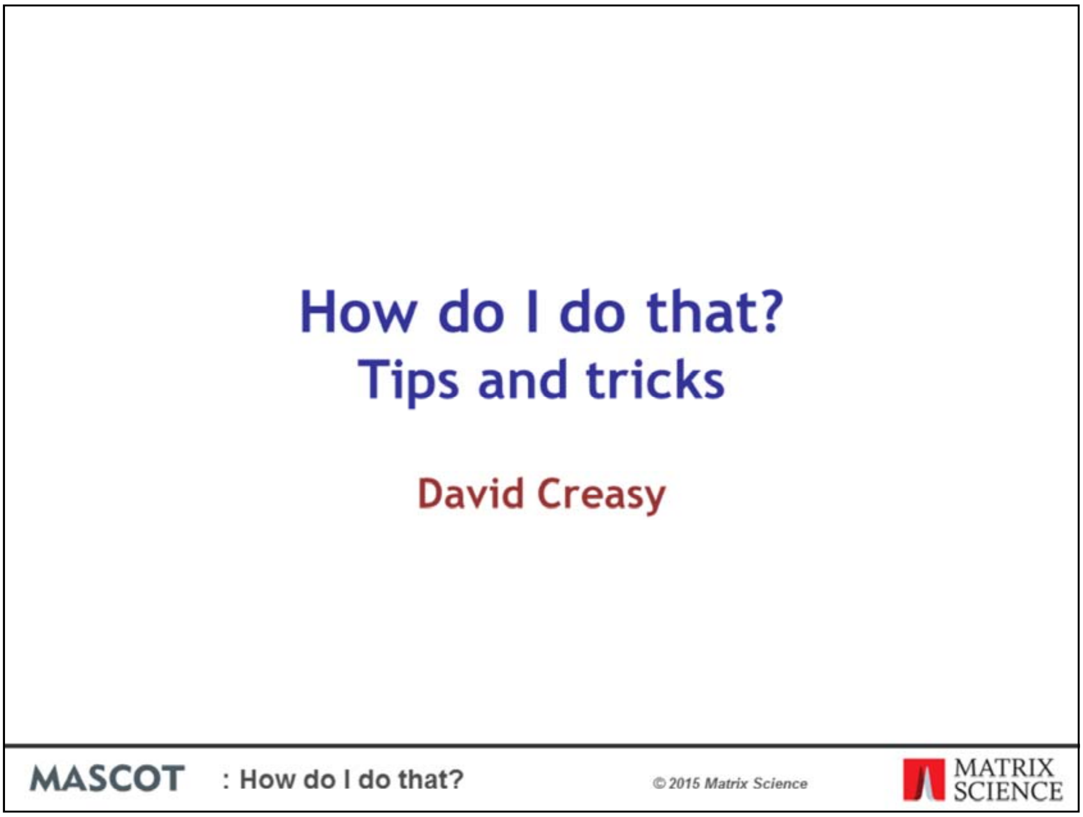
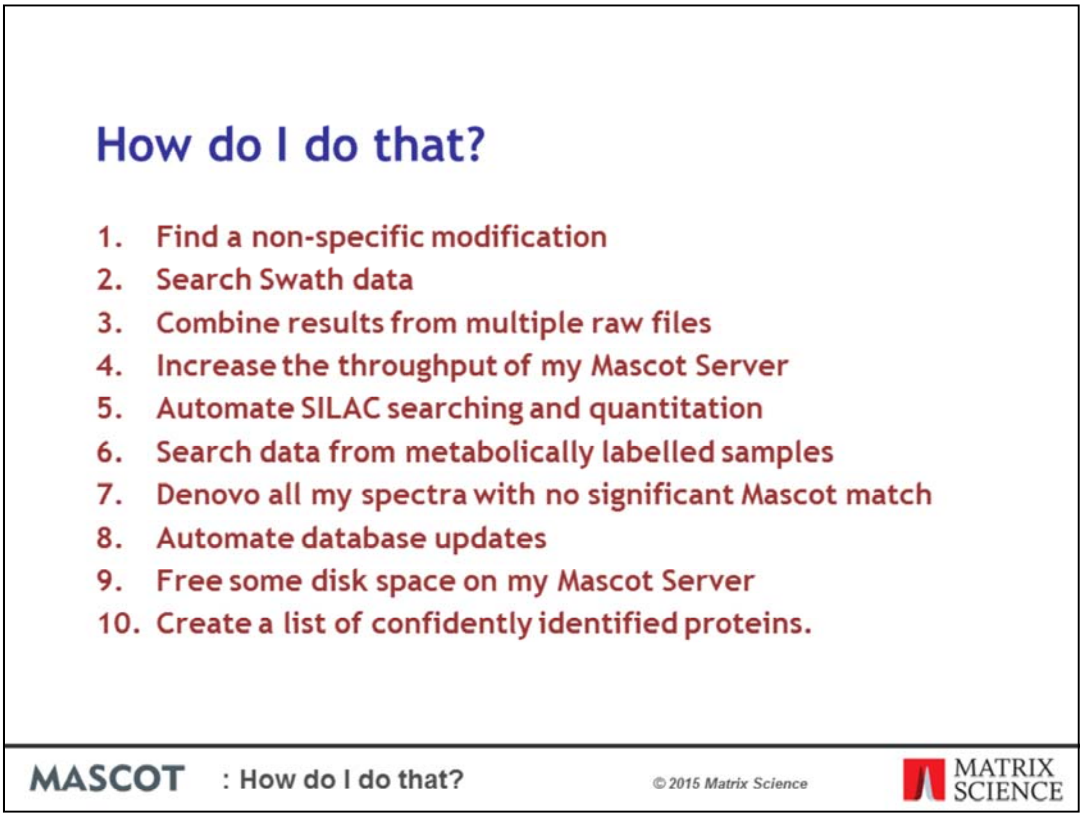


How do I do that?
Tips and tricks

David Creasy

MASCOT : How do I do that?

MATRIX SCIENCE

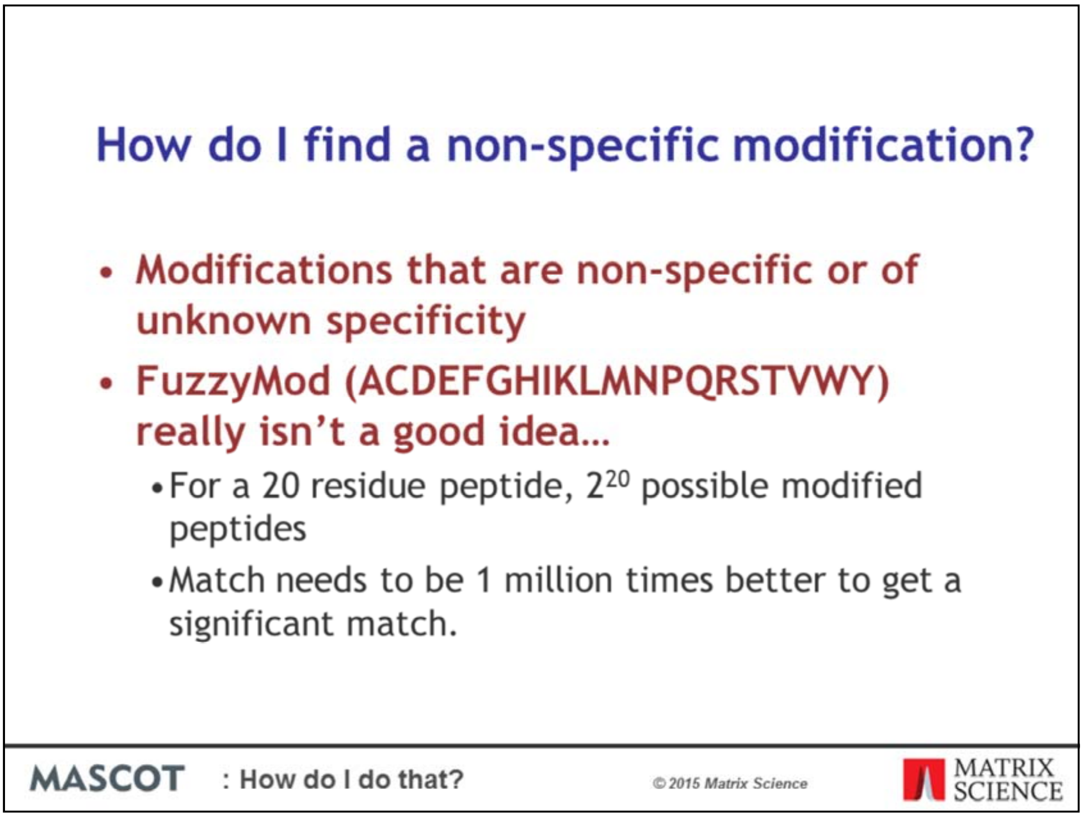


How do I do that?

1. Find a non-specific modification
2. Search Swath data
3. Combine results from multiple raw files
4. Increase the throughput of my Mascot Server
5. Automate SILAC searching and quantitation
6. Search data from metabolically labelled samples
7. Denovo all my spectra with no significant Mascot match
8. Automate database updates
9. Free some disk space on my Mascot Server
10. Create a list of confidently identified proteins.

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

I'll be covering 12 short topics today. Most of these come up regularly as support questions and some of them have been entries on our blog over the last couple of years. I'll start with how to find a non-specific modification.



How do I find a non-specific modification?

- Modifications that are non-specific or of unknown specificity
- FuzzyMod (ACDEFGHIKLMNPQRSTVWY) really isn't a good idea…
  - For a 20 residue peptide, 2^{20} possible modified peptides
  - Match needs to be 1 million times better to get a significant match.

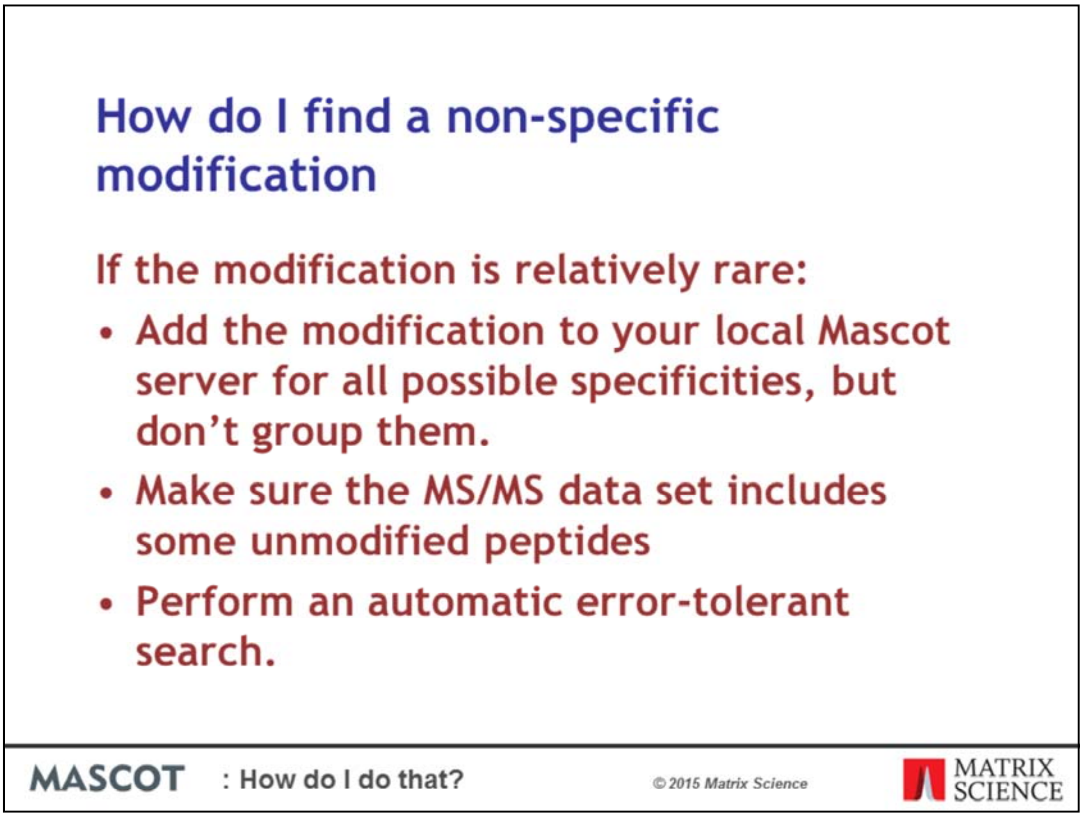
MASCOT : How do I do that?

MATRIX SCIENCE

Sometimes, we need to search for modifications that are non-specific or of unknown specificity. You might be tempted to create a modification such as FuzzyMod (ACDEFGHIKLMNPQRSTVWY). If you try to use such a modification in a search, Mascot will almost certainly run out of memory or address space and crash because of the combinatorial explosion.

Consider the numbers. If modification of any residue is possible then, for a single 20 residue peptide from the database, there are 2^20 possible modified peptides that need to be tested to see if they fit to the precursor mass and, if they do, matched to the MS/MS spectrum. This increases the search space by a factor of 1 million. Even if the code can handle this in a reasonable amount of memory, you can't escape from the fact that any match needs to be 1 million times better than if you were searching without the modification to be statistically significant.

If multiple modifications per peptide are expected, this is a very difficult problem for database search, and is probably intractable. A truly non-specific modification is going to produce a population of peptides which won't all be modified in the same way. Many arrangements of the modification will be represented, and sets of these will be isobaric. Thus, each MS/MS spectrum will contain fragments from a mixture of precursors and so will be of poor quality in terms of database matching.



How do I find a non-specific modification

If the modification is relatively rare:

- Add the modification to your local Mascot server for all possible specificities, but don’t group them.
- Make sure the MS/MS data set includes some unmodified peptides
- Perform an automatic error-tolerant search.

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

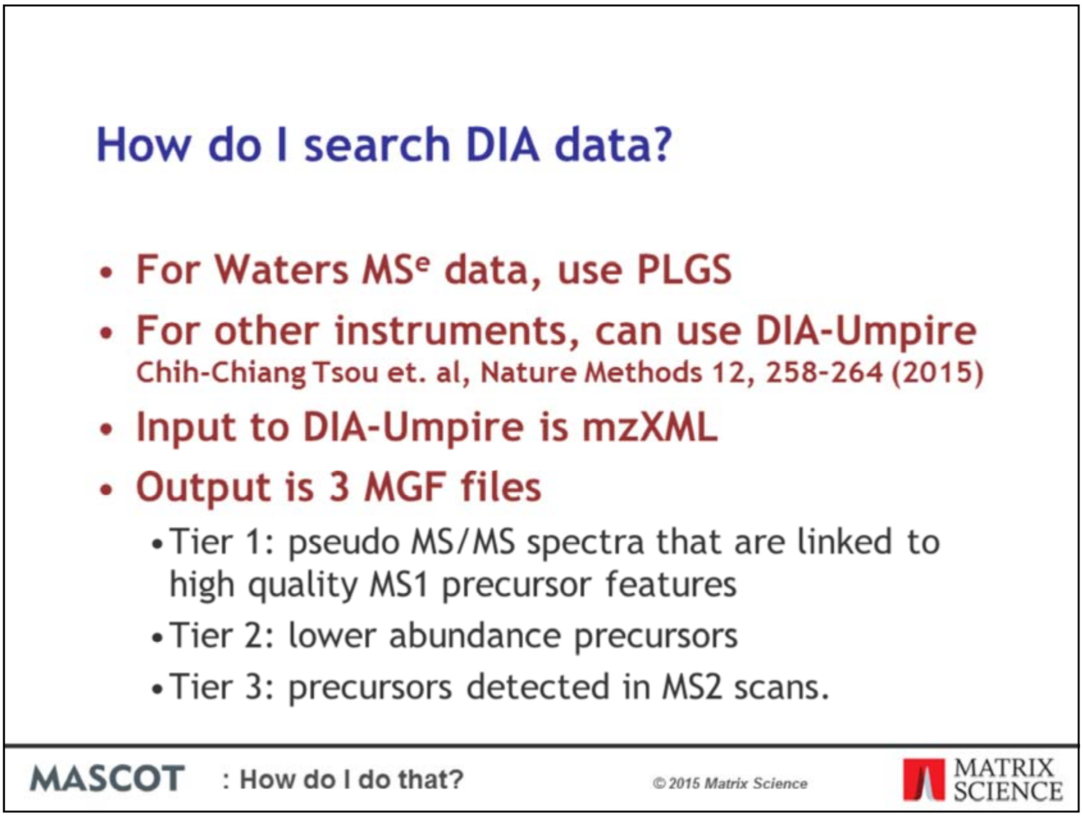
However, if the modification is relatively rare, and you don’t expect more than one instance on most peptides, such as a cross-linker being used to interrogate protein-protein interactions, the problem is much simpler:

Add the modification to your local Mascot server for all possible specificities, but don’t group them. This creates up to 20 separate new modifications.

Make sure the MS/MS data set includes some unmodified peptides so that you get a hit to the protein. If necessary, spike in some unmodified protein.

Perform an automatic error-tolerant search.

If the spectra are good, the highest scoring match for a peptide will have the modification at the correct location. If the spectra are not so good, you may get matches with similar scores for a range of possible locations. Note that the error-tolerant search will not give matches to peptides that carry the new modification with different specificities; a peptide with two FuzzyMod (H) would be OK but you wouldn’t match a peptide with FuzzyMod (H) and FuzzyMod (M).



How do I search DIA data?

- For Waters MS^e data, use PLGS
- For other instruments, can use DIA-Umpire
  Chih-Chiang Tsou et. al, Nature Methods 12, 258-264 (2015)
- Input to DIA-Umpire is mzXML
- Output is 3 MGF files
  - Tier 1: pseudo MS/MS spectra that are linked to high quality MS1 precursor features
  - Tier 2: lower abundance precursors
  - Tier 3: precursors detected in MS2 scans.

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

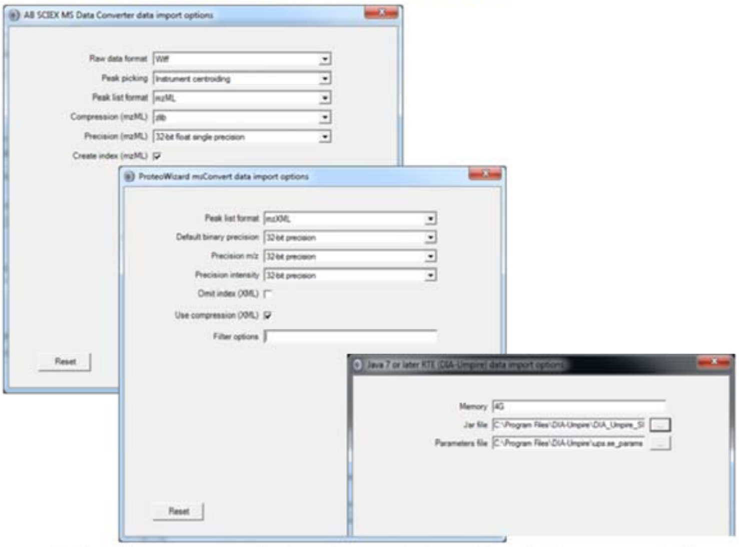
For Waters MS^e data, use PLGS to convert to MGF and search with Mascot

DIA-Umpire is a new, open source Java program that enables untargeted peptide and protein identification and quantitation using DIA data. A detailed description can be found in this Nature Methods paper. The DIA-Umpire signal extraction module deconvolutes the DIA data to create a conventional DDA-type peak list, suitable for database searching. The software is intended to be applicable to DIA from any instrument, although the examples in the publication are mostly Swath on an AB Sciex 5600, with one example of data from a Q Exactive Plus.

The choice of mzXML as the input format is slightly unfortunate, as this became obsolete in 2008, when HUPO PSI released mzML. The suggested workflow for AB Sciex data is to convert Wiff to mzML using the AB Sciex MS Data Converter then convert mzML to mzXML using msconvert.

The output from DIA-Umpire is three MGF files for each input file: Tier 1 (*_Q1.mgf) representing pseudo MS/MS spectra that are linked to high quality MS1 precursor features (3 or more detected isotope peaks), tier 2 (*_Q2.mgf) representing lower abundance precursors (2 detected isotope peaks only), and tier 3 (*_Q3.mgf) representing precursors detected in MS2 scans.

How do I search DIA data?



AB SCIEX MS Data Converter data import options

Raw data format: Wiff
Peak picking: Instrument centroiding
Peak list format: mzML
Compression (mzML): zlib
Precision (mzML): 32-bit float single precision
Create index (mzML): ☑

ProteoWizard msConvert data import options

Peak list format: mzXML
Default binary precision: 32-bit precision
Precision m/z: 32-bit precision
Precision intensity: 32-bit precision
Omit index (XML): □
Use compression (XML): ☑
Filter options:

Java 7 or later RTE (DIA-Umpire) data import options

Memory: 4G
Jar file: C:\Program Files\DIA-Umpire\DIA_Umpire_SI
Parameters file: C:\Program Files\DIA-Umpire\ups.se_params

http://www.matrixscience.com/blog/searching-dia-data-especially-swath.html

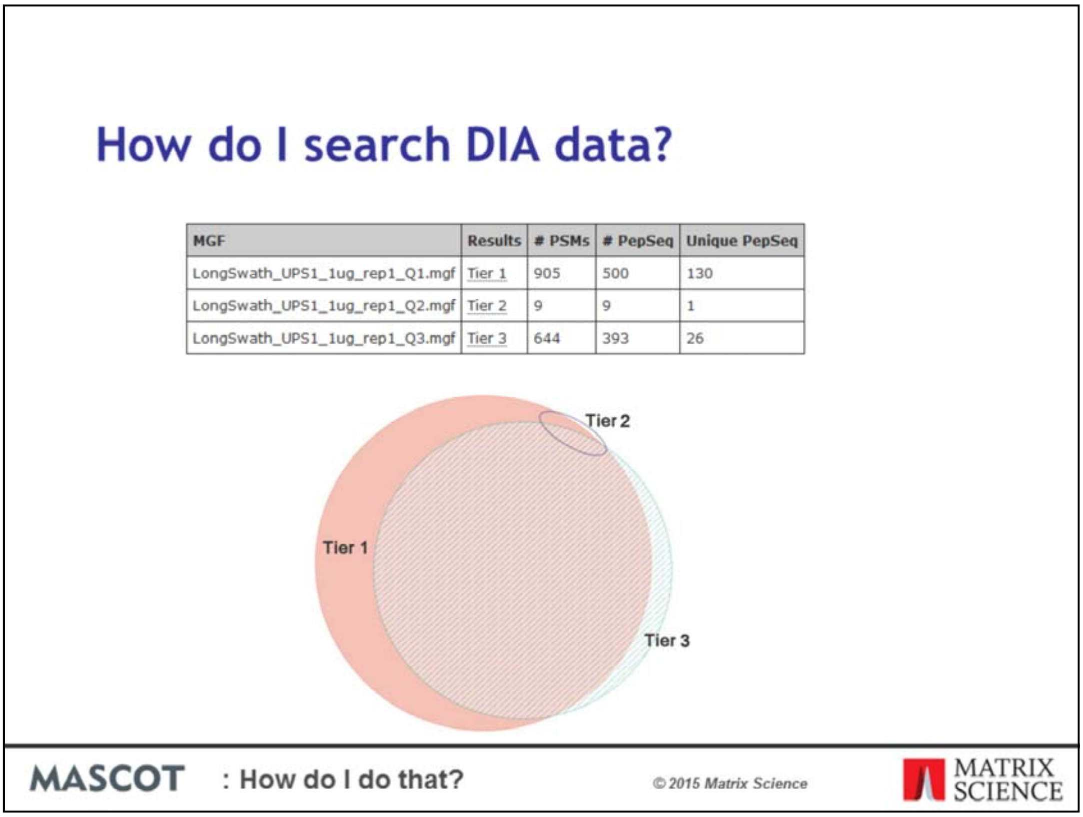
The file conversion requirements make automation tricky. Mascot Daemon 2.5 supports both AB Sciex MS Data Converter and msconvert, but not as multiple steps in a file conversion chain. It's a bit tricky to set up, but you can create four separate real-time monitor tasks:

- The first one to converts Wiff to mzML. You will need to get it to monitor a folder somewhere on your system and search for *.wiff files. These are suitable parameters for the AB SCIEX MS Data Converter
- Next, create a second real-time monitor task to convert mzML to mzXML using the ProteoWizard mzConvert import filter. These are suitable parameters for the import filter. You will need to monitor the directory where the first task, for example: C:\ProgramData\Matrix Science\Mascot Daemon\MGF\20 Wiff to mzML
- The third task will create MGF files from the mzXML using DIA-Umpire. Once again it needs to monitor the folder from the previous task, and these are the relevant options for DIA-Umpire
- A final task is required to search each of the MGFs. In this case you need to look for the *.mgf files in the second task because DIA-Umpire doesn't allow the output directory for the MGFs to be specified. , so there has to be an additional task to pick up the MGFs from the same directory as the mzXML.

You need to set up some search parameters for the first three tasks that will only search a small database. The first task will succeed with the search, but for the second two, you should just ignore the errors in the Daemon event log from the searches which are bound to fail.

You can download an updated daemon_di_utils.xml, containing a definition for DIA-Umpire and with the mzXML output option added to the msconvert definition. Note that there is an error in the DIA-Umpire instructions: when using msconvert to transfrom mzML to mzXML, do not specify peak picking because this causes msconvert to throw an exception.

Download links at: http://www.matrixscience.com/blog/searching-dia-data-especially-swath.html



How do I search DIA data?

MGF	Results	# PSMs	# PepSeq	Unique PepSeq
LongSwath_UPS1_1ug_rep1_Q1.mgf	Tier 1	905	500	130
LongSwath_UPS1_1ug_rep1_Q2.mgf	Tier 2	9	9	1
LongSwath_UPS1_1ug_rep1_Q3.mgf	Tier 3	644	393	26

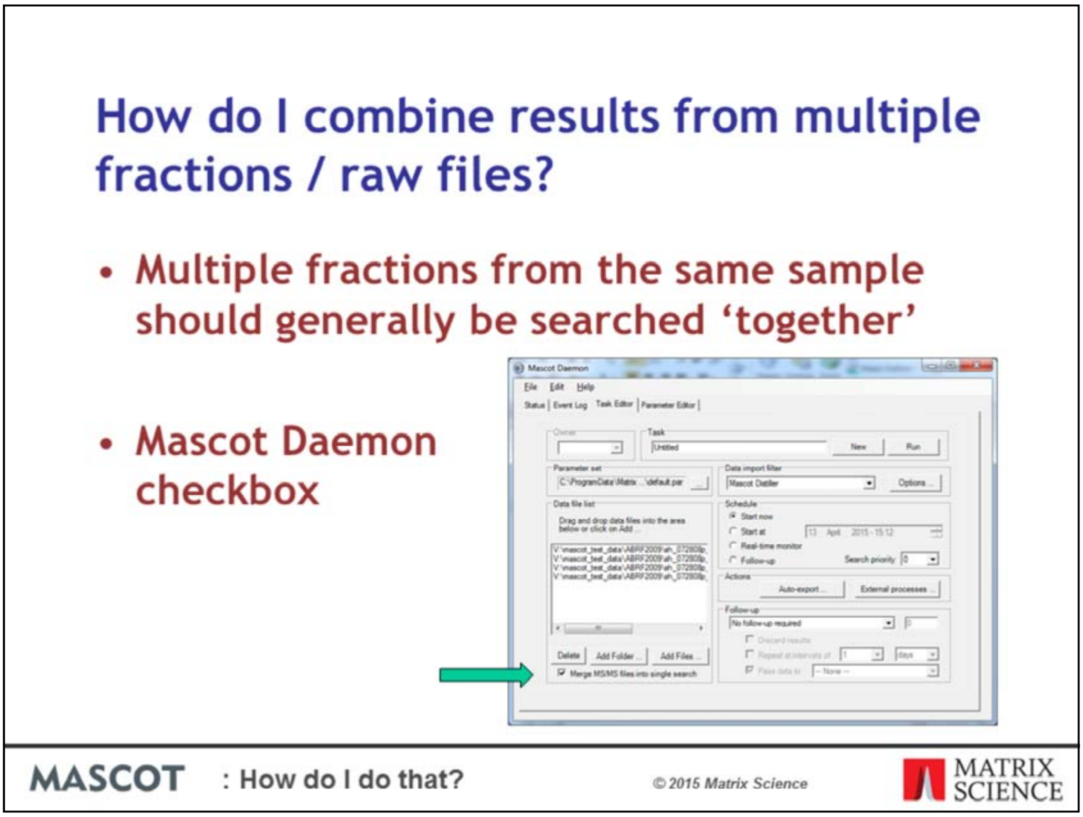
MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

Example data files described in the text and supplemental information can be downloaded from SourceForge and more are available from Pride.

I'll show the results from just one data set, the example LongSwath_UPS1_1ug_rep1.mzXML using the parameters (ups.se_params) suggested in the paper. This yielded three MGF files. If you remember from the earlier slide, the tier 1 mgf represents the pseudo MS/MS spectra that are linked to high quality MS1 precursor features, the tier 2 represents lower abundance precursors, and tier 3 represents precursors detected in MS2 scans. All three peak lists were searched against human proteins in SwissProt.

Counts are at 1% FDR and Unique PepSeq is the count of peptide sequences not found in either of the other tiers. For this particular analysis, almost nothing is found in tier 2. An area-proportional Venn diagram (eulerAPE) for counts of distinct peptide sequences at 1% FDR looks like this.

The DIA-Umpire documentation states that "These spectra are written to separate files, because they must be searched separately against a protein database as a consequence of differences in FDR estimates for these varying quality data." Using Mascot, the same search parameters can be used for all three, the score distributions for tier 1 and 3 are very similar, and the significance threshold for 1% FDR is essentially the same for tier 1 and 3, so no particular need for separate searches. If this picture applies generally, you might choose to search only the tier 1 peak list or, for best coverage, merge the tier 1 and 3 peak lists.



How do I combine results from multiple fractions / raw files?

- Multiple fractions from the same sample should generally be searched 'together'
- Mascot Daemon checkbox

Mascot Daemon

Task: Untitled

Parameter set: C:\ProgramData\Matrix ...\default.par

Data import filter: Mascot Distiller

Data file list: Drag and drop data files into the area below or click on Add ...

V:\mascot_test_data\ABRF2009\ah_072808p_
V:\mascot_test_data\ABRF2009\ah_072808p_
V:\mascot_test_data\ABRF2009\ah_072808p_
V:\mascot_test_data\ABRF2009\ah_072808p_

☑ Merge MS/MS files into single search

Schedule:
● Start now
○ Start at 13 April 2015 - 15:12
○ Real-time monitor
○ Follow-up

Search priority: 0

Follow-up: No follow-up required

□ Discard results
□ Repeat at intervals of 1 days
☑ Pass data to: -- None --

MASCOT : How do I do that?

© 2015 Matrix Science

MATRIX SCIENCE

This is quite a common support question.

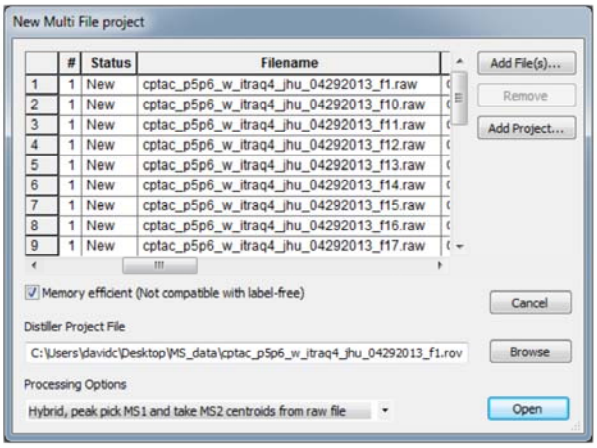
If you are taking fractions from the same sample, we can't think of any reason why you wouldn't want to combine all the data. Peptides from the same protein are likely to be found in different fractions. It's obviously possible to search all the data independently and then use Excel to combine the results, but this would be incredibly tedious.

Using our standard software, there are at least two ways to do this. The first uses Mascot Daemon. It's simply a case of using the check box on the task editor tab.

You have to use "Start now" or "Start at" with a list of files. Obviously can't use "Real time monitor" because it would have no way of knowing when it has all the files available and so when to submit the search.

This works by producing a peak list for all of the files and then concatenating all the peak lists into one big MGF file. One possible problem with this approach is that if the resulting peak list is bigger than 4Gb and if your Mascot Server runs on Windows, then you will hit upload limit size for the IIS Web server on Windows. One option is to use Apache on Windows or Linux.

How do I combine results from multiple fractions / raw files?



New Multi File project

	#	Status	Filename
1	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f1.raw
2	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f10.raw
3	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f11.raw
4	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f12.raw
5	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f13.raw
6	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f14.raw
7	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f15.raw
8	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f16.raw
9	1	New	cptac_p5p6_w_itraq4_jhu_04292013_f17.raw

☑ Memory efficient (Not compatible with label-free)

Distiller Project File

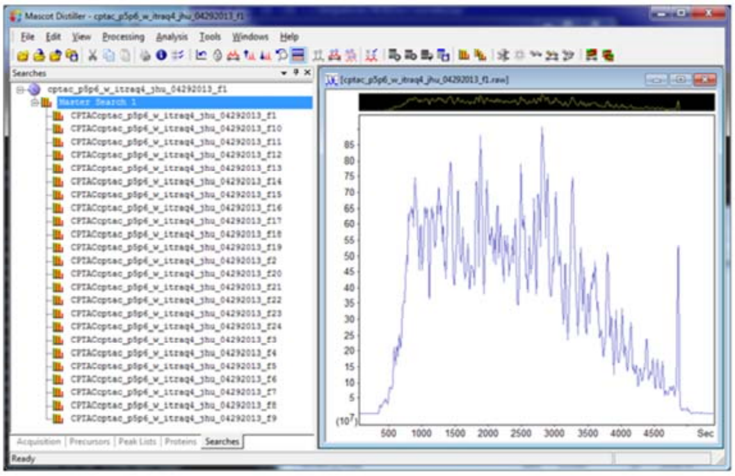
C:\Users\davidc\Desktop\MS_data\cptac_p5p6_w_itraq4_jhu_04292013_f1.rov

Processing Options

Hybrid, peak pick MS1 and take MS2 centroids from raw file

The second option is to use Mascot Distiller. In Distiller, choose a new Multi File project, and then select the required files. If you deselect the “Memory efficient” checkbox, then all the peak lists will be concatenated together in a similar way to using Mascot Daemon. If you leave it checked, and then “Process and Search”, a search will be run for each raw data file and you won’t have any of the limitations with files more than 4Gb.

How do I combine results from multiple fractions / raw files?


Mascot Distiller - cptac_p5p6_w_itraq4_jhu_04292013_f1
File Edit View Processing Analysis Tools Windows Help
Searches
cptac_p5p6_w_itraq4_jhu_04292013_f1
Master Search 1
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f1
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f10
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f11
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f12
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f13
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f14
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f15
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f16
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f17
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f18
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f19
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f2
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f20
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f21
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f22
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f23
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f24
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f3
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f4
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f5
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f6
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f7
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f8
CPTACcptac_p5p6_w_itraq4_jhu_04292013_f9
Acquisition | Precursors | Peak Lists | Proteins | Searches
Ready
[cptac_p5p6_w_itraq4_jhu_04292013_f1.raw]
85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5
(10^7)
500 1000 1500 2000 2500 3000 3500 4000 4500 Sec

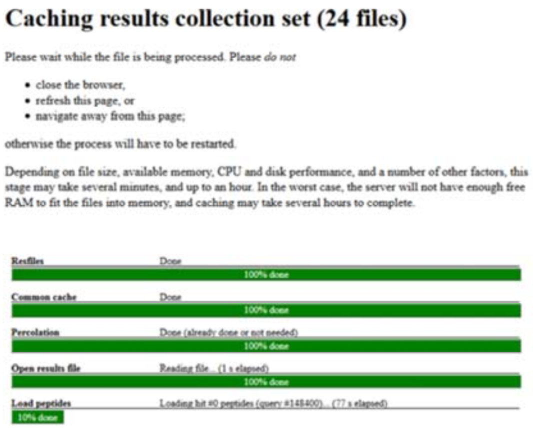
MASCOT : How do I do that?

© 2015 Matrix Science

MATRIX SCIENCE

After all the searches have completed, Distiller will combine the results files and the proteins tab will show the result of the combined searches. On the Searches tab, you can view the individual search results for each fraction, or you can click on the "Master Search" and view the combined search in a browser:

How do I combine results from multiple fractions / raw files?



Caching results collection set (24 files)

Please wait while the file is being processed. Please *do not*

- close the browser,
- refresh this page, or
- navigate away from this page;

otherwise the process will have to be restarted.

Depending on file size, available memory, CPU and disk performance, and a number of other factors, this stage may take several minutes, and up to an hour. In the worst case, the server will not have enough free RAM to fit the files into memory, and caching may take several hours to complete.

Resfiles Done
100% done

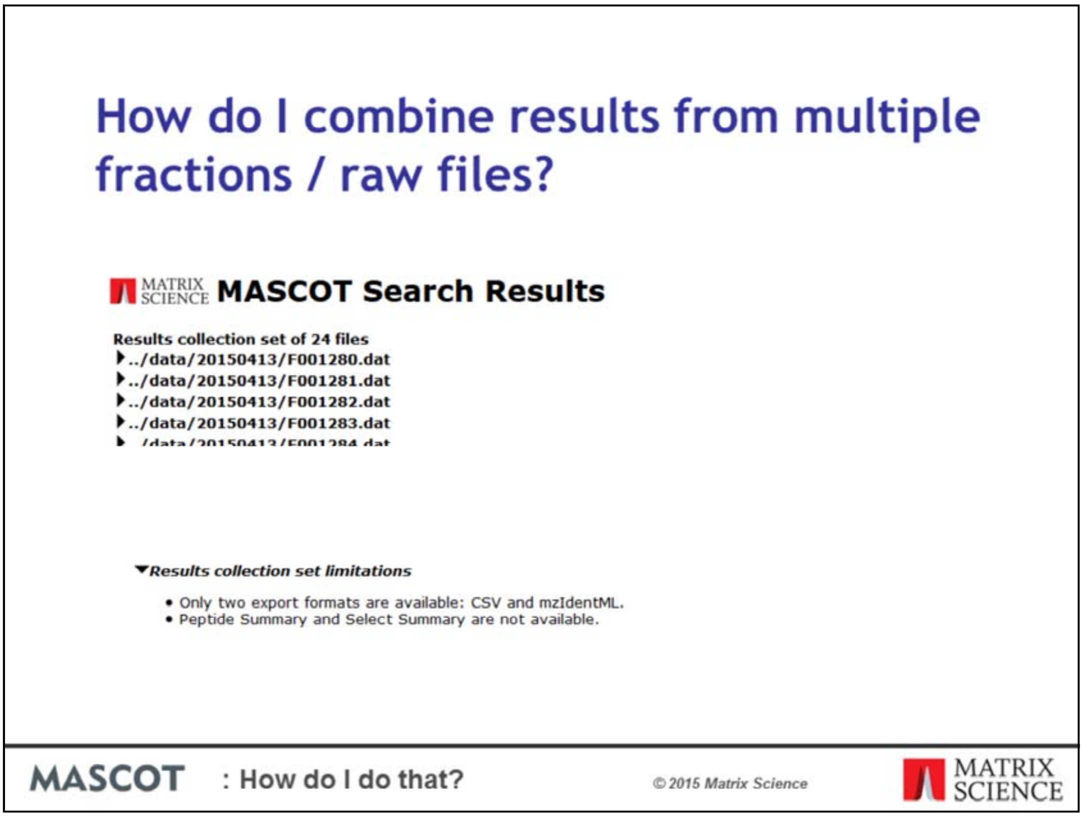
Common cache Done
100% done

Percolation Done (already done or not needed)
100% done

Open results file Reading file... (1 s elapsed)
100% done

Load peptides Loading hit #0 peptides (query #148400)... (77 s elapsed)
10% done

Creating a cache file for the results collection may take a little while, but this only needs to be done once.



How do I combine results from multiple fractions / raw files?

MASCOT Search Results

Results collection set of 24 files

```
▸../data/20150413/F001280.dat
▸../data/20150413/F001281.dat
▸../data/20150413/F001282.dat
▸../data/20150413/F001283.dat
▸../data/20150413/F001284.dat
```

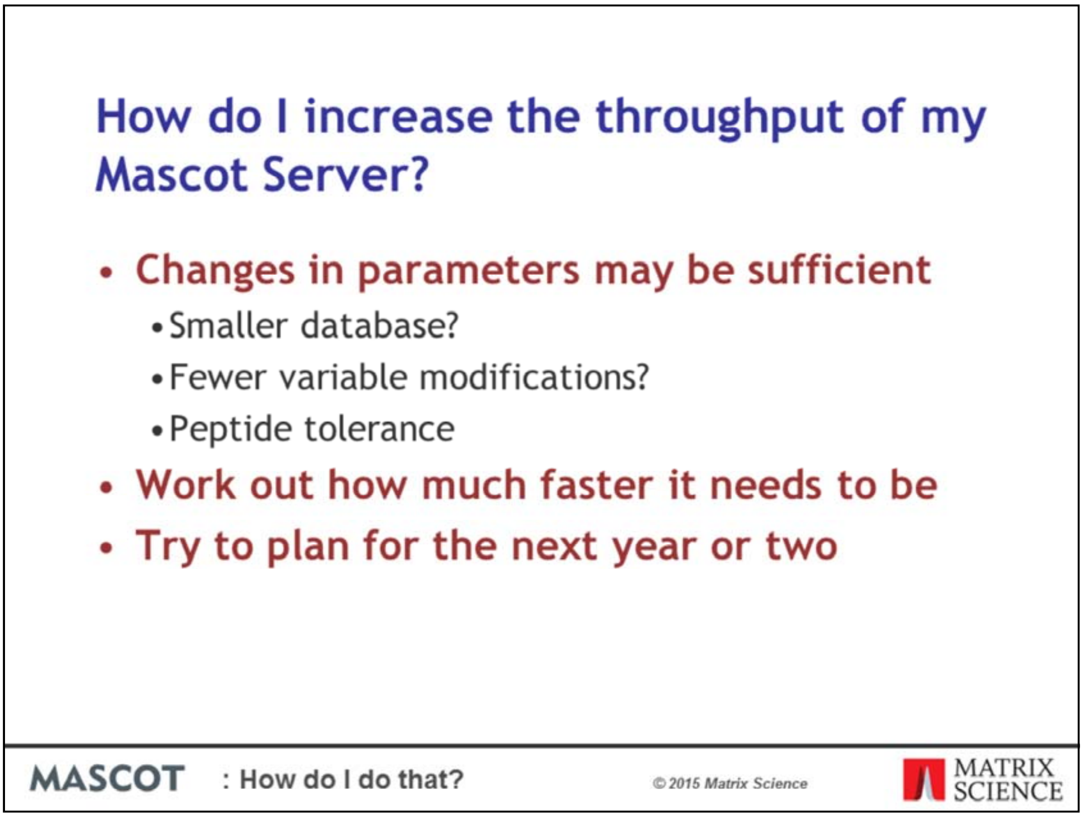
▾***Results collection set limitations***

- Only two export formats are available: CSV and mzIdentML.
- Peptide Summary and Select Summary are not available.

MASCOT : How do I do that? © 2015 Matrix Science

You then end up with a report that is the same as if the MGF files had all been concatenated.

The only limitations are that not all export formats are available, and you have to use the Protein Family Summary, although it wouldn’t be advisable to use either of the other summary reports for a large data set anyway.



How do I increase the throughput of my Mascot Server?

- Changes in parameters may be sufficient
  - Smaller database?
  - Fewer variable modifications?
  - Peptide tolerance
- Work out how much faster it needs to be
- Try to plan for the next year or two

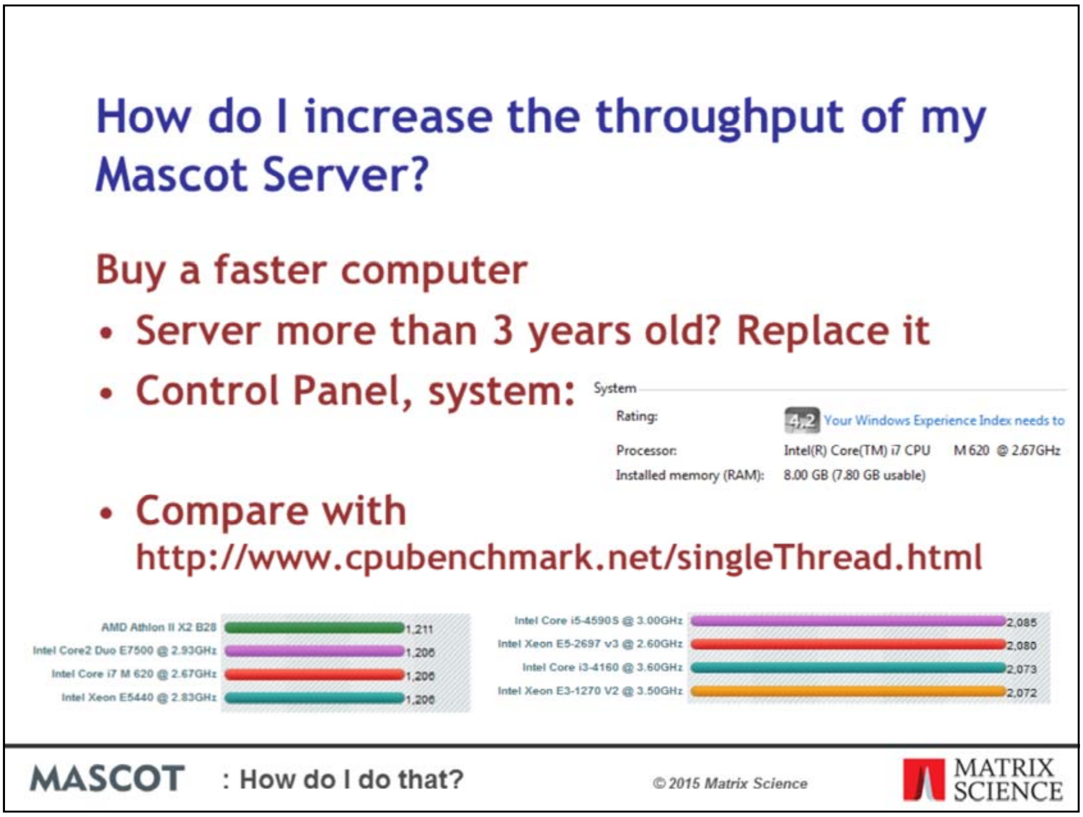
MASCOT : How do I do that?

MATRIX SCIENCE

As instruments get faster, and databases get larger, your Mascot searches will obviously take longer and eventually you will need to deal with this. It may well be that you can just change some of the search parameters. For example, you may be able to use a different database, or restrict searches by taxonomy. Or, your standard protocol may include modifications that you rarely see and you could consider dropping these. It's also worth checking that the peptide tolerance you are using is appropriate. The peptide tolerance affects the search speed, but reducing the fragment tolerance will have negligible effect.

Assuming it's still not fast enough, you need to calculate, or at least estimate how much faster you need the searches to be. Obviously, 'A lot faster ' is not very scientific!

It goes without saying that it's worth trying to plan for the next year or two, although this can be very hard to judge.



How do I increase the throughput of my Mascot Server?

Buy a faster computer

- Server more than 3 years old? Replace it
- Control Panel, system:

System

Rating: 4,2 Your Windows Experience Index needs to

Processor: Intel(R) Core(TM) i7 CPU M 620 @ 2.67GHz

Installed memory (RAM): 8.00 GB (7.80 GB usable)

- Compare with http://www.cpubenchmark.net/singleThread.html

AMD Athlon II X2 B28 1,211
Intel Core2 Duo E7500 @ 2.93GHz 1,206
Intel Core i7 M 620 @ 2.67GHz 1,206
Intel Xeon E5440 @ 2.83GHz 1,206
Intel Core i5-4590S @ 3.00GHz 2,085
Intel Xeon E5-2697 v3 @ 2.60GHz 2,080
Intel Core i3-4160 @ 3.60GHz 2,073
Intel Xeon E3-1270 V2 @ 3.50GHz 2,072

MASCOT : How do I do that?

© 2015 Matrix Science

MATRIX SCIENCE

The first thing to do is to get some details about your current server hardware. If you are anything like me, there's a good chance you can't remember how old your system is. I thought my laptop was about 2 years old, but looked it up to find that's it coming up to 5 years old. If it's 3 or more years old, it's almost certainly worth replacing it.

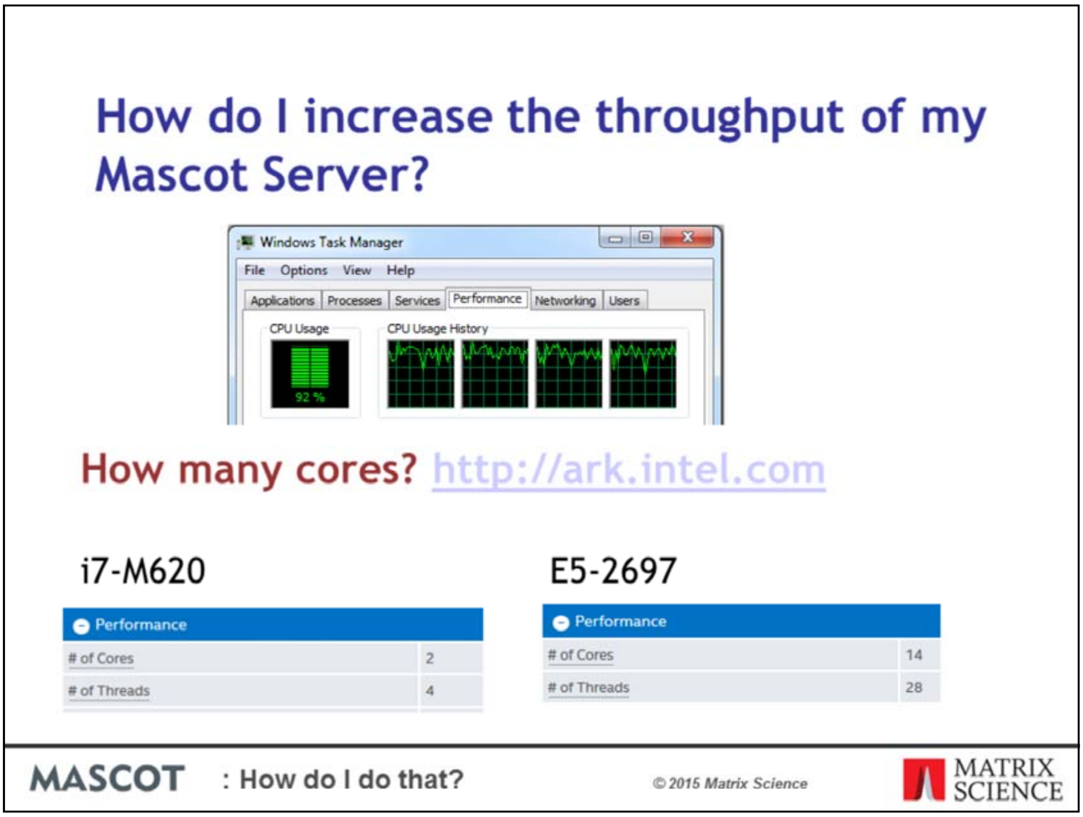
Regardless of age, it's worth checking what processor(s) you have. In Windows, from the Control Panel, System you should be able to find the model and speed.

In Linux, you can: cat /proc/cpuinfo

You can see that my laptop has an M620 running at 2.67GHz.

Lookup the value on the cpubenchmark web site to get a Single Thread Rating. For my laptop, I found this information:

My colleagues at Matrix Science told me I was daft to run Mascot Server on a laptop, so I looked on the Lenovo website for a workstation and found one with the option of either 1 or 2 E5-2697 processors. I looked this up and it has a Single Thread Rating of nearly double that of my laptop, so I would expect Mascot to run twice as fast.



How do I increase the throughput of my Mascot Server?

How many cores? http://ark.intel.com

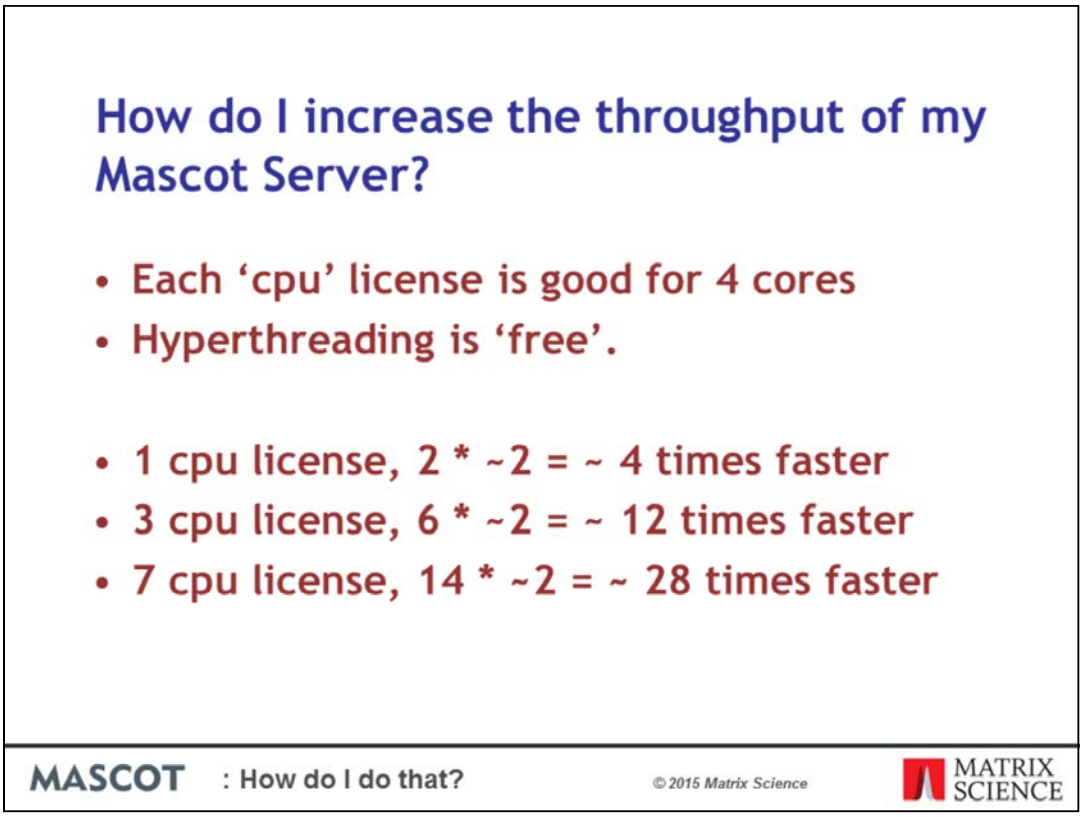
i7-M620

Performance	
# of Cores	2
# of Threads	4

E5-2697

Performance	
# of Cores	14
# of Threads	28

It’s not just about the speed of a single core, it’s the number of cores. The place not to look, is Windows Task manager. From here, it looks like I have 4 cores. For Intel processors, definitive information can be found on ark.intel.com, and if I look for my processor, it says there are just 2 cores, but 4 threads. The system that I’m looking at buying has 14 cores and 28 threads per processor.



How do I increase the throughput of my Mascot Server?

- Each ‘cpu’ license is good for 4 cores
- Hyperthreading is ‘free’.

- 1 cpu license, 2 * ~2 = ~ 4 times faster
- 3 cpu license, 6 * ~2 = ~ 12 times faster
- 7 cpu license, 14 * ~2 = ~ 28 times faster

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

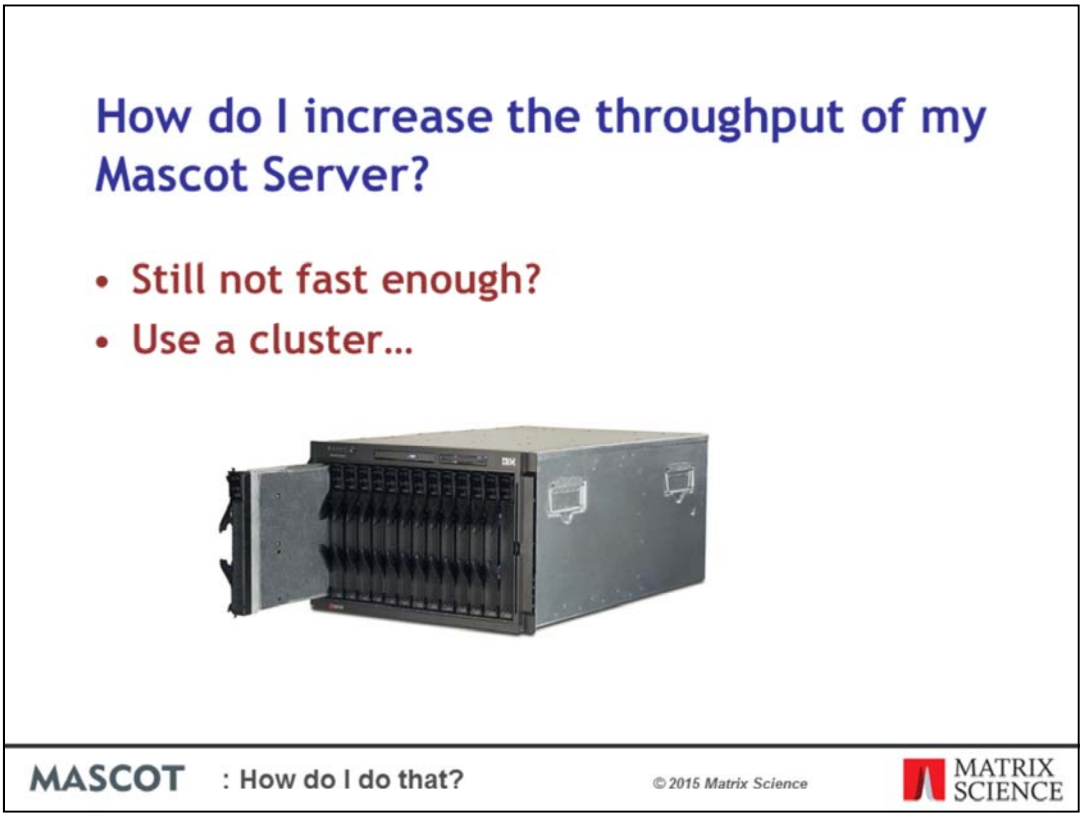
We license by the ‘cpu’ and each license is good for 4 cores. If your processors have hyperthreading, we don’t count these pseudo cores as they don’t improve performance much.

The ~2 in each case is from the single thread benchmark where we saw that the new processors had a value of 2080 compared with my system which has a score of just 1026.

I’m not making full use of my 1 cpu license when I’m running Mascot Server on my laptop, because it only has 2 cores. If I run Mascot on this new server, it will use 4 cores, so the searches will run a total of about 4 times faster.

If that’s not good enough, I could upgrade my license to a 3 cpu license and run searches 12 times faster.

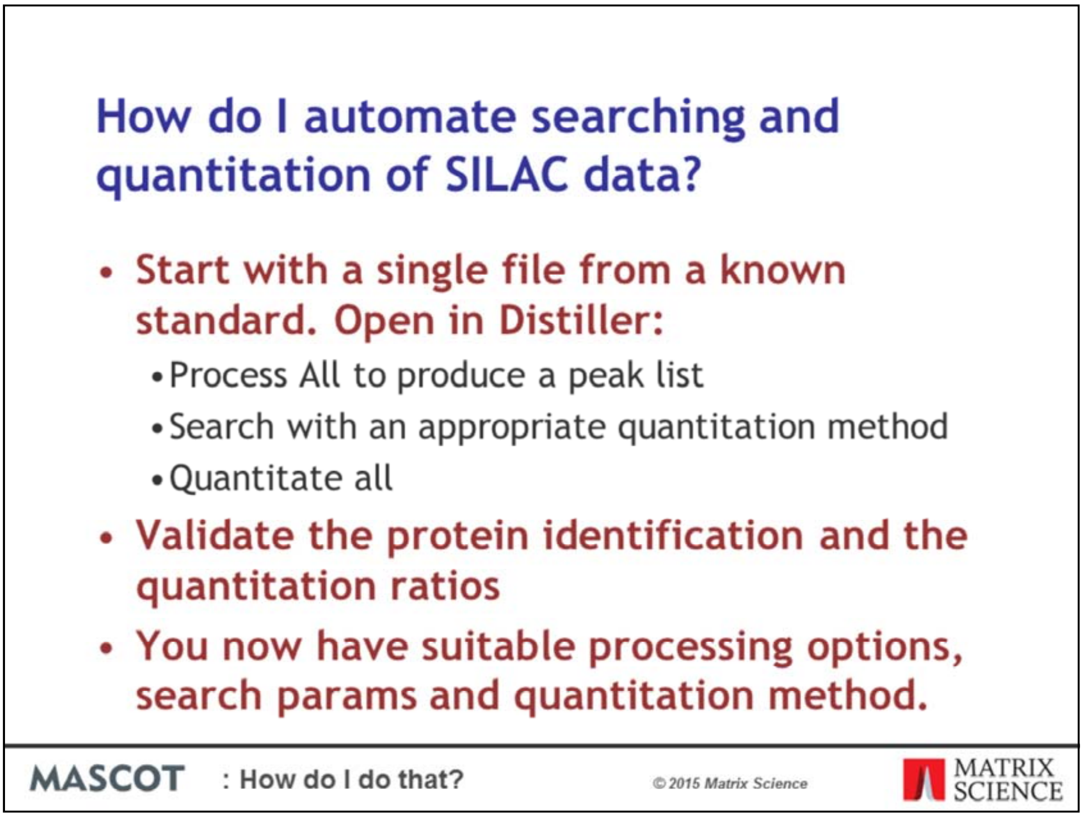
The maximum performance improvement I could get using this new server is with a 7 cpu license, where I would use all 28 cores.



How do I increase the throughput of my Mascot Server?

- Still not fast enough?
- Use a cluster…

If that’s still not fast enough, you’ll probably need to consider setting up a cluster. This can be as simple as just adding another couple of PCs connected by a LAN, or you could purchase a blade center.



How do I automate searching and quantitation of SILAC data?

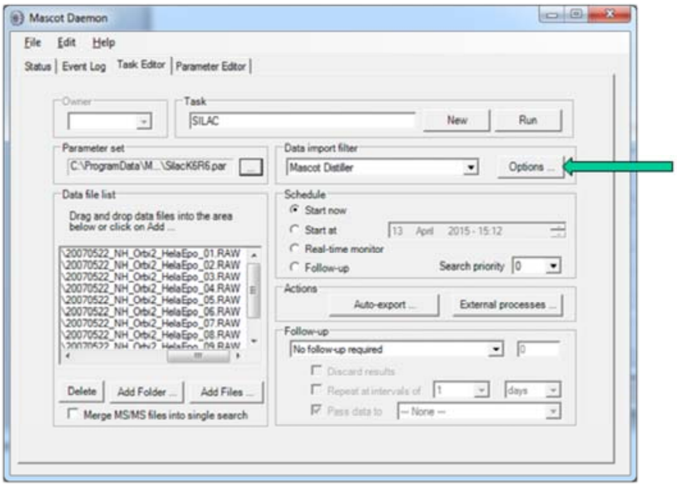
- Start with a single file from a known standard. Open in Distiller:
  - Process All to produce a peak list
  - Search with an appropriate quantitation method
  - Quantitate all
- Validate the protein identification and the quantitation ratios
- You now have suitable processing options, search params and quantitation method.

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

This may sound like I'm teaching grandmother to suck eggs, but it's surprising that some people jump straight into using a complex technique without testing with a standard sample. We also recommend that you start with a single file – if it's a simple standard, you should be able to use a single fraction. Open the file in the Distiller application and choose suitable processing options to produce the peak list. When you first open the file, you should be presented with suitable options from the list. If there's nothing suitable for your instrument, we will happily help create a new processing options file.

Next, submit the search from Distiller choosing the appropriate quantitation method. You must select a quantitation method, so don't just chose the relevant modifications. Once the search is complete, chose "Quantitate All" from the menu. Since you know what the results should be, validate the protein identification and the quantitation results. You may need to modify these. If you have trouble at this stage, we may ask you to send the .rov file and the raw file and we can help.

How do I automate searching and quantitation of SILAC data?



Owner:

Task: SILAC

Parameter set: C:\ProgramData\M...\SilacK6R6.par

Data import filter: Mascot Distiller

Data file list

Drag and drop data files into the area below or click on Add ...

```
\20070522_NH_Orbi2_HelaEpo_01.RAW
\20070522_NH_Orbi2_HelaEpo_02.RAW
\20070522_NH_Orbi2_HelaEpo_03.RAW
\20070522_NH_Orbi2_HelaEpo_04.RAW
\20070522_NH_Orbi2_HelaEpo_05.RAW
\20070522_NH_Orbi2_HelaEpo_06.RAW
\20070522_NH_Orbi2_HelaEpo_07.RAW
\20070522_NH_Orbi2_HelaEpo_08.RAW
\20070522_NH_Orbi2_HelaEpo_09.RAW
```

□ Merge MS/MS files into single search

Schedule

● Start now

○ Start at 13 April 2015 - 15:12

○ Real-time monitor

○ Follow-up

Search priority: 0

Actions

Follow-up: No follow-up required 0

□ Discard results

□ Repeat at intervals of 1 days

☑ Pass data to -- None --

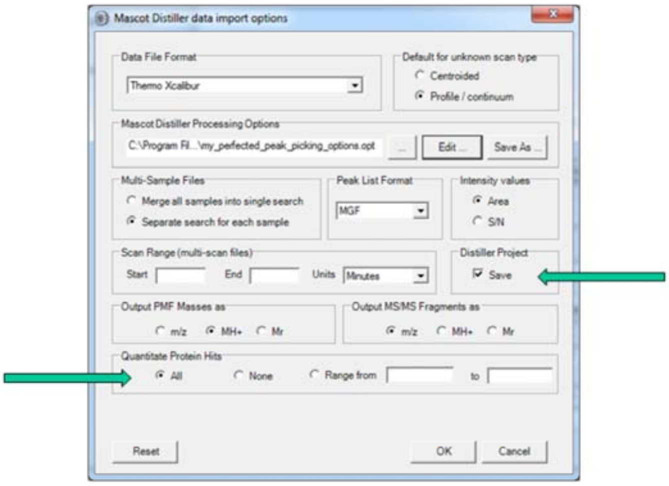
Once you have all your parameters nailed down, you can use Mascot Daemon to automate the process for many files. I’ve assumed here that you are now running multiple fractions.

You can either choose a real time monitor task, or you can use a “Start now” task and select the files.

Using the parameter editor tab, create a parameter set for the search parameters, and then make sure you select it here.

Use the Mascot Distiller import filter, and then choose the options:

How do I automate searching and quantitation of SILAC data?



Mascot Distiller data import options

Data File Format: Thermo Xcalibur

Default for unknown scan type: ○ Centroided ● Profile / continuum

Mascot Distiller Processing Options: C:\Program Fil...\my_perfected_peak_picking_options.opt

Multi-Sample Files: ○ Merge all samples into single search ● Separate search for each sample

Peak List Format: MGF

Intensity values: ● Area ○ S/N

Scan Range (multi-scan files): Start End Units Minutes

Distiller Project: ☑ Save

Output PMF Masses as: ○ m/z ● MH+ ○ Mr

Output MS/MS Fragments as: ● m/z ○ MH+ ○ Mr

Quantitate Protein Hits: ● All ○ None ○ Range from to

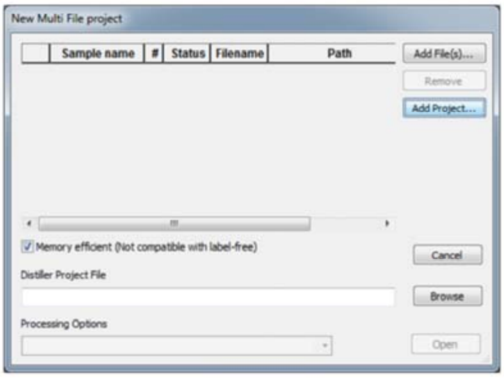
For the Distiller import options, choose your options file that you prepared earlier when using Distiller.

Make sure that you save the Distiller Project

Select Quantitate All Protein hits.

How do I automate searching and quantitation of SILAC data?

Once it's all done, you need to combine the results in Mascot Distiller:



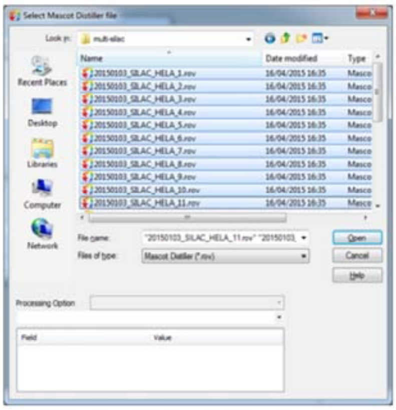
New Multi File project

	Sample name	#	Status	Filename	Path

☑ Memory efficient (Not compatible with label-free)

Distiller Project File

Processing Options



Select Mascot Distiller file

Look in: multi-silac

Name	Date modified	Type
20150103_SILAC_HELA_1.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_2.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_3.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_4.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_5.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_6.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_7.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_8.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_9.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_10.rov	16/04/2015 16:35	Masco
20150103_SILAC_HELA_11.rov	16/04/2015 16:35	Masco

File name: "20150103_SILAC_HELA_11.rov" "20150103_

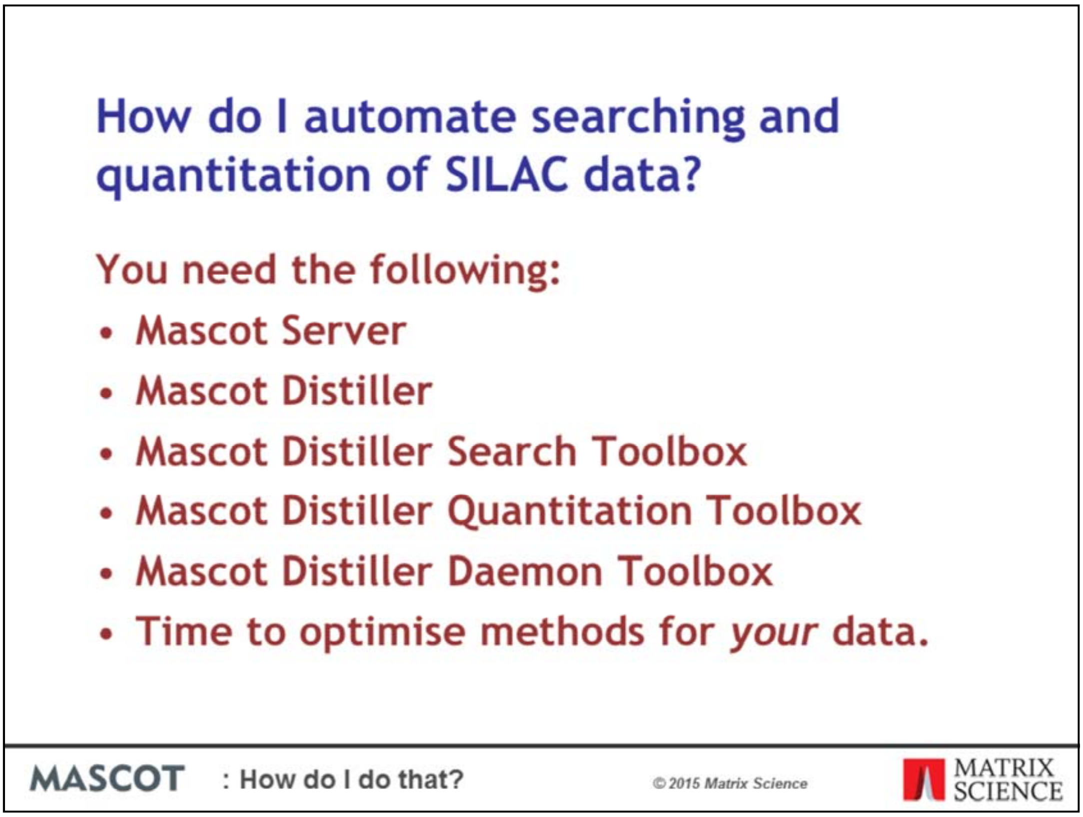
Files of type: Mascot Distiller (*.rov)

Processing Option

Field	Value

If these are fractions from a single sample, you'll want to combine them. Choose "New Multi File Project" from the file menu and then "Add Project". Select all the relevant Mascot Distiller Project files.

Next, from the Analysis menu, choose to Quantitate all hits, and the results will be produced pretty quickly.



How do I automate searching and quantitation of SILAC data?

You need the following:

- Mascot Server
- Mascot Distiller
- Mascot Distiller Search Toolbox
- Mascot Distiller Quantitation Toolbox
- Mascot Distiller Daemon Toolbox
- Time to optimise methods for *your* data.

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

You need the following:

Mascot Server

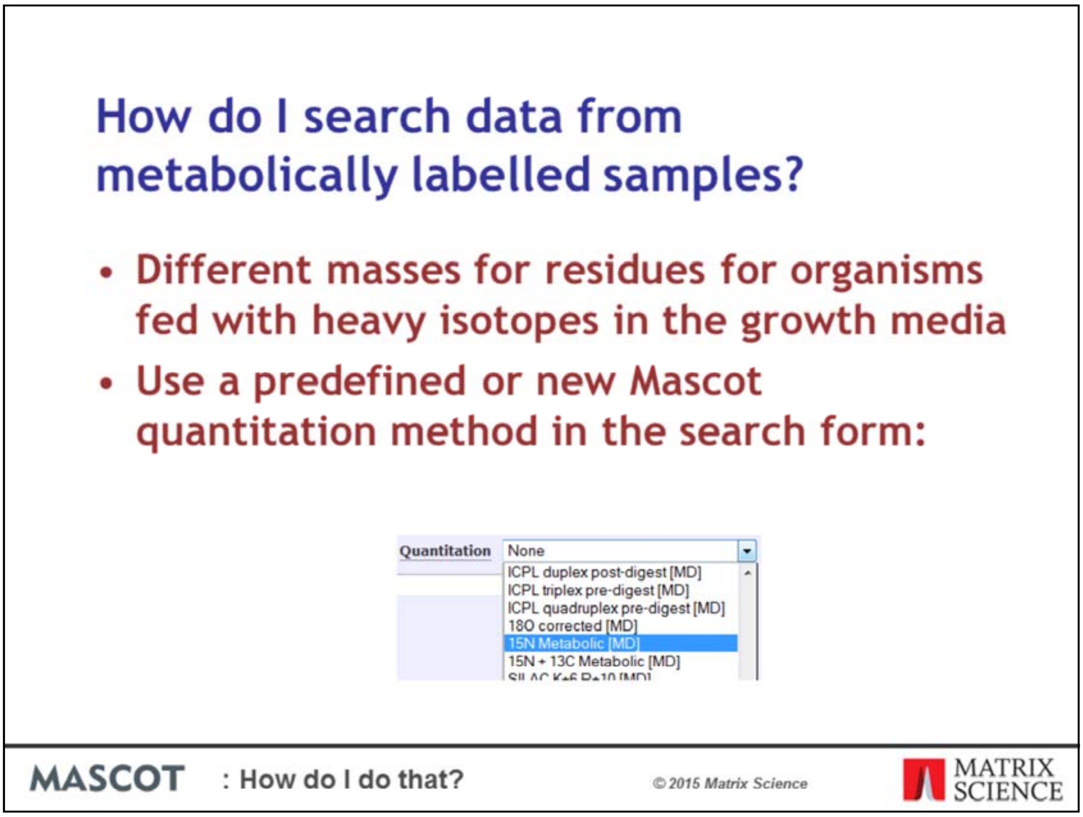
Mascot Distiller

Mascot Distiller Search Toolbox

Mascot Distiller Quantitation Toolbox

Mascot Distiller Daemon Toolbox

Time to optimise methods for *your* data.



How do I search data from metabolically labelled samples?

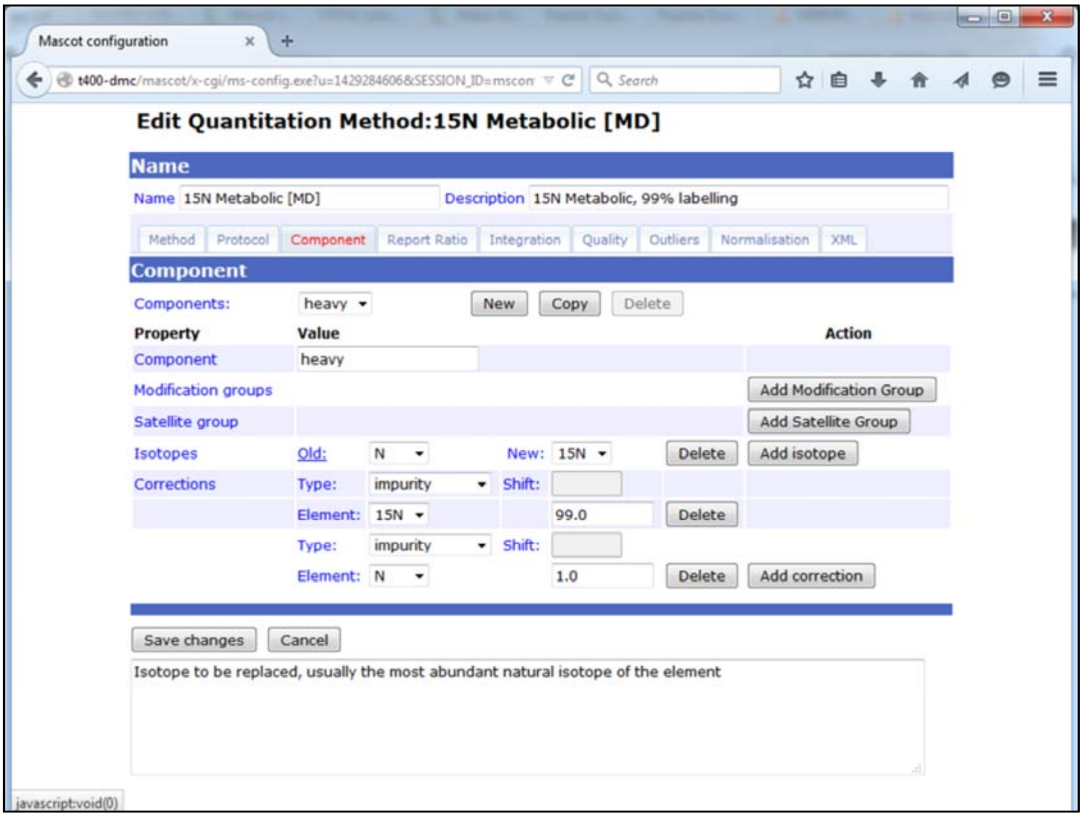
- Different masses for residues for organisms fed with heavy isotopes in the growth media
- Use a predefined or new Mascot quantitation method in the search form:

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

A fairly common question we get, is how to search data from metabolically labelled samples, for example 15N data. The problem is that all of the residues from the organisms grown with heavy isotopes will have a different mass. You could of course re-define all the residue masses in Mascot Server, but a much better way is to use a quantitation method, even if you aren't going to use Mascot Distiller to quantify the results.

The names of the quantitation methods should be obvious, and for a Metabolic labelled sample, you will notice that the name has the [MD] at the end, indicating that the quantitation would need to be done in Mascot Distiller.

It is worth looking at the method in the method editor:



Edit Quantitation Method:15N Metabolic [MD]

Name

Name: 15N Metabolic [MD] Description: 15N Metabolic, 99% labelling

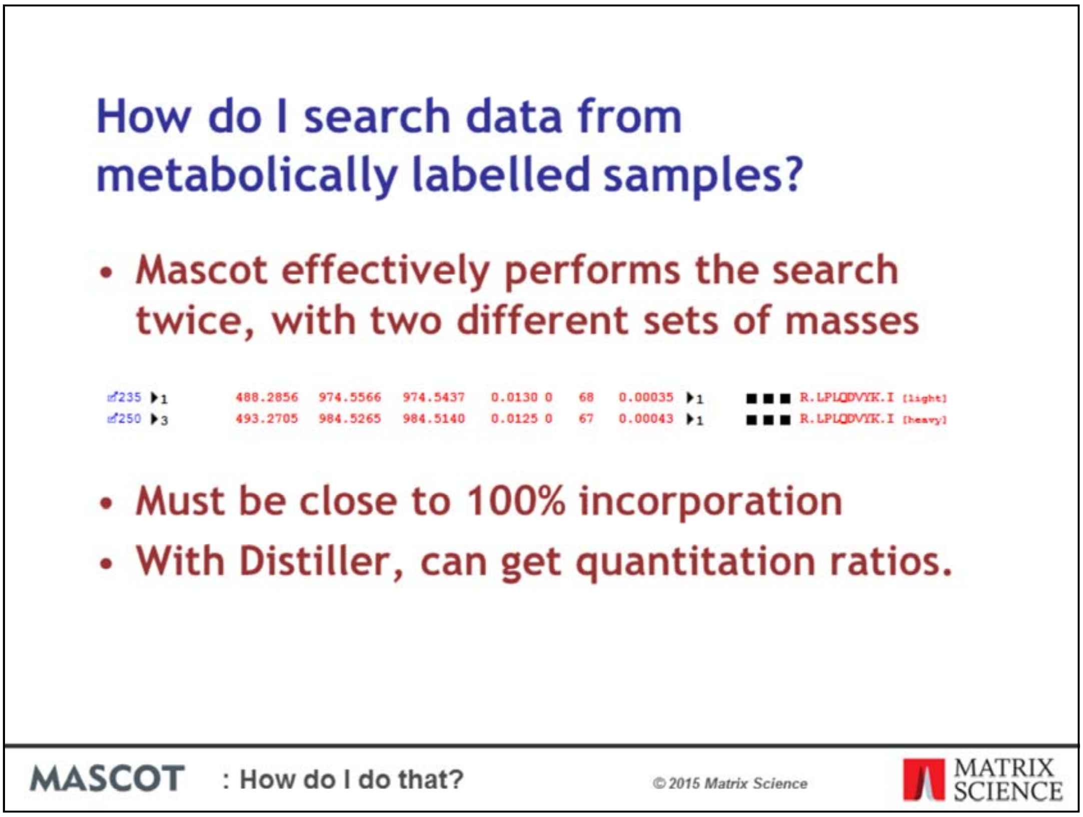
Component

Components: heavy

Property	Value				Action
Component	heavy				
Modification groups					Add Modification Group
Satellite group					Add Satellite Group
Isotopes	Old: N	New: 15N		Delete	Add isotope
Corrections	Type: impurity	Shift:			
	Element: 15N	99.0		Delete	
	Type: impurity	Shift:			
	Element: N	1.0		Delete	Add correction

Isotope to be replaced, usually the most abundant natural isotope of the element

For 15N, you will see that there are two components and the heavy one is defined with the isotope 15N rather than standard N. You will also see that we can define the impurity. Here, we have a 99% labelling efficiency.



How do I search data from metabolically labelled samples?

- Mascot effectively performs the search twice, with two different sets of masses

```
235 ▶1     488.2856  974.5566  974.5437  0.0130 0  68  0.00035 ▶1   ■■■ R.LPLQDVYK.I [light]
250 ▶3     493.2705  984.5265  984.5140  0.0125 0  67  0.00043 ▶1   ■■■ R.LPLQDVYK.I [heavy]
```

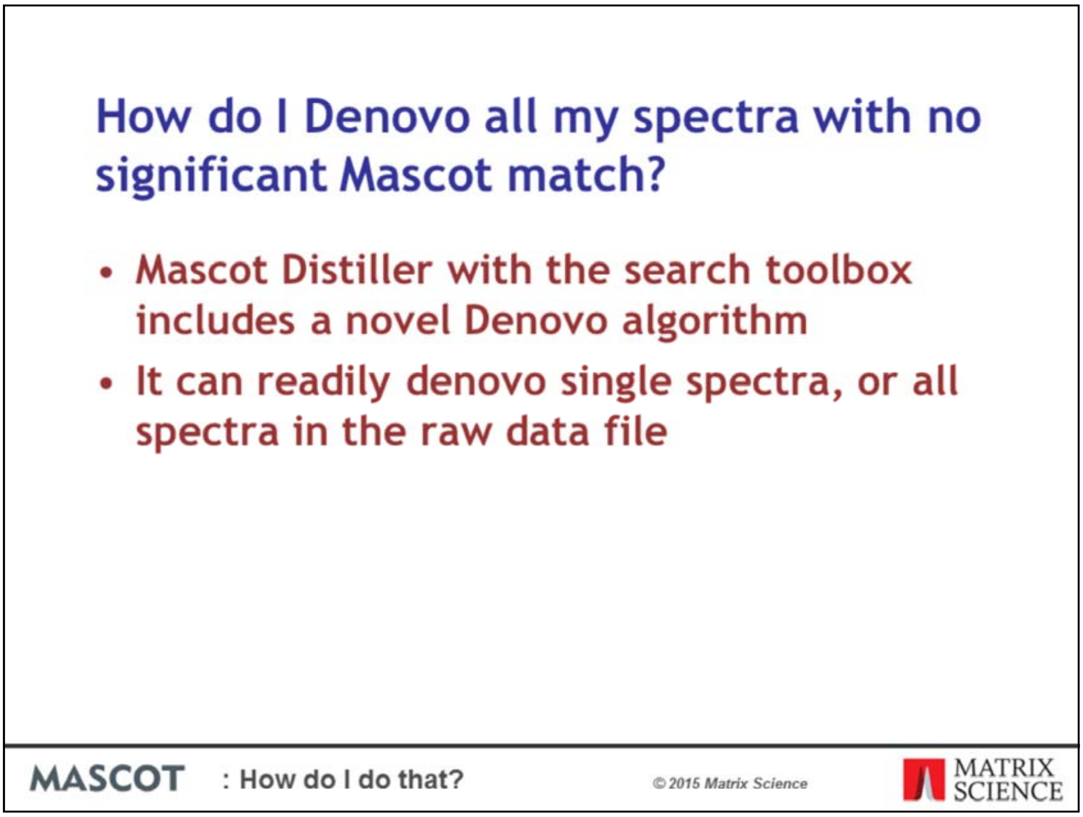
- Must be close to 100% incorporation
- With Distiller, can get quantitation ratios.

MASCOT
: How do I do that?
MATRIX SCIENCE

Mascot effectively performs the search twice, with two different sets of masses. In the search engine, the heavy and light matches are shown using whatever labels you defined in the quantitation editor

The search engine doesn't use the incorporation rates from the quantitation editor – those are just used for the Distiller quantitation. Therefore, with a lower incorporation rate, the masses will be slightly inaccurate. A 98% rate should just about be OK, but you will need to increase the tolerances, possibly to +/- 2 or 3 Da and it's likely that the peak picking won't work well. It's normally worth the extra effort to get the highest incorporation rate possible

And if you load and process the data through Distiller, that will of course calculate quantitation information from the MS1 scans.



How do I Denovo all my spectra with no significant Mascot match?

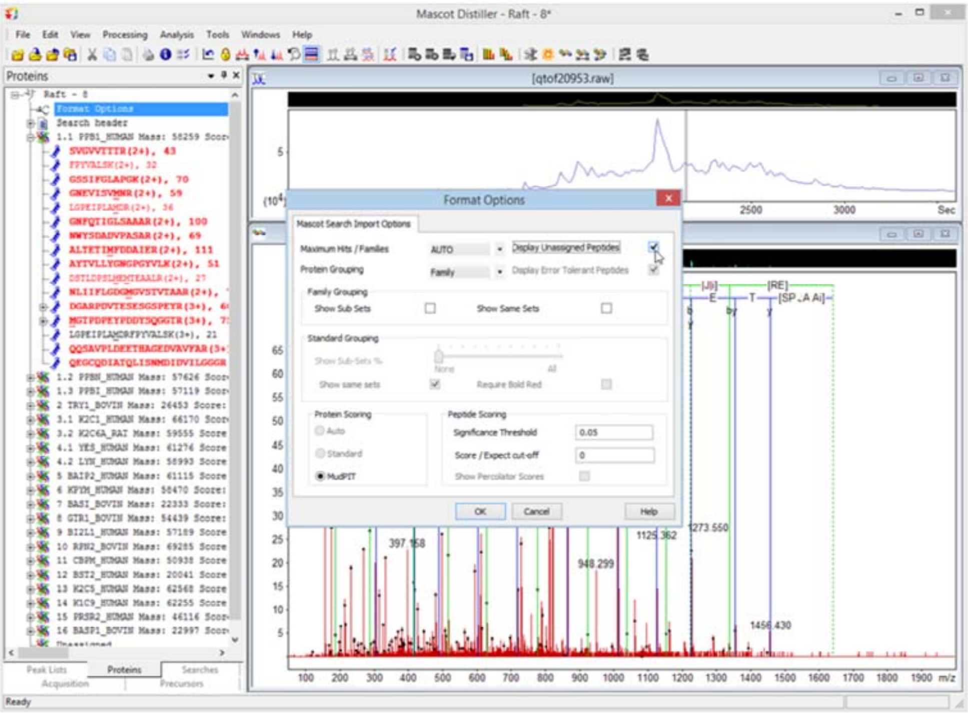
- Mascot Distiller with the search toolbox includes a novel Denovo algorithm
- It can readily denovo single spectra, or all spectra in the raw data file

MASCOT : How do I do that?

MATRIX SCIENCE

Mascot Distiller with the search toolbox includes a novel Denovo algorithm that can be used when you have high quality spectra, but are unable to get a significant match from a database search.

It can readily denovo single spectra, or all spectra in the raw data file, but a common requirement is to be able to search all spectra with no significant match from Mascot Server.



Format Options

Mascot Search Import Options

Maximum Hits / Families: AUTO

Protein Grouping: Family

Display Unassigned Peptides ☑

Display Error Tolerant Peptides ☑

Family Grouping

Show Sub Sets □ Show Same Sets □

Standard Grouping

Show Sub-Sets % None All

Show same sets ☑ Require Bold Red □

Protein Scoring

○ Auto

○ Standard

● MudPIT

Peptide Scoring

Significance Threshold: 0.05

Score / Expect cut-off: 0

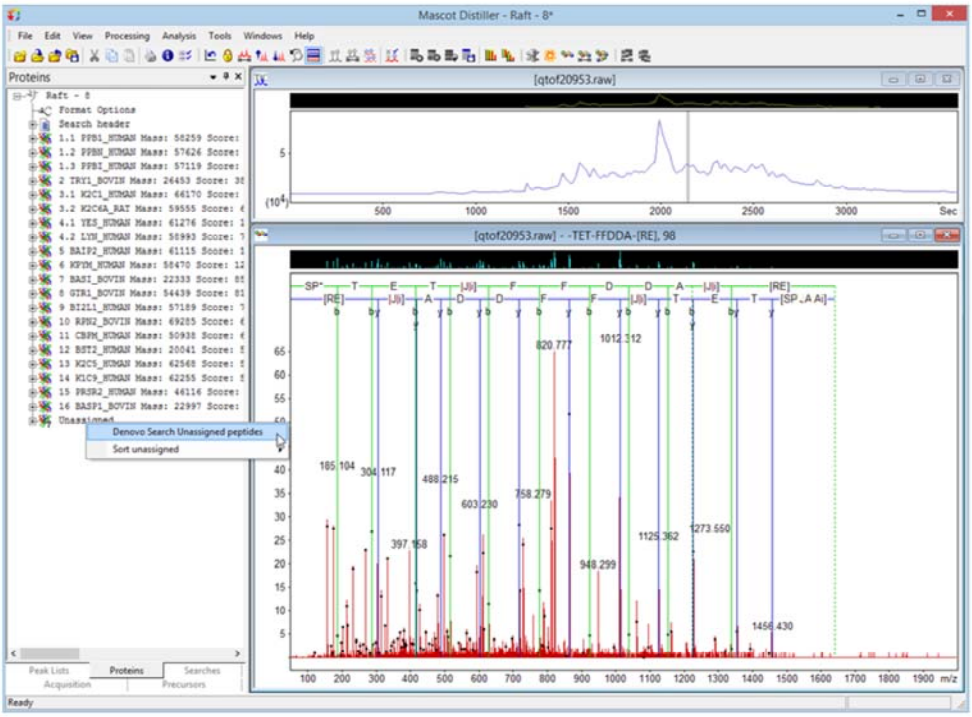
Show Percolator Scores □

OK Cancel Help

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

To de novo sequence a complete peak list collection, or the peak lists in the currently displayed TIC range, use the context menu obtained by right-clicking the root (world) node of the peak lists tree.

The most efficient way to de novo only those spectra that failed to give decent matches in the Mascot search is to switch to the proteins tree, click on Format Options, and choose to load the unassigned queries.



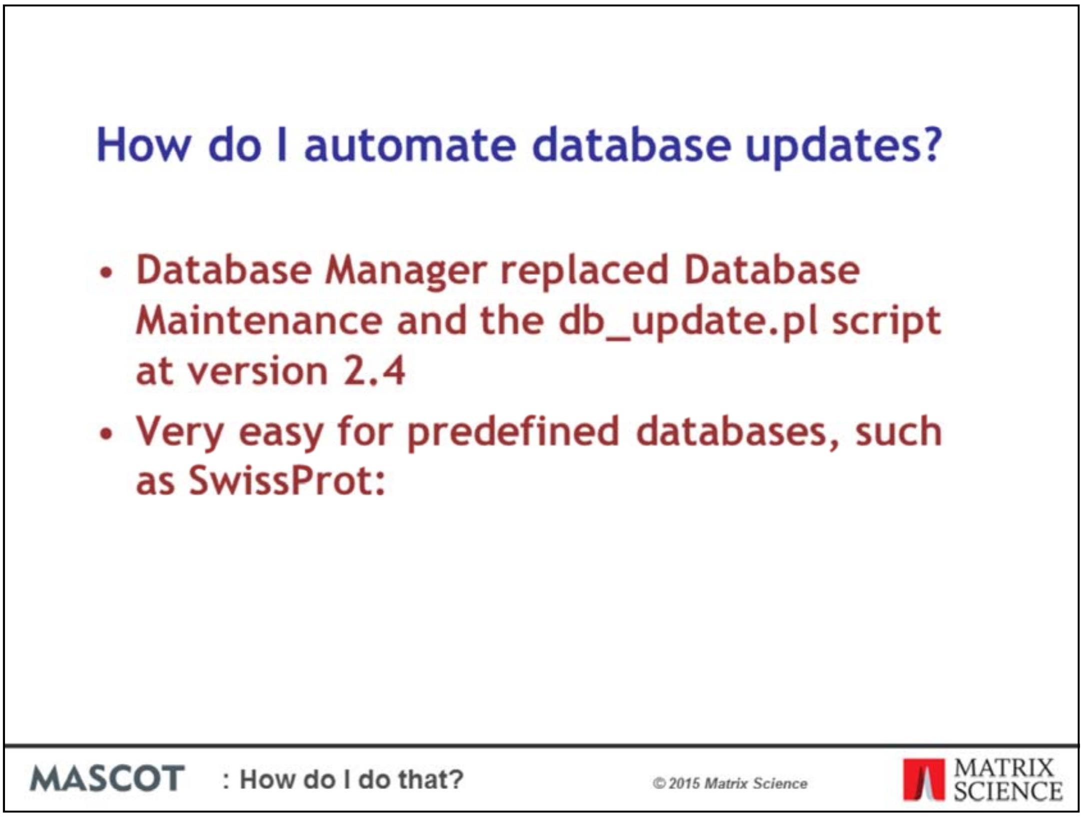
```
Raft - 8
  Format Options
  Search header
  1.1 PPB1_HUMAN Mass: 58259 Score:
  1.2 PPBN_HUMAN Mass: 57626 Score:
  1.3 PPBI_HUMAN Mass: 57119 Score:
  2 TRY1_BOVIN Mass: 26453 Score: 3
  3.1 K2C1_HUMAN Mass: 66170 Score:
  3.2 K2C6A_RAT Mass: 59555 Score: 
  4.1 YES_HUMAN Mass: 61276 Score: 1
  4.2 LYN_HUMAN Mass: 58993 Score: 7
  5 BAIP2_HUMAN Mass: 61115 Score: 1
  6 KPYM_HUMAN Mass: 58470 Score: 12
  7 BASI_BOVIN Mass: 22333 Score: 8
  8 GTR1_BOVIN Mass: 54439 Score: 81
  9 BI2L1_HUMAN Mass: 57189 Score: 7
  10 RPN2_BOVIN Mass: 69285 Score: 
  11 CBPM_HUMAN Mass: 50938 Score: 
  12 BST2_HUMAN Mass: 20041 Score: 5
  13 K2C5_HUMAN Mass: 62568 Score: 5
  14 K1C9_HUMAN Mass: 62255 Score: 5
  15 PRSR2_HUMAN Mass: 46116 Score:
  16 BASP1_BOVIN Mass: 22997 Score:
  Unassigned
```

MASCOT : How do I do that?

© 2015 Matrix Science

MATRIX SCIENCE

Use the context menu obtained by right-clicking the root node or unassigned node to de novo just the unassigned queries.

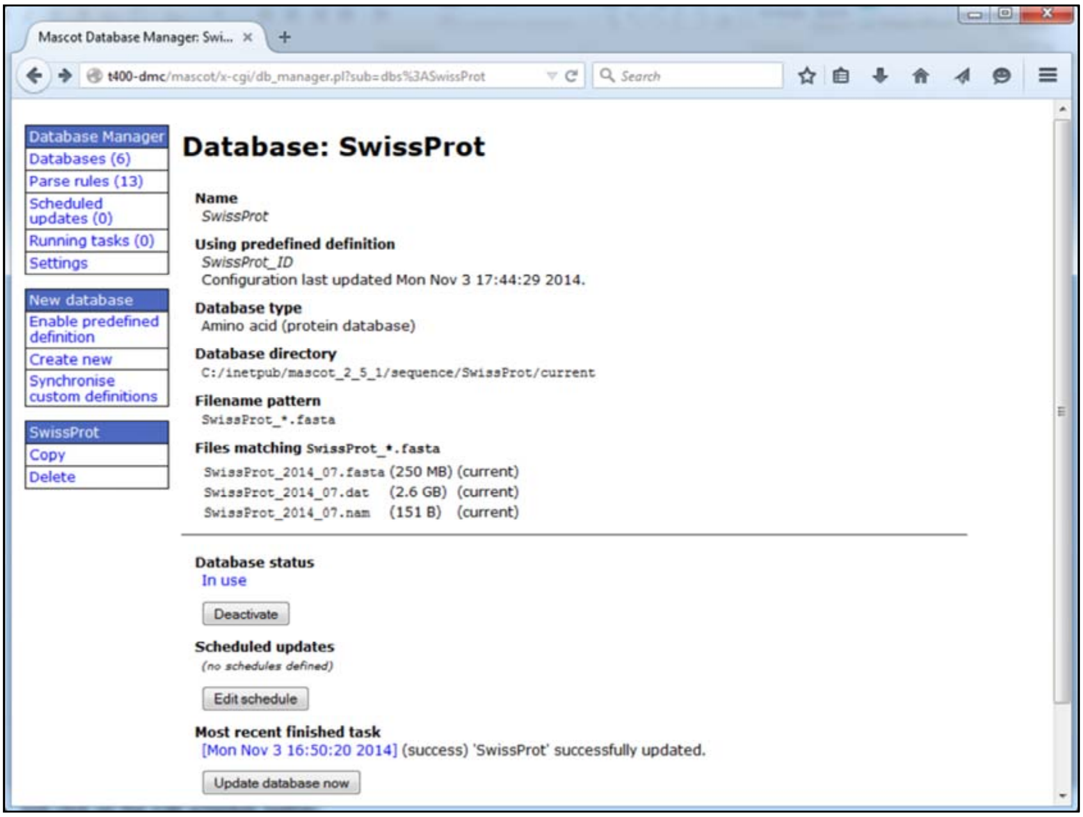


How do I automate database updates?

- Database Manager replaced Database Maintenance and the db_update.pl script at version 2.4
- Very easy for predefined databases, such as SwissProt:

MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

It is normally best to update the databases you are searching every month or two. This is less critical for some species like Human, because there are fewer updates to human proteins in the databases. Of course, you can do this manually, but it’s obviously preferable to automate the process. In versions of Mascot prior to version 2.4, it took a little bit of setting up using the db_update.pl script. For version 2.4 and later, it’s really very easy using Database Manager. I’ll show an example using SwissProt which is one of the predefined databases:



Database: SwissProt

Name
SwissProt

Using predefined definition
SwissProt_ID
Configuration last updated Mon Nov 3 17:44:29 2014.

Database type
Amino acid (protein database)

Database directory
`C:/inetpub/mascot_2_5_1/sequence/SwissProt/current`

Filename pattern
`SwissProt_*.fasta`

Files matching `SwissProt_*.fasta`

```
SwissProt_2014_07.fasta (250 MB) (current)
SwissProt_2014_07.dat   (2.6 GB) (current)
SwissProt_2014_07.nam   (151 B)  (current)
```

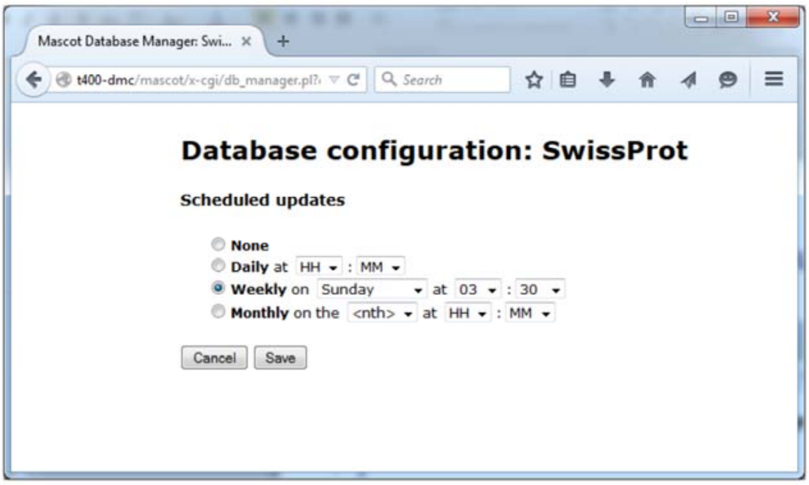
Database status
In use

Scheduled updates
(no schedules defined)

Most recent finished task
[Mon Nov 3 16:50:20 2014] (success) 'SwissProt' successfully updated.

Just click on the Edit schedule button:

How do I automate database updates?


Mascot Database Manager: Swi...
t400-dmc/mascot/x-cgi/db_manager.pl?
Search

Database configuration: SwissProt

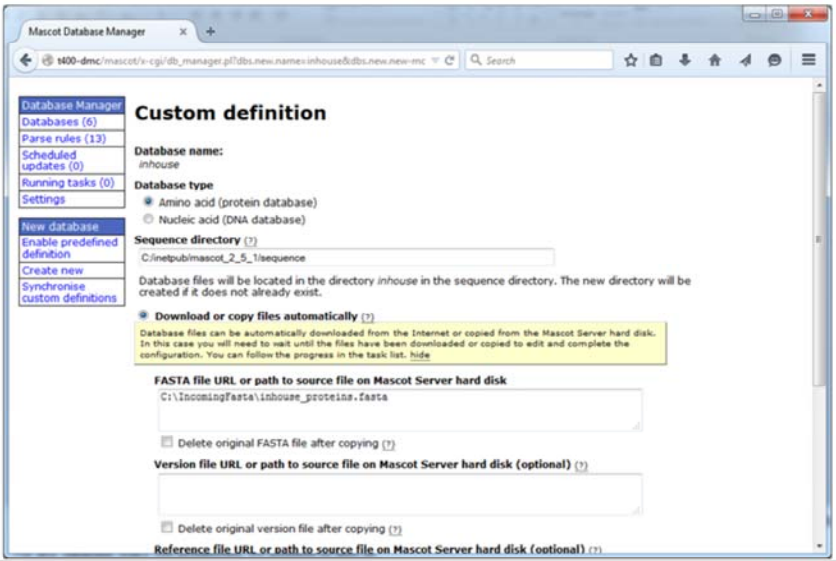
Scheduled updates

○ **None**
○ **Daily** at HH : MM
● **Weekly** on Sunday at 03 : 30
○ **Monthly** on the <nth> at HH : MM

Cancel Save

SwissProt is only updated about once per month but you’ll notice that this has been set for a quiet time at 3:30am on Sunday. Database Manager keeps a record of the time, date and size of the last download, so it only downloads a file if it’s newer than the one on the Mascot Server and therefore it’s not an issue to check for a new version every week.

How do I automate database updates?



Custom definition

Database name:
inhouse

Database type

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

Sequence directory (?)
C:/inetpub/mascot_2_5_1/sequence

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

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

Database files can be automatically downloaded from the Internet or copied from the Mascot Server hard disk. In this case you will need to wait until the files have been downloaded or copied to edit and complete the configuration. You can follow the progress in the task list. hide

FASTA file URL or path to source file on Mascot Server hard disk
C:\IncomingFasta\inhouse_proteins.fasta

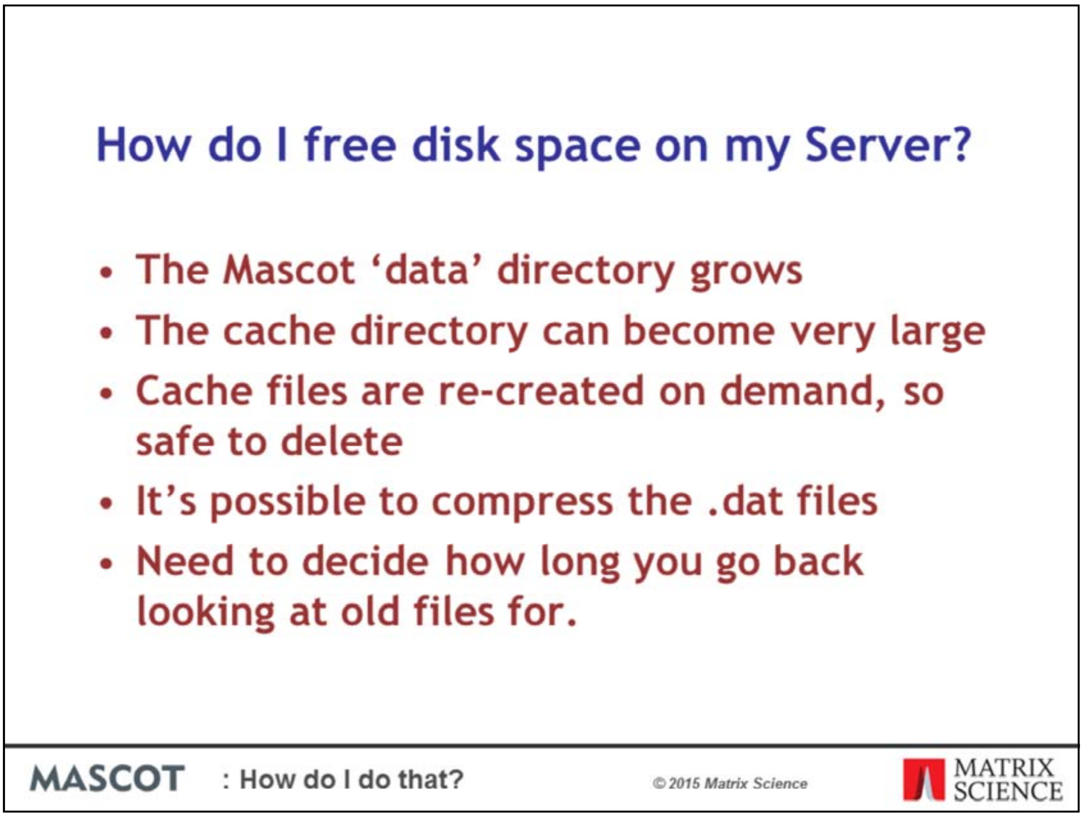
□ Delete original FASTA file after copying (?)

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

□ Delete original version file after copying (?)

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

For any database that’s not predefined, it’s possible to add a path to a file to download from the local hard disk or by using a URL. Once this has been defined, you can set up a schedule in exactly the same way as for SwissProt.



How do I free disk space on my Server?

- The Mascot 'data' directory grows
- The cache directory can become very large
- Cache files are re-created on demand, so safe to delete
- It's possible to compress the .dat files
- Need to decide how long you go back looking at old files for.

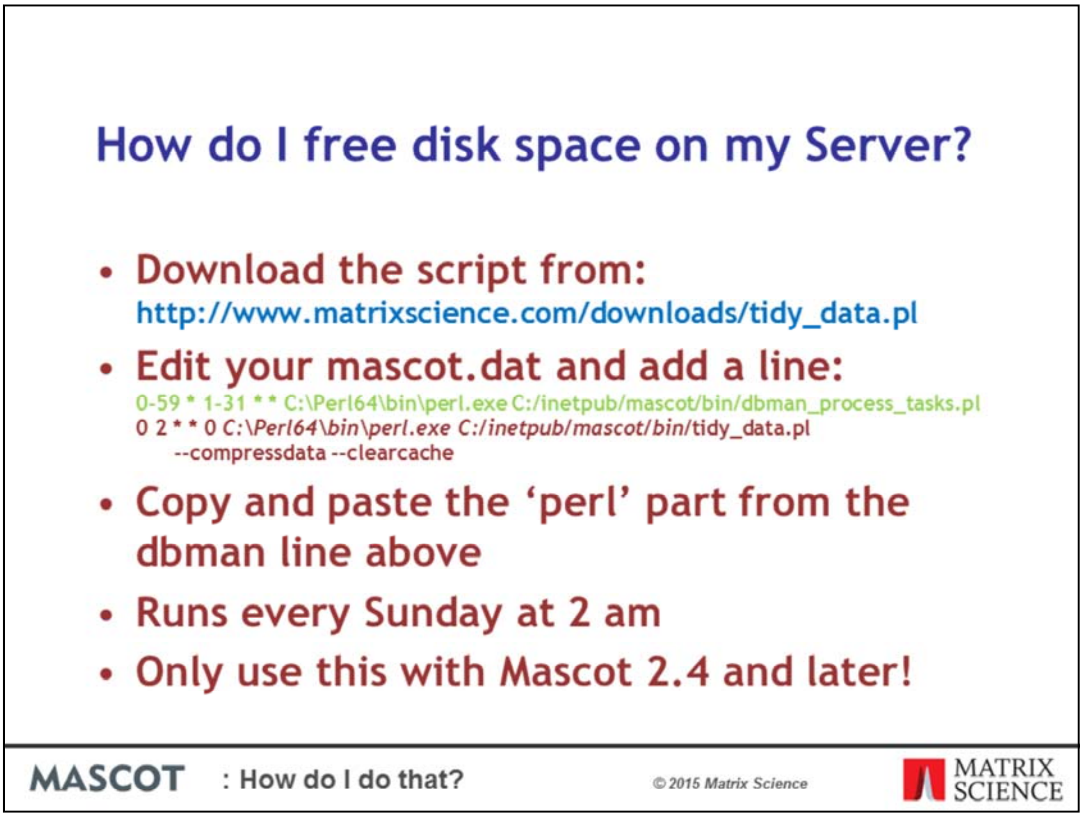
MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

We are often asked what is safe to delete or move on the Mascot Server and many customers note that their data directory has become huge.

Much of the space is taken up by the results files. All of the programs on the Mascot Server that need to read a results file will un-compress a file if necessary. This means that it's a good idea to compress all the old results files. So, if you want to go back and look at an old results file that has been compressed, it will just take a bit longer, but it will still be available.

Mascot uses cache files to speed up report loading and navigation and by default these files are saved in the 'cache' subdirectory under the data directory. If you delete these files, then Mascot will re-create them when you look at a report, so it's totally safe to delete all of them. However, if you routinely go back and look at files for several weeks after searching them, it makes sense to just delete ones that are older than one month.

You can obviously do both of these manually, but it makes more sense to automate this process.



How do I free disk space on my Server?

- Download the script from:
  http://www.matrixscience.com/downloads/tidy_data.pl
- Edit your mascot.dat and add a line:

```
0-59 * 1-31 * * C:\Perl64\bin\perl.exe C:/inetpub/mascot/bin/dbman_process_tasks.pl
0 2 * * 0 C:\Perl64\bin\perl.exe C:/inetpub/mascot/bin/tidy_data.pl
    --compressdata --clearcache
```

- Copy and paste the ‘perl’ part from the dbman line above
- Runs every Sunday at 2 am
- Only use this with Mascot 2.4 and later!

MASCOT : How do I do that?

MATRIX SCIENCE

There’s a script that will do all of this for you, and we’ll probably include it with Mascot 2.6

You can download it from here. Remember that we’ll put these slides on our public web site, so no need to make extensive notes.

Next, you need to edit the bottom of your mascot.dat and you’ll need to do this using a text editor like notepad. You should find a line like this one in green which you shouldn’t alter. Insert a new line below. You can copy and paste the line from this presentation, but you will need to change this part in italics to be the same as the line above if you didn’t install Mascot or Perl in the standard places, or if you are running on Linux.

The “0 2 * * 0” says when the script should run. The parameters are the same as for the Linux cron utility and are described in chapter 6 of the Mascot Installation and Setup manual. However, the

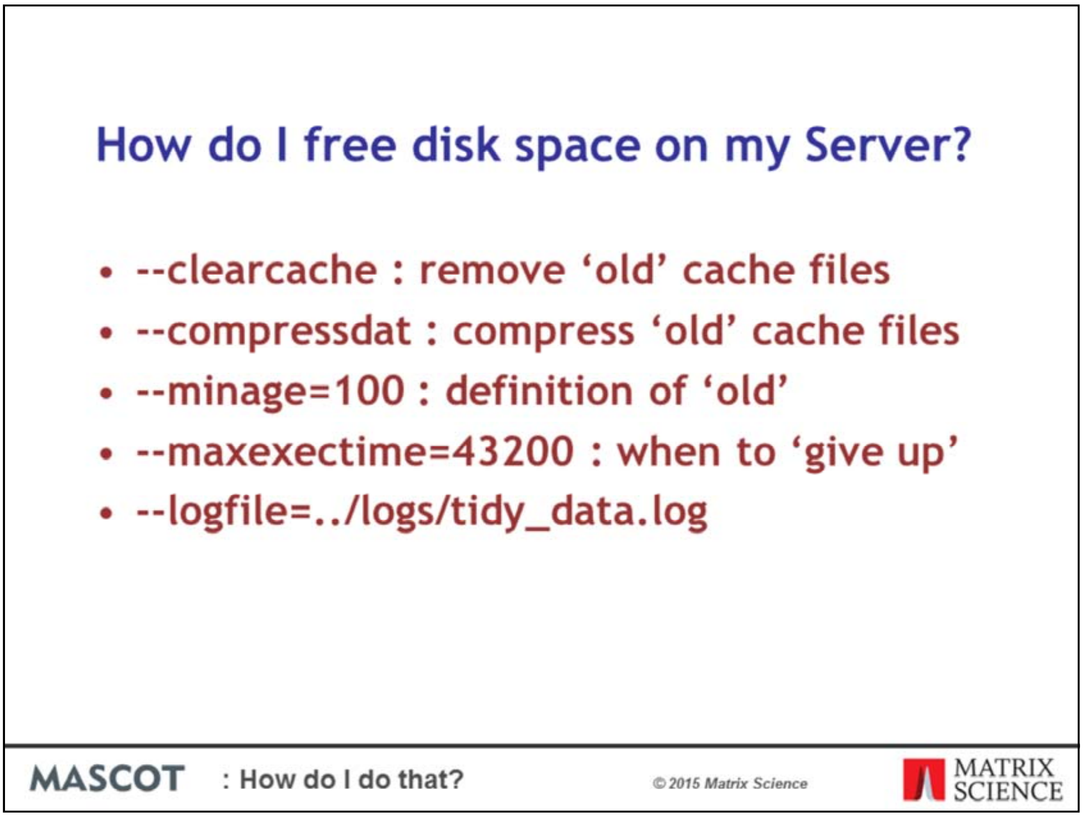
‘0’ means 0 minutes past the hour

‘2’ means 2 o’clock in the morning

‘*’ and ‘*’ means any day of the month and any month of the year

‘0’ means the first day of the week, that is Sunday.

One word of warning is that you should only use this script with Mascot 2.4 and later. The cron section of mascot.dat wasn’t implemented in earlier versions.



How do I free disk space on my Server?

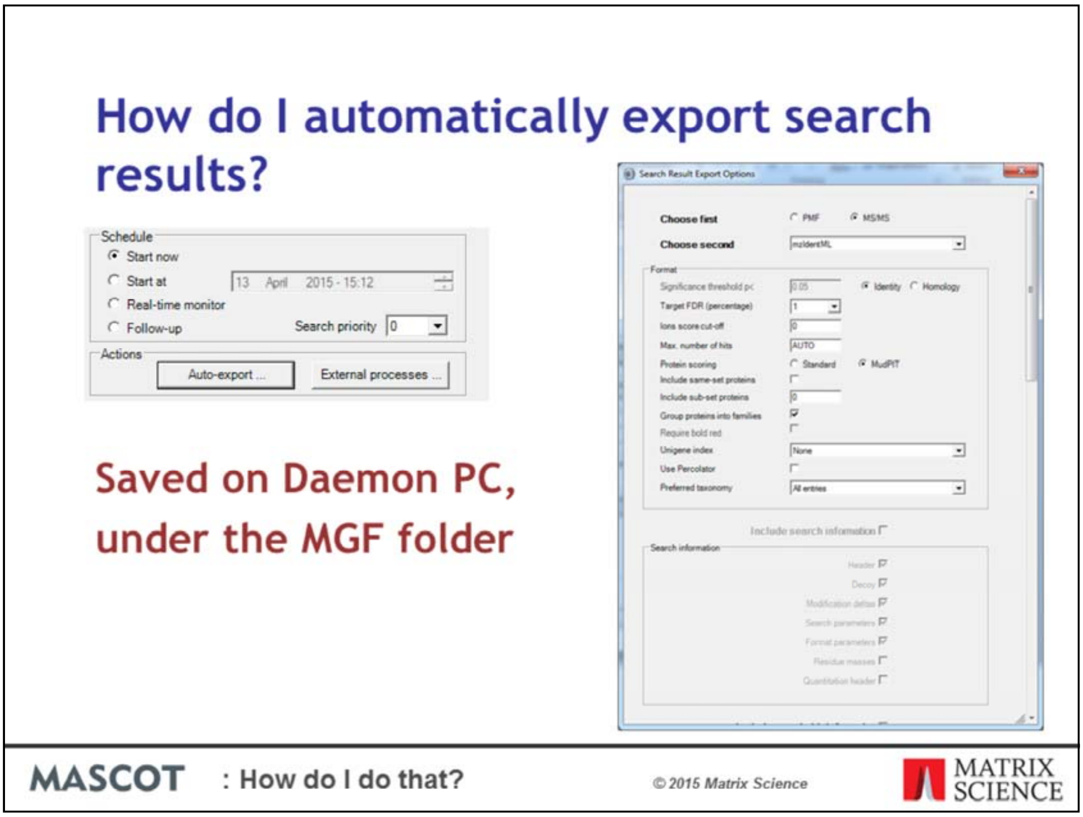
- --clearcache : remove ‘old’ cache files
- --compressdat : compress ‘old’ cache files
- --minage=100 : definition of ‘old’
- --maxexectime=43200 : when to ‘give up’
- --logfile=../logs/tidy_data.log

You may have noticed that there are a couple of parameters for the script. If you don’t want to compress your .dat files, but you do want to clear the cache remove the –compressdat

There are some other parameters. What do we mean by ‘old’. You can specify this by adding –minage=100, so only files older than 100 days will be deleted or compressed. The default is 100, so if you are happy with that value, there’s no need to add the parameter

When you first run this script, it may take a very long time to run if there are lots of files to compress. Assuming that you set this running at 2 am on Sunday morning, you probably don’t want it to still be running on Monday when you come into work. So, you can specify how long it should run before stopping. The default is 12 hours, which is 43,200 seconds.

You can also set the location and logging level for this, but there’s probably no reason to change the defaults.



How do I automatically export search results?

Schedule
- ● Start now
- ○ Start at 13 April 2015 - 15:12
- ○ Real-time monitor
- ○ Follow-up

Search priority 0

Actions

Search Result Export Options

Choose first ○ PMF ● MS/MS

Choose second mzIdentML

Format

Significance threshold p<	0.05	● Identity ○ Homology
Target FDR (percentage)	1	
Ions score cut-off	0	
Max. number of hits	AUTO	
Protein scoring	○ Standard	● MudPIT
Include same-set proteins	□	
Include sub-set proteins	0	
Group proteins into families	☑	
Require bold red	□	
Unigene index	None	
Use Percolator	□	
Preferred taxonomy	All entries	

Include search information □

Search information
- Header ☑
- Decoy ☑
- Modification deltas ☑
- Search parameters ☑
- Format parameters ☑
- Residue masses □
- Quantitation header □

Saved on Daemon PC, under the MGF folder

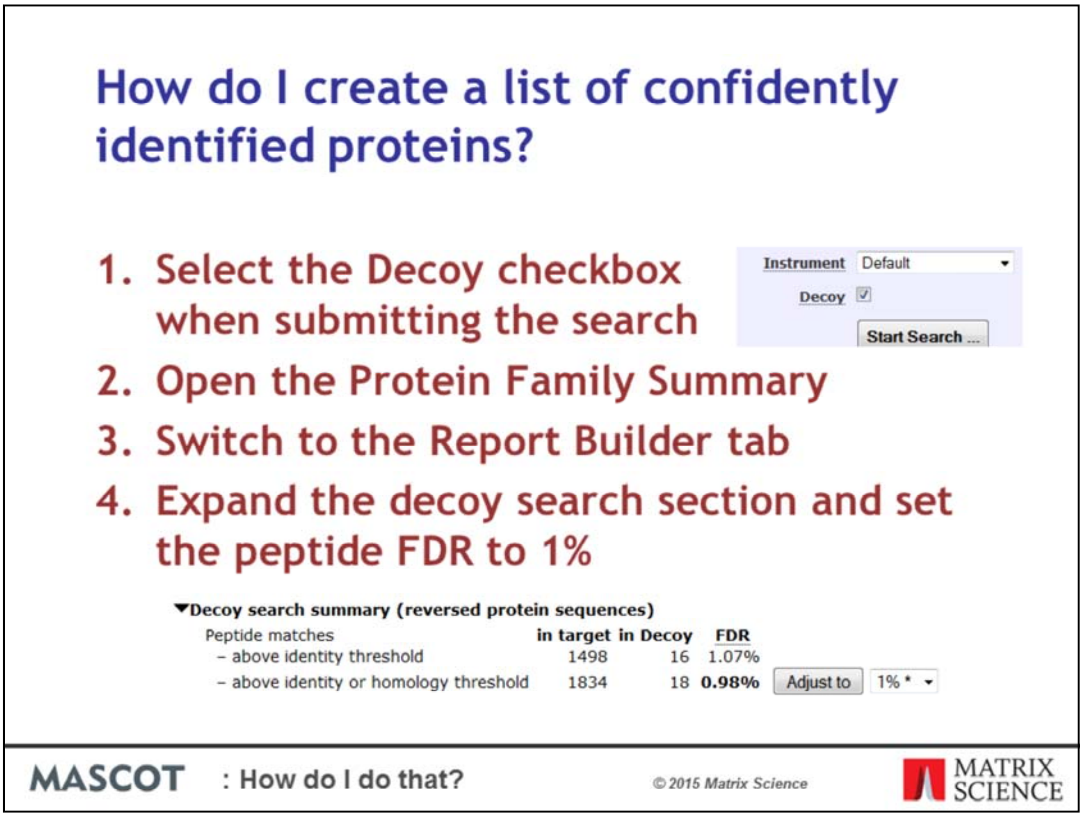
MASCOT : How do I do that?

© 2015 Matrix Science

MATRIX SCIENCE

We had this as a tip in one of our talks at the 2013 ASMS user meeting. Back then, there were 7 slides describing how to do this as it was quite complex. Because of that, we added an easy way to do it in Mascot Daemon 2.5, but you may have missed it. On the task editor page, there's a new button for "Auto-export". Just click this button and you can select the export file format that you want and also what you want exported. There are different options for PMF and MS/MS options, so make sure that you choose the type of search that you are doing first. The different options then depend on the output format that you have chosen.

The files are saved in the MGF folder – normally in a subdirectory under:
C:\ProgramData\Matrix Science\Mascot Daemon\MGF



How do I create a list of confidently identified proteins?

1. Select the Decoy checkbox when submitting the search
2. Open the Protein Family Summary
3. Switch to the Report Builder tab
4. Expand the decoy search section and set the peptide FDR to 1%

Instrument Default ▾
Decoy ☑
Start Search ...

▾Decoy search summary (reversed protein sequences)

Peptide matches	in target	in Decoy	FDR		
– above identity threshold	1498	16	1.07%		
– above identity or homology threshold	1834	18	**0.98%**	Adjust to	1% * ▾

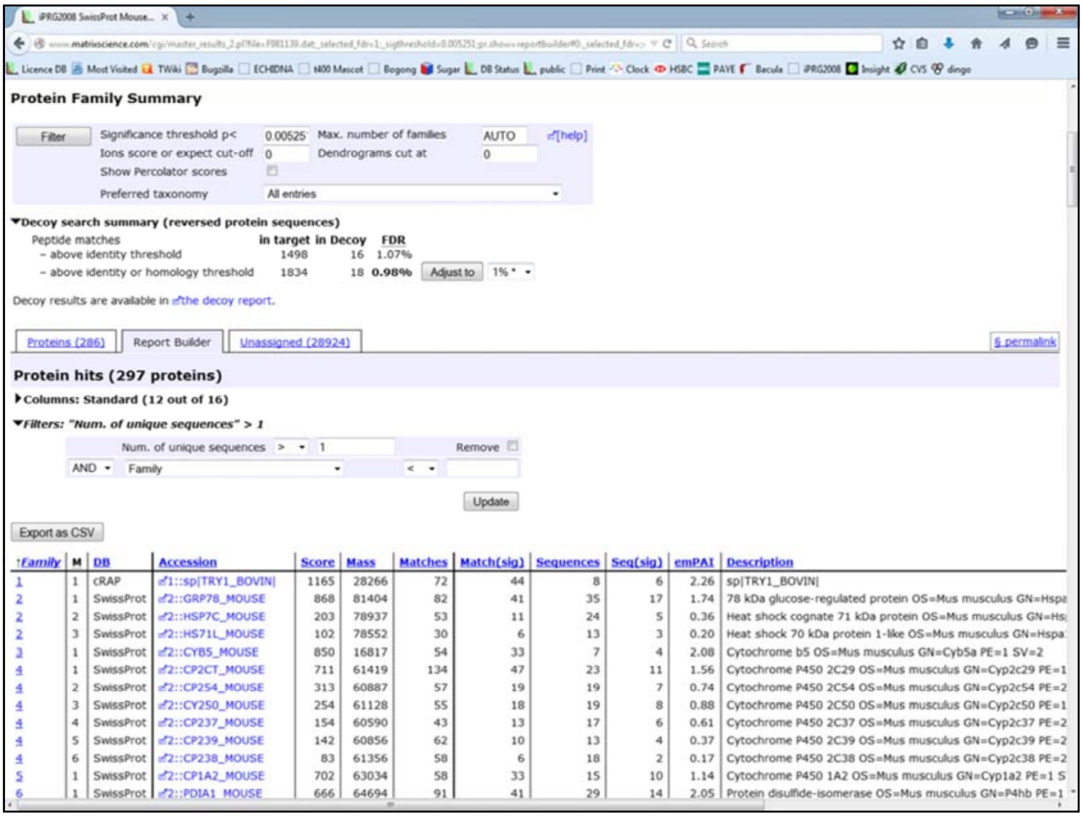
MASCOT : How do I do that? © 2015 Matrix Science MATRIX SCIENCE

Creating a list of confidentially identified proteins is very simple.

Select the Decoy checkbox in the search form when submitting the search. Assuming there are more than 300 ms-ms spectra, and assuming you haven’t changed any defaults, the Protein Family Summary will display by default. If it doesn’t, switch to that report

Next, switch to the Report Builder tab.

Then, expand the decoy search section and set the peptide FDR to 1%


iPRG2008 SwissProt Mouse...
www.matrixscience.com/cgi/master_results_2.pl?file=F981139.dat;_selected_fdr=1;_sigthreshold=0.005251;pr.show=reportbuilder#0;_selected_fdr
Licence DB
Most Visited
TWiki
Bugzilla
ECHIDNA
t400 Mascot
Bogong
Sugar
DB Status
public
Print
Clock
HSBC
PAYE
Bacula
iPRG2008
Insight
CVS
dingo

Protein Family Summary

Filter

Significance threshold p<	0.00525	Max. number of families	AUTO
Ions score or expect cut-off	0	Dendrograms cut at	0
Show Percolator scores	□		
Preferred taxonomy	All entries		

[help]

▼**Decoy search summary (reversed protein sequences)**

Peptide matches	in target	in Decoy	FDR
– above identity threshold	1498	16	1.07%
– above identity or homology threshold	1834	18	**0.98%**

Adjust to 1% *

Decoy results are available in the decoy report.

Proteins (286) | Report Builder | Unassigned (28924) | § permalink

Protein hits (297 proteins)

▶**Columns: Standard (12 out of 16)**

▼***Filters: "Num. of unique sequences" > 1***

Num. of unique sequences > 1 Remove □

AND Family <

Update

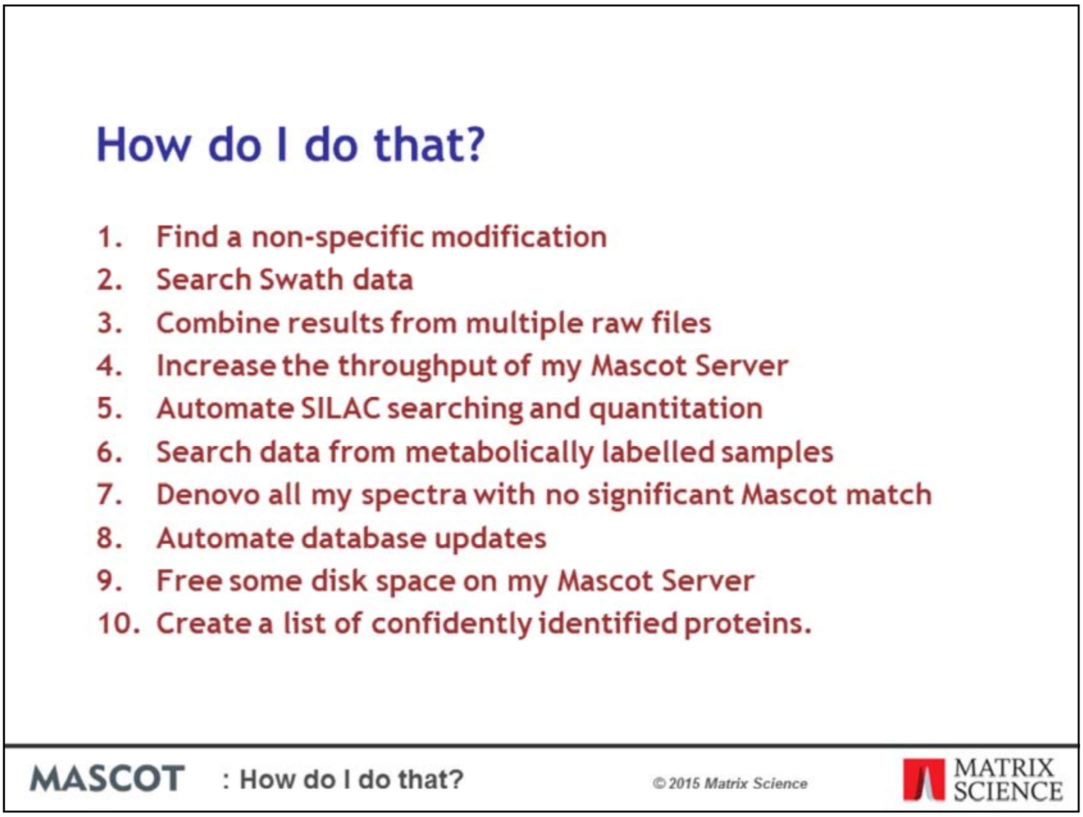
Export as CSV

↑Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	cRAP	1::sp|TRY1_BOVIN|	1165	28266	72	44	8	6	2.26	sp|TRY1_BOVIN|
2	1	SwissProt	2::GRP78_MOUSE	868	81404	82	41	35	17	1.74	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa
2	2	SwissProt	2::HSP7C_MOUSE	203	78937	53	11	24	5	0.36	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hs
2	3	SwissProt	2::HS71L_MOUSE	102	78552	30	6	13	3	0.20	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa
3	1	SwissProt	2::CYB5_MOUSE	850	16817	54	33	7	4	2.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
4	1	SwissProt	2::CP2CT_MOUSE	711	61419	134	47	23	11	1.56	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1
4	2	SwissProt	2::CP254_MOUSE	313	60887	57	19	19	7	0.74	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2
4	3	SwissProt	2::CY250_MOUSE	254	61128	55	18	19	8	0.88	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1
4	4	SwissProt	2::CP237_MOUSE	154	60590	43	13	17	6	0.61	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2
4	5	SwissProt	2::CP239_MOUSE	142	60856	62	10	13	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2
4	6	SwissProt	2::CP238_MOUSE	83	61356	58	6	18	2	0.17	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2
5	1	SwissProt	2::CP1A2_MOUSE	702	63034	58	33	15	10	1.14	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 S
6	1	SwissProt	2::PDIA1_MOUSE	666	64694	91	41	29	14	2.05	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1

Expand the filters section and set ‘Num of significant unique sequences’ > 1

Optionally, expand the columns section and choose which columns you require and their order

Print the table or export it as CSV



How do I do that?

1. Find a non-specific modification
2. Search Swath data
3. Combine results from multiple raw files
4. Increase the throughput of my Mascot Server
5. Automate SILAC searching and quantitation
6. Search data from metabolically labelled samples
7. Denovo all my spectra with no significant Mascot match
8. Automate database updates
9. Free some disk space on my Mascot Server
10. Create a list of confidently identified proteins.