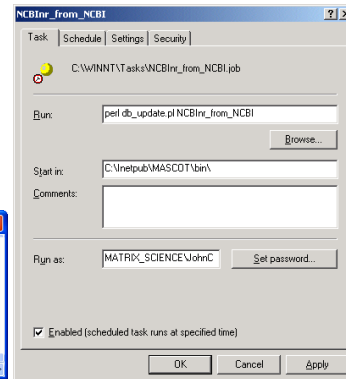
## Testing

- Execute the script from a command / shell prompt and from a directory other than the one in which the script is located
- Single log file - location is defined by the variable \$local\_log\_file in the script header

## Automation

- Unix Cron
- Windows Scheduled Tasks.

```
matrix@pubwww2: /home/matrix
matrix@pubwww2:~$
matrix@pubwww2:~$ crontab -l
# Send output to root
# All Mascot databases are updated on Sundays
00 4 * * Sun $HOME/site/bin/db_update.pl NCBInr_from_NCBI > /dev/null 2>&1
00 8 * * Sun $HOME/site/bin/db_update.pl EST_human_from_NCBI > /dev/null 2>&1
00 10 * * Sun $HOME/site/bin/db_update.pl EST_mouse_from_NCBI > /dev/null 2>&1
#00 12 * * Sun $HOME/site/bin/db_update.pl EST_others_from_NCBI > /dev/null
00 17 * * Sun $HOME/site/bin/db_update.pl Sprot_varsplc_from_EBI > /dev/null 2>&1
00 20 * * Sun $HOME/site/bin/db_update.pl MSDB_from_EBI > /dev/null 2>&1
matrix@pubwww2:~$
```



Before adding a new db\_update.pl entry to Unix Cron or Windows Scheduled Tasks, it is essential to test it. To test the functionality, you should execute the script at a shell or command prompt from a directory other than the one in which the script is located. This will ensure that directory permissions are correct and the paths can be resolved.

A single log file is maintained for all instances of db\_update.pl. The location is defined by \$local\_log\_file in the script header.

Once the script has been found to function correctly for a particular definition block, an entry can be added to Cron or Windows Scheduled Tasks. As a rule, you should stagger database updates through Mascot server quiet periods. Trying to update all the databases simultaneously will prolong the download times and may slow down any Mascot searches currently in progress. On our public web site, we update the databases every week. It is rarely worth doing it more frequently than this

Each file downloaded by FTP is listed in a file called .history, located in the corresponding incoming directory. This is used to prevent a given file being downloaded more than once. If you want to defeat this mechanism, simply delete or edit the .history file.



## Database Tips 1

**Always test a new database configuration**

**Check the statistics file after a new database has been compressed**

**Be selective when locking databases into memory**

- Only the smaller databases, which are searched regularly, should be locked in memory

**Can distribute sequence databases to empty drives**

**Include a date or version stamp in the database filename**

- That's what the wild card is there for

**Don't forget the taxonomy files**

- And ensure they are kept up to date!

**MASCOT** : *Sequence Database Administration* © 2007-2010 Matrix Science



This slide summarises some of the important tips mentioned earlier.

Always test a new database configuration

Check the statistics file after a new database has been compressed

Be selective when locking databases into memory

Only the smaller databases, which are searched regularly, should be locked in memory

Can distribute sequence databases to empty drives

Include a date or version stamp in the database filename

That's what the wild card is there for

Don't forget the taxonomy files

And ensure they are kept up to date!

## Database Tips 2

### Before adding a new database, select a similar one as template

- When the new database is added, all the rules are copied from the currently selected database

### Try to get the database name right first time!

- The name is case sensitive but Windows doesn't have case sensitive file names

### Don't download a database file onto your desktop

- It will then probably have insufficient permissions to be read by the database maintenance utility

### Check proxy settings in mascot.dat for remote reference files

### For manual updates, download to incoming, move to current

### If automatic updates have stopped, check help page.

To add a new database, simply press the 'New Definition' button in the database maintenance utility. This sets the defaults from the currently displayed definition, so it makes most sense to select one that is very similar to the new database

This is rather complex to describe and only affects Windows users, but try and get the database name correct first time. If you choose a name in all upper case, and then change it to lower case, you will need to delete the relevant 'test' file in the data\test directory.

Windows security can cause havoc! If you download a fasta file to your desktop and then move it to the current directory, chances are that the database maintenance utility won't be able to see it. You will need to set read permissions for everybody

If you select to display reference files from a remote source, then you may need to set proxy settings in the options section of the mascot.dat file. See the mascot installation and setup manual for further details

When performing a manual update, download the files to the incoming directory, rename them if necessary and then move them to the current directory

If automatic updates for one database suddenly stop for some reason, it's likely that the database provider has changed something. It's very likely that we have already noticed, and may have changed our help pages already

## Common Mistakes

### Forgetting the wild card in the database filename

- Even if you don't want to use a time-stamp or version number, you must still include a wild card

### Putting the wild card in the filename extension

- NCBIInr\_\*.fasta or NCBIInr\*.fasta, **not** NCBIInr.\*

### Using spaces or special characters in the database path

- Spaces in paths may be legal in Windows, but they shouldn't be!

### Using back slashes in the database path

- Even on Windows, all paths must use forward slashes

### Out of date taxonomy files

- Aim for > 99% accuracy

### Creating a sequence database with inconsistent title syntax

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

Conversely, here are some of the most common mistakes encountered by our technical support desk.

### Forgetting the wild card in the database filename

Even if you don't want to use a time-stamp or version number, you must still include a wild card

### Putting the wild card in the filename extension

NCBIInr\_\*.fasta or NCBIInr\*.fasta, **not** NCBIInr.\*

### Using spaces or special characters in the database path

Spaces in paths may be legal in Windows, but they shouldn't be!

### Using back slashes in the database path

Even on Windows, all paths must use forward slashes

### Out of date taxonomy files

Aim for > 99% accuracy

### Creating a sequence database with inconsistent title syntax

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