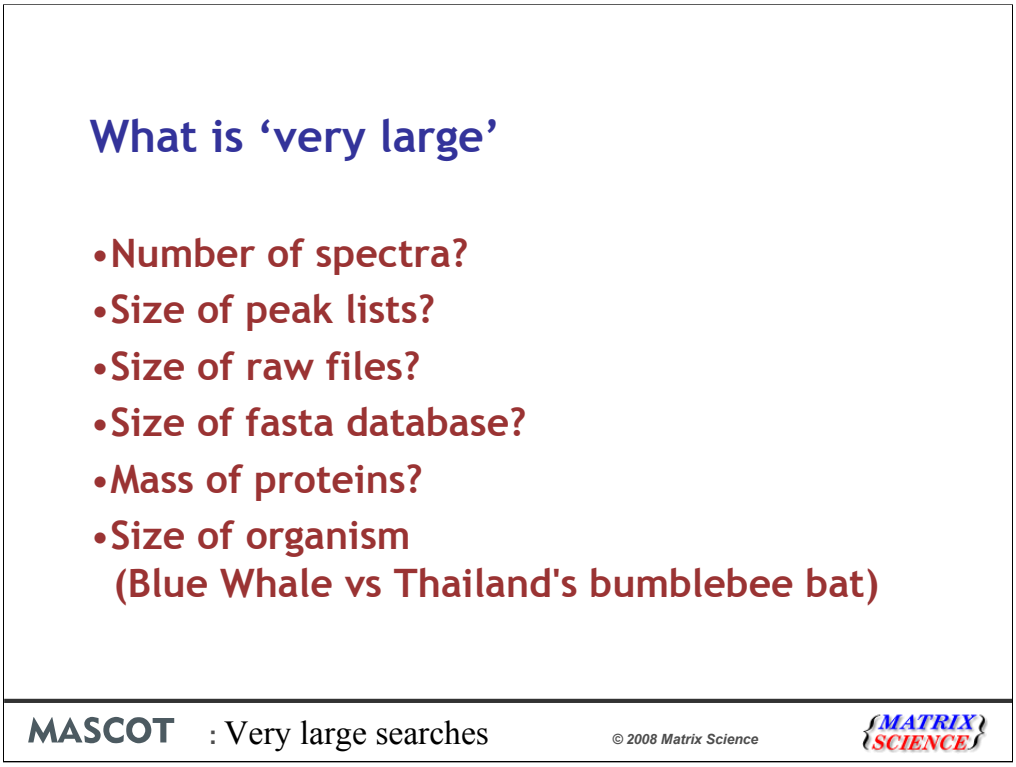
Challenges of very large searches

MASCOT

MATRIX SCIENCE



The first challenge is actually decide what we mean by very large. And before we try and put any number on this, we need to decide what we are going to measure.

In general, ‘large’ is related to the number of ms-ms spectra which should of course be related to the size (in GB) of the peak lists, but we’ll return to that later.

From a particular instrument, you’d expect the size of the raw files to be proportional to these, but don’t expect sizes between different instrument manufacturers to be similar. Also, remember that some instruments can save as profile or centroided

I’ll also be discussing the size of the fasta file because this can cause some issues. Also, it will obviously take longer to search the complete human genome than say IPI.

And finally, we can’t even rule out the size of the proteins when looking at these problems - although they should have a minimal effect.

Although sample preparation from a huge or a tiny organism may present a challenge, at least this has no bearing to the data processing.

Why

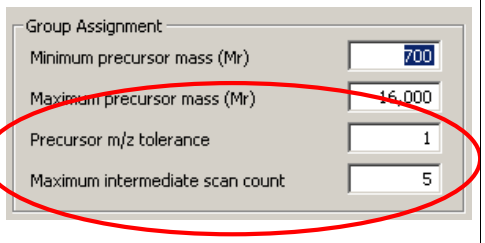
- Because I like a challenge!
- Multiple files from the same sample using MudPIT cannot be considered in isolation.

By the end of the talk you may wonder why any one ever want to perform large searches.For some people, they seem to just want a challenge, I'd recommend climbing Mount Everest as it's something you can brag to more people about.

For anyone performing MudPIT type searches, there is little choice because you need to see all the results together - the different SCX fractions cannot be considered in isolation. It's often best to merge all the peak lists together and perform a single search.

Merging ms-ms spectra

- Most data systems have option to merge ms-ms spectra from same precursor mass
- For MudPIT, this ideally needs to be performed across multiple raw files



- Distiller 2.2 has option to merge from MudPIT raw files based on a retention time window.

One problem (or some might say one advantage) of the MudPIT approach is that the same peptide will normally be analysed multiple times. So you are likely to see the same peptide in multiple SCX fractions.

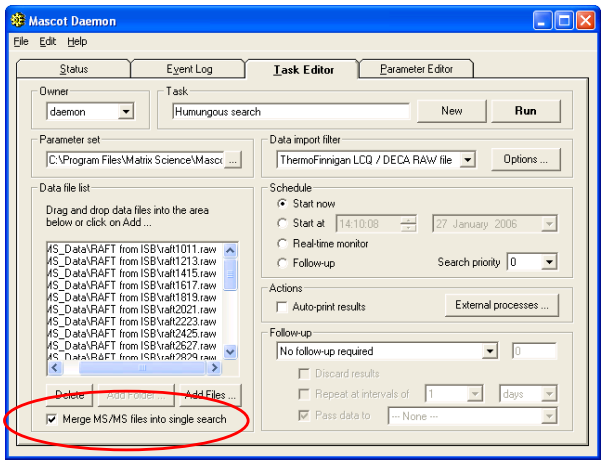
Most data systems have the option to merge ms-ms spectra with similar precursors that have eluted within a specified time window. The advantages of this are improved signal to noise, a smaller peak list and fewer duplicate peptides so reports are less cluttered.

Ideally, this should be performed across multiple data sets, but this can't be done with Mascot at the moment. The next release of Distiller has this option, by specifying a retention time window that is used to determine whether to merge similar spectra in different raw files.

Other software systems compare spectra with similar precursor masses using a cross correlation type of algorithm, and merge these.

Combining data files

- Can use Mascot Daemon to process and merge MudPIT fractions
- Use Distiller or a file specific data import filter



For the moment, the smartest way to merge files, like fractions from a MudPIT run, is using Mascot Daemon. Just tick the box at the bottom left.

The batch can be peak lists or raw files

Note that Mascot Daemon 2.1 had a file size limit of 2 GB. This was lifted in 2.2, and we have successfully merged and searched a 6 GB file

Combining data files

Concatenating peak lists:

- DTA or PKL

  Download merge.pl from the Matrix Science Xcalibur help page
  `http://www.matrixscience.com/help/instruments_xcalibur.html`

  Retains filename as scan title

```
BEGIN IONS
TITLE=raft3031.1706.1706.2.dta
CHARGE=2+
PEPMASS=1243.577388
451.1228 5080
487.4352 3283
550.4203 5087
```

MASCOT : Very large searches MATRIX SCIENCE

If you don't want to use Daemon, you can merge peak lists manually.

For DTA or PKL, you can download a script from our web site.

A nice feature of this script is that it puts the filename into the scan title, so you can tell which fraction a particular spectrum came from. The scan titles are displayed in the yellow pop-ups on the Mascot result report

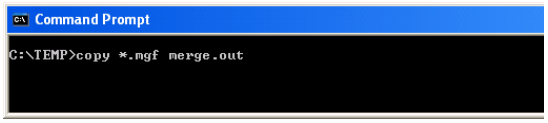
You can't merge mzData files because XML doesn't work like that.

Combining data files

Concatenating peak lists:

- MGF

Windows: copy



Unix: cat

As long as MGF files contain only peak lists, you don't need a script. Just use copy or cat

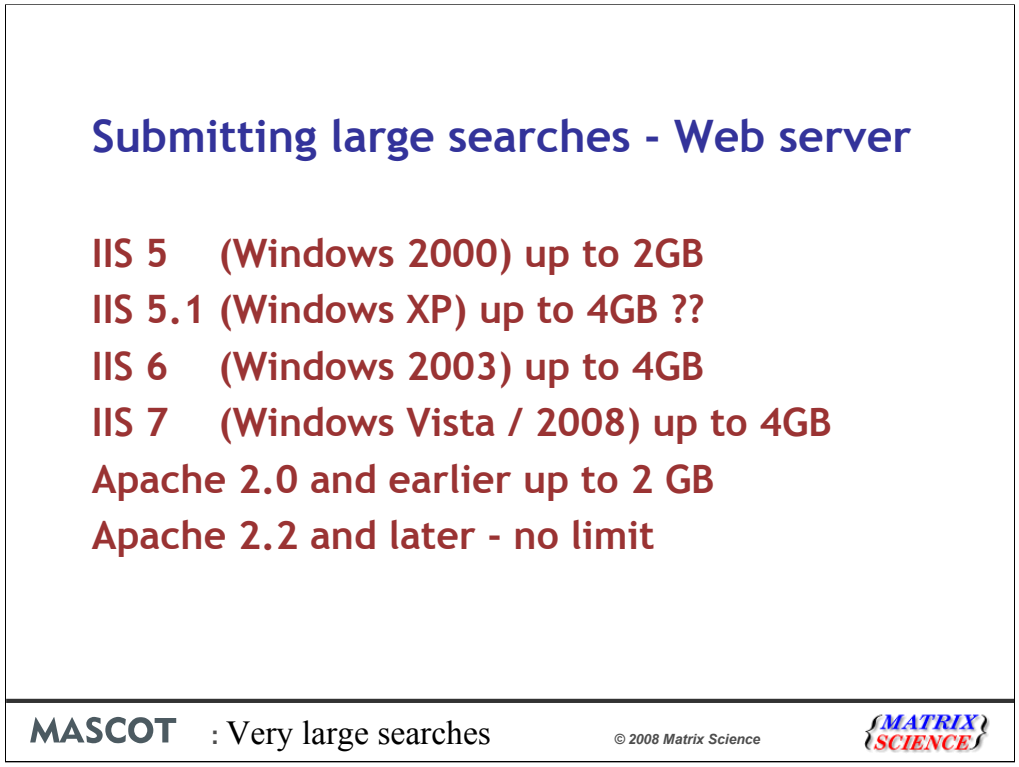
If the MGF files have search parameters at the beginning, you'll need to remove these before merging the files.

Combining data files

- Average spectrum might contain 100 real peaks
- Each peak might require ~ 20 bytes
  967.41590 [tab] 470.20193 [newline]
- 2 GB should be sufficient for ~ 1 million spectra
- If your peak list is orders of magnitude larger than 2kB / spectrum, then something is not right!

In talking to Mascot users, it is clear that peak lists files are often much bigger than they should be. In other words, the peak detection is not very good. If you do a back of the envelope calculation, you can see that 2 GB should be enough for approximately 1 million spectra.

If you intend to do a lot of large searches, its worth getting the peak detection right. Shipping unnecessarily large files around wastes both time and disk space

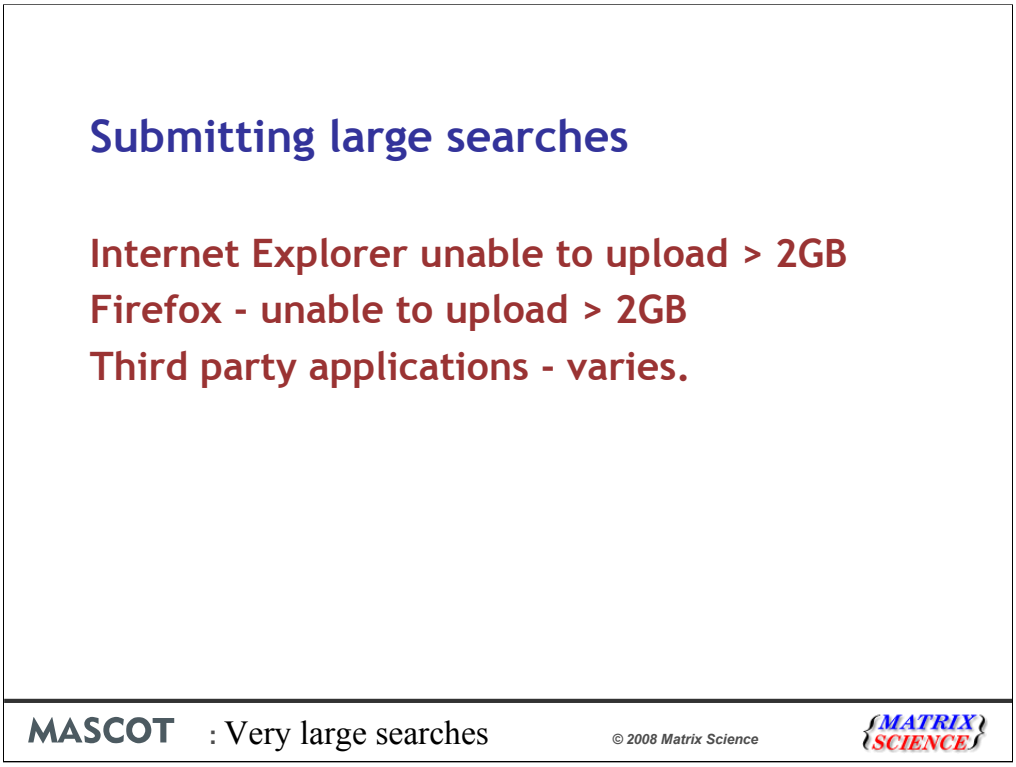


We've combined our files to produce one huge peak list. How do we submit this to Mascot?

In most cases, the submission is done using a web browser, such as Internet Explorer or Firefox to the Mascot web server. As we all know, there are lots of advantages to using this technology, but one disadvantage is that some web servers won't accept huge uploads. This list shows the published limits for the most commonly used web servers.

In practice, if you use Mascot Daemon version 2.2 or later, then the limit with all versions of IIS is 4Gb.

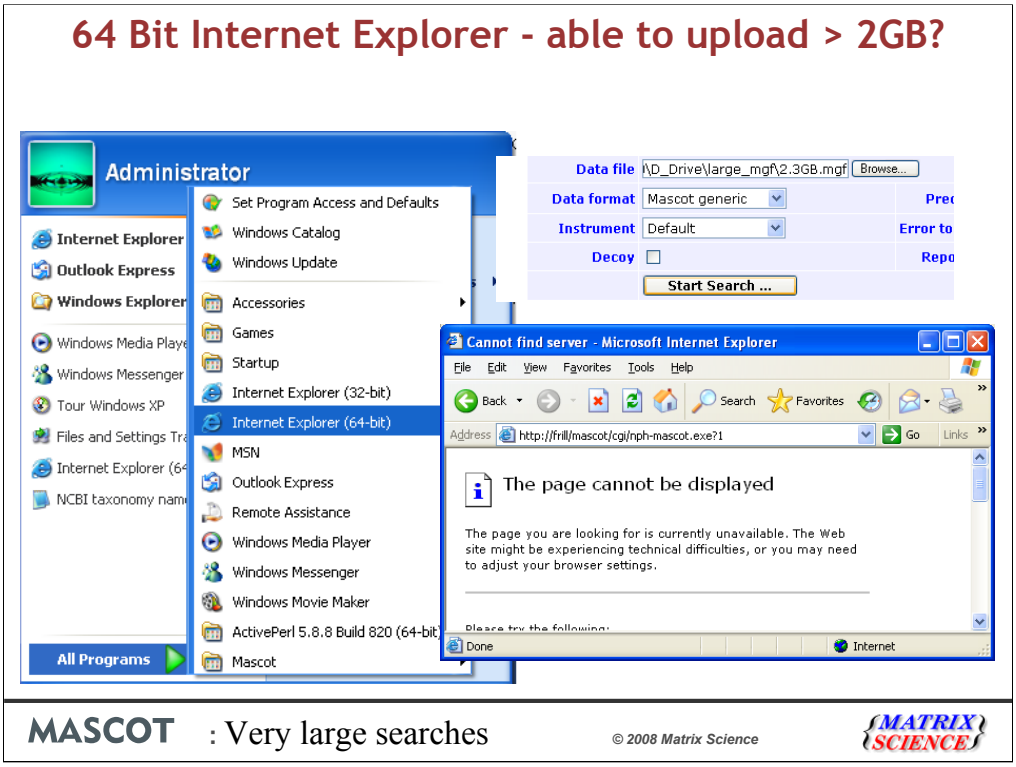
The question marks with XP are because there's conflicting information on the Microsoft web site and as we'll see in a minute, there's no easy way to test this.



If you are going to submit a search using a web browser, then there is a limit of 2GB

Third party applications from Instrument manufacturers will vary. For those that display a the search form in a web browser like window, the limits will be the same as for Internet Explorer.

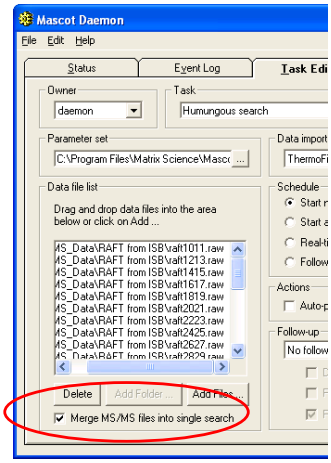
At this point, some of you may be thinking that 64bit Windows or 64bit Linux may provide an answer. Let's try it:



In 64 bit Windows XP, there's an option to run 32 bit or 64 bit Internet Explorer. If you try either with a 2.3 GB file, it fails.

Submitting large searches

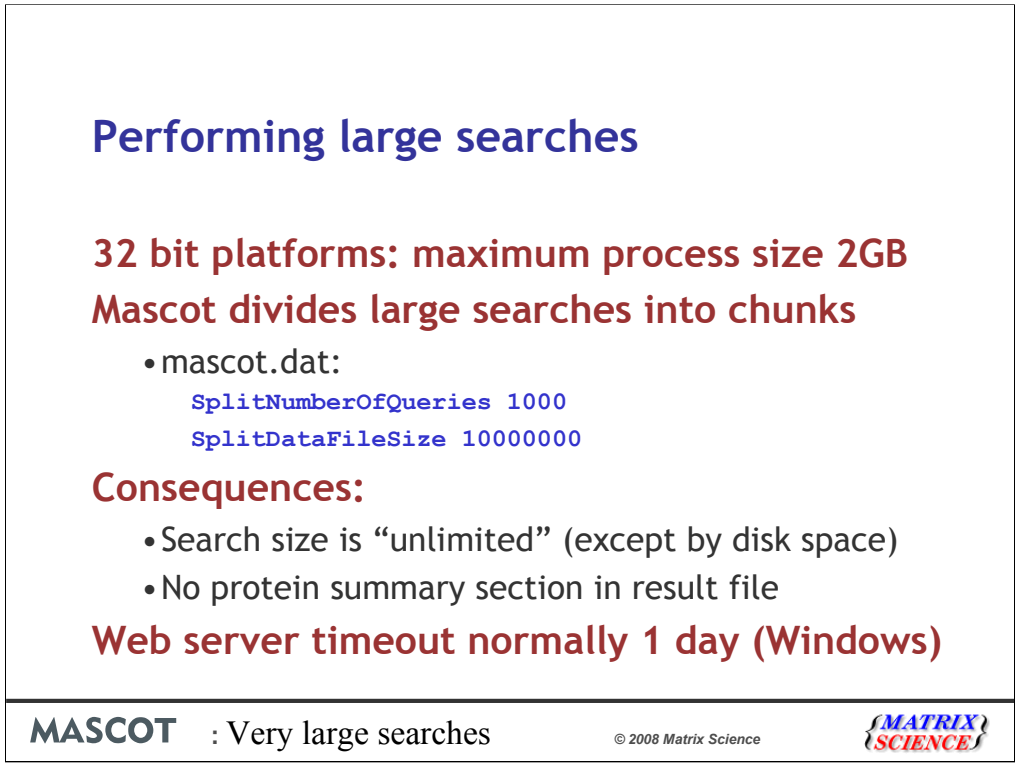
- Using Daemon, unlimited file size for upload
- Tested 6GB uploads.



MASCOT : Very large searches

MATRIX SCIENCE

So, the best thing is to use Mascot daemon, but remember that there is still a limit of 4GB for all IIS servers.



Next step is the actual search, and the good news is that there is no real problem here.

32 bit platforms, like most Windows and Linux installations, have a maximum process size of 2 GB on Windows or 3GB on Linux. To get around this limit, Mascot divides large searches into smaller chunks, so as to avoid having everything in memory at the same time. The parameters to control this are SplitNumberOfQueries and SplitDataFileSize in the Options section of mascot.dat

One consequence of splitting a search is that there is no protein summary section in the result file. This is not a problem, because no-one wants a protein summary report for a large MS/MS search.

One thing to watch out for with a huge search is that the Web Server timeout is set to one day by default (on a Windows server). We've had a couple of support calls where a customer with a huge job found that there search had got to about 95% complete and then died - rather frustrating.

Reporting large search results

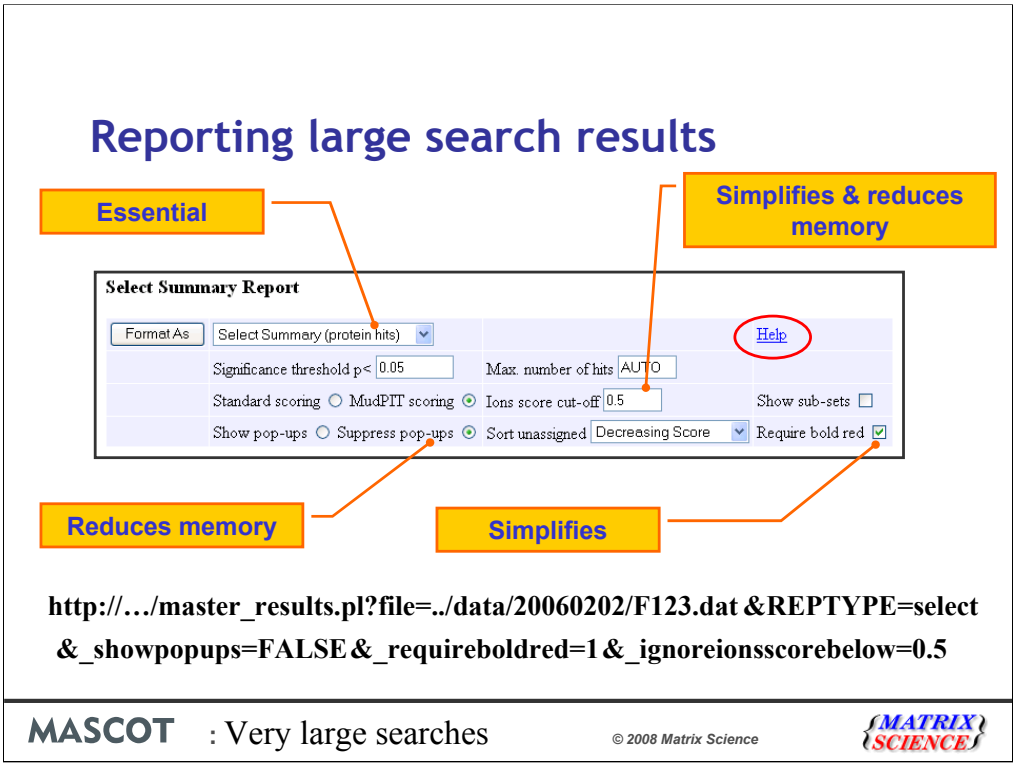
Not so easy!

- Web server or web browser timeouts
  - Try generating report at a command prompt
- Out of memory on server
  - Windows: Task Manager, Unix: top
  - Fix is to use 64 bit platform and lots of RAM
- Out of memory on client
- Browser response becomes unacceptably slow
  - Internet Explorer 5.5 particularly bad.

Running large searches is easy. Reporting the results is not so easy.

One problem is that it can be quite hard to see where the problem is. The symptom, if you like is that in the browser, if it's IE the world keeps spinning and you are eventually left with a blank screen. At this point, it's not clear whether the problem is on the Mascot server or is on your computer.

Common problems are timeouts and running out of memory. These problems may be on the server side or they may be on the client side. Here are some troubleshooting tips



With very large searches, it becomes important to minimise the size of the result report. The key format controls are:

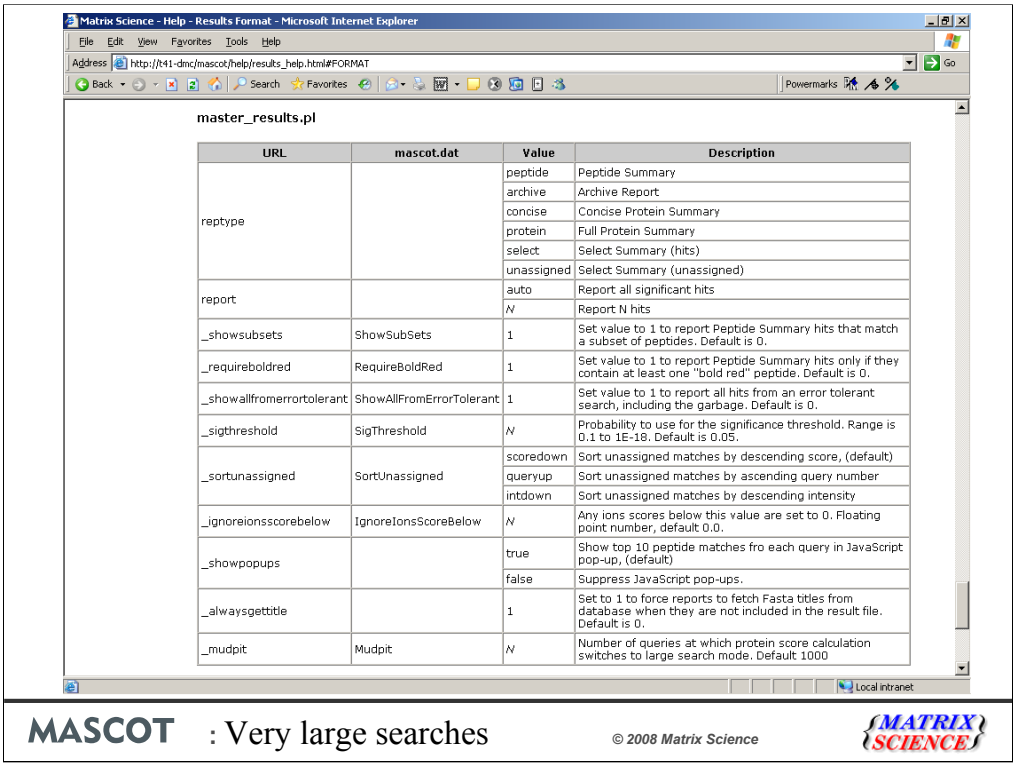
•Ensure you are using the Select report. If you are using a third party client that has specified Peptide summary or Protein summary, add this to the URL before opening the file: &REPTYPE=select

•Get rid of the yellow pop-ups: &_showpopups=FALSE

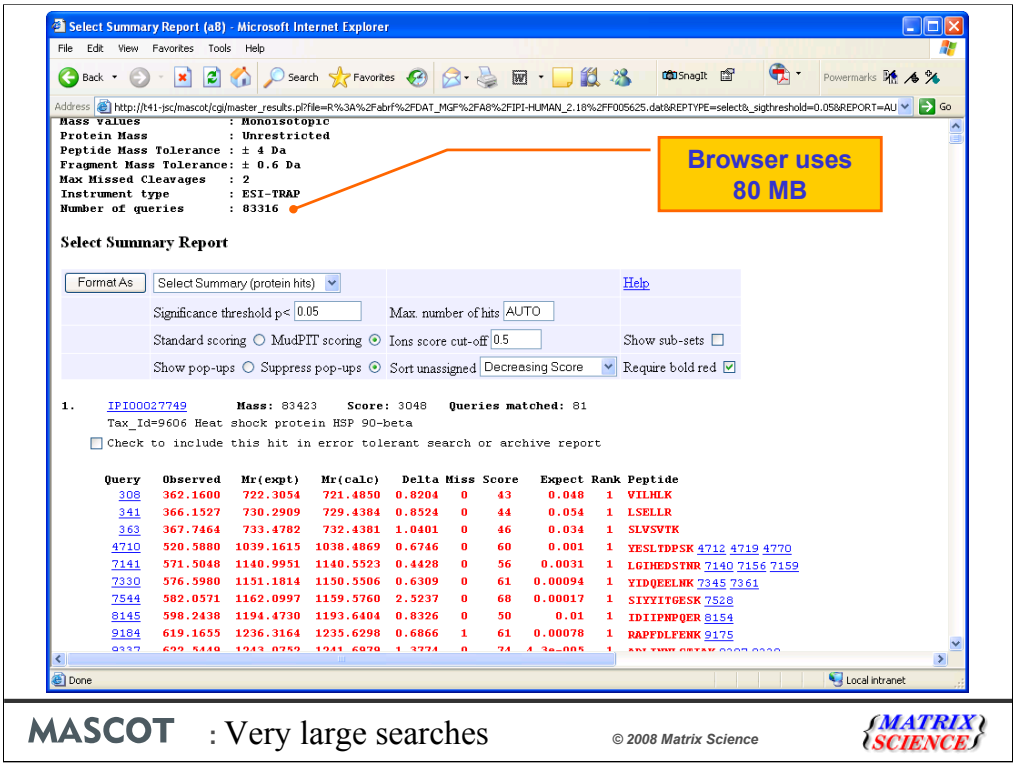
•Set number of hits to AUTO, so that you don't list lots of low scoring protein hits: &REPORT=AUTO

•Setting require bold red and an expect value cut-off will minimise the number of hits: &_ignoreionsscorebelow=0.5&_requireboldred=1

Note that the ions score cut-off is just that when the value is 1 or greater. When the value is between 0 and 1, it is an expect cut-off, which is much more useful. I usually set this to 0.5 to get rid of all the junk matches.

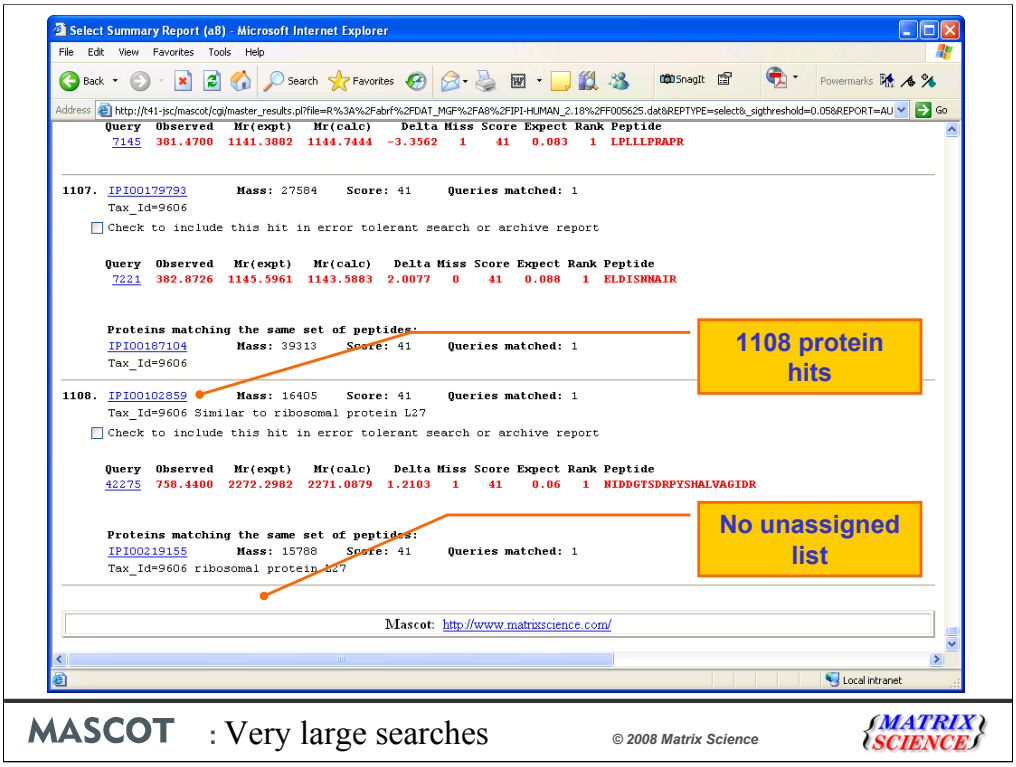


If you can't remember these URL parameters, just click on the help link



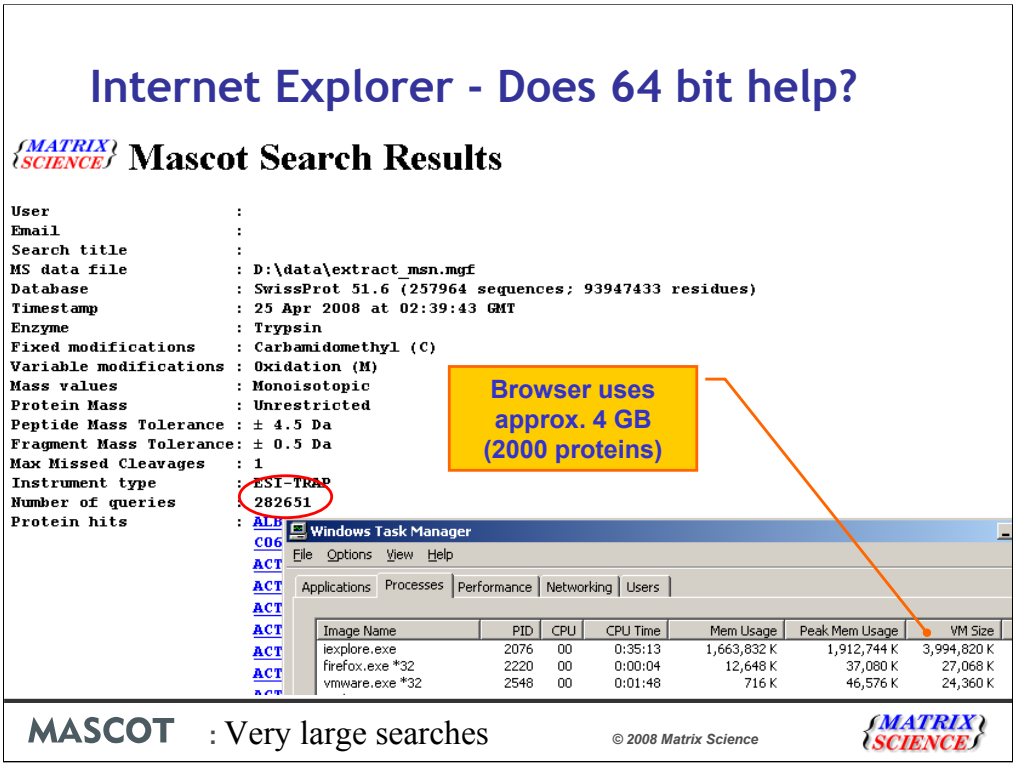
With these settings, a Select Summary for a typical mudPIT search of 80 thousand spectra uses about 80 MB of RAM when displayed in Internet Explorer 6.0.

One factor that keeps the select summary concise is that multiple matches to the same peptide are collapsed onto a single line.



But, the main reason why the select report is more memory efficient is that it omits the unassigned list. This is the list of weak and random matches which don't make it into the body of the report, where peptide matches are assigned to proteins. For MudPIT, this is usually around 95% of all queries! So the exact size of the select summary depends mainly on how many queries got strong matches.

If you want to see the unassigned list, this is available as a separate report

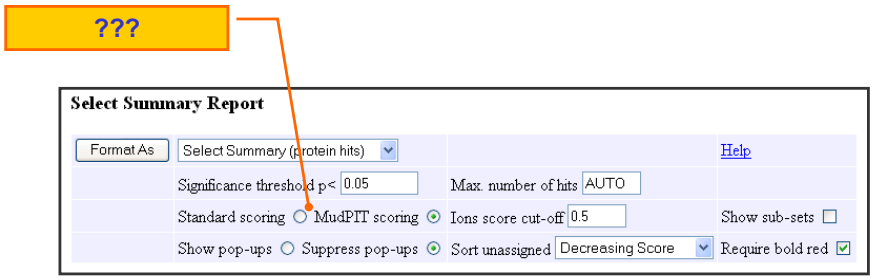


However, a search with 280,000 queries and 2000 proteins takes up approximately 4GB RAM using 64 bit IE7. What's worse is that it takes 35 seconds to respond after pressing page down.

If you create the html file, using the command line as I described earlier, the html file is about 800MB.

One of the big claims for Firefox 3.0 is that it is much less memory hungry than Internet Explorer - however, it seems to be similar to me.

Reporting large search results



MASCOT : Very large searches

MATRIX SCIENCE

What do we mean by Standard scoring and MudPIT scoring?

Protein Scores for MS/MS Searches

Standard protein score

- the sum of the ions scores
- excluding the scores for duplicate matches, which are shown in parentheses
- correction to reduce the contribution of low-scoring random matches

183. IPI00141647 **Mass:** 3011421 **Score:** 47 **Queries matched:** 5
Tax_Id=9606 titin isoform N2-B
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
76	424.6860	847.3575	845.4355	1.9220	1	17	27	8	NDGGSRIK
118	446.7500	891.4854	889.4691	2.0164	0	17	25	4	GGIQIMAGK
358	366.3330	1095.9773	1092.6179	3.3594	0	23	5.7	8	YISSLEILR
569	439.3649	1315.0730	1313.6615	1.4115	0	26	2.9	1	EPVLYDTHVNK
1182	870.8864	1739.7583	1741.8886	-2.1303	0	15	27	3	VTAVNEYGPGVPTDVPK

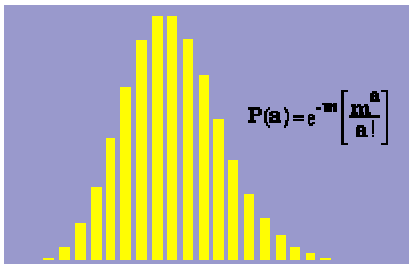
MASCOT : Very large searches MATRIX SCIENCE

With standard peptide summary scoring, the protein score is essentially the sum of the ions scores of all the non-duplicate peptides. Where there are duplicate peptides, the highest scoring peptide is used. A correction is applied based on the number of candidate peptides that were tested. This correction is very small unless it is a very large protein, like here, or a no-enzyme search

Despite this correction, as this example shows, we can still get a protein score of 47 even though none of the peptide matches are significant

Protein Scores for MS/MS Searches

- Even if you only have random matches, you can still get multiple matches to a protein.
- The distribution of random matches depends on the ratio between the number of spectra and the number of entries in the database
- Poisson distribution



MASCOT : Very large searches

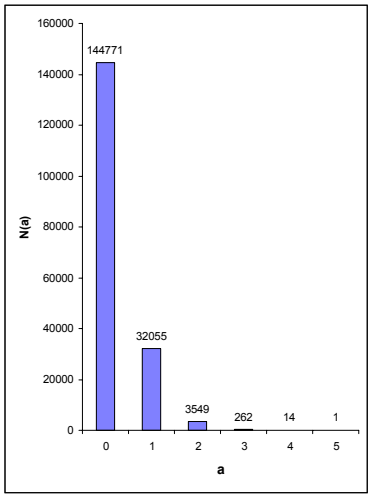
MATRIX SCIENCE

Even if peptide matches are random, you can still get multiple matches to a single protein. How likely this is depends on the ratio between the number of spectra and the number of entries in the database. We can predict whether this will be a serious problem or not using a function called a Poisson distribution.

If average number of events per interval is *m*, then the Poisson distribution gives us the probability of observing *a* events in a particular interval.

Protein Scores for MS/MS Searches

Shotgun / MudPIT
- 20 SCX fractions
- 160,000 scans total
- 80,000 after processing
- 40,000 random matches in search of Swiss-Prot (180652 entries)



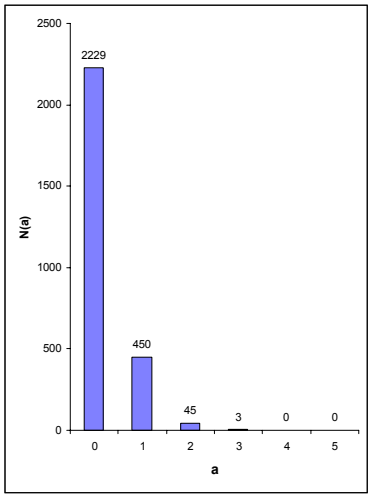
MASCOT : Very large searches

MATRIX SCIENCE

For this MudPIT search, 262 proteins are expected to pick up 3 random matches by chance. 1 protein will pick up 5

Protein Scores for MS/MS Searches

Small database
- 30 minute run
- 1500 scans total
- 1200 after processing
- 550 random matches in search of Swiss-Prot using drosophila taxonomy filter (2727 entries)



MASCOT : Very large searches

MATRIX SCIENCE

The problem isn't limited to large searches. It is the ratio between the number of spectra and the number of entries in the database that matters. So, a small search against a small database can give similar numbers

Protein Scores for MS/MS Searches

MudPIT protein score

- The sum of the excess of the ions score over the identity or homology threshold for each query
- Plus 1 x the average threshold

178. IPI00001639 **Mass:** 98420 **Score:** 46 **Queries matched:** 3
Tax_Id=9606 Importin beta-1 subunit
☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
22	386.4956	770.9767	770.4286	0.5481	0	22	8.5	3	DPSVVVR
914	779.7214	1557.4282	1555.8205	1.6077	1	23	4.9	2	TVSPDRLELEAAQK
1359	918.3068	1834.5991	1832.8839	1.7152	0	46	0.024	1	GDQENVMPDVMLVQPR

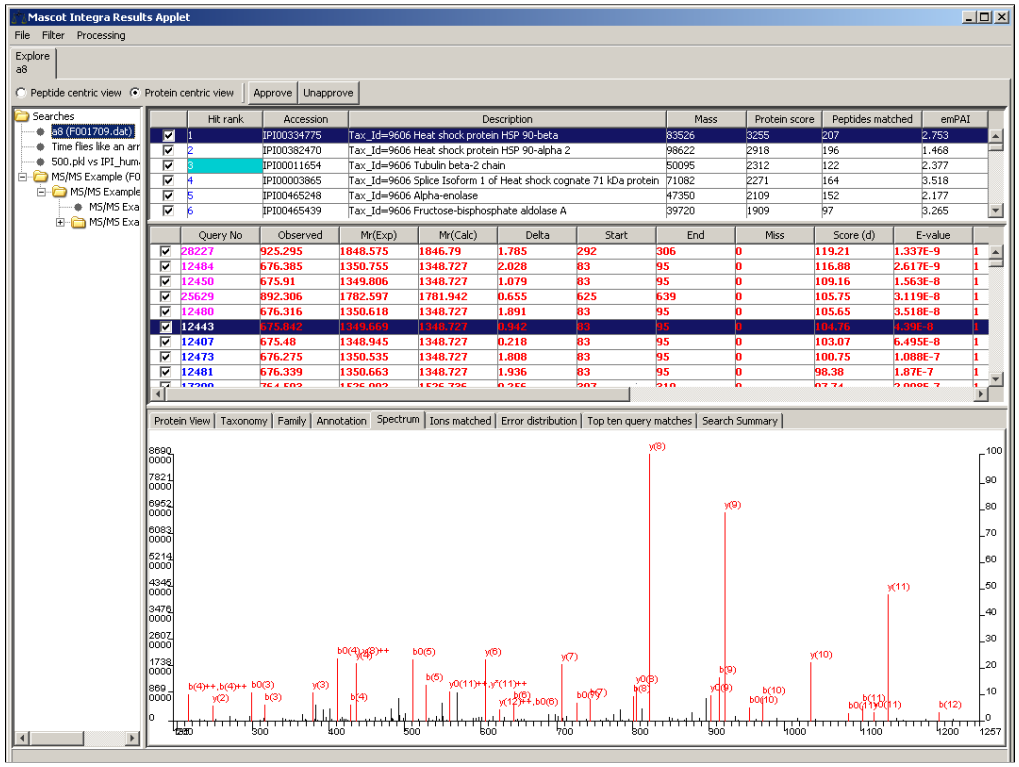
MASCOT : Very large searches MATRIX SCIENCE

For MudPIT scoring, the score for each peptide is not its absolute score, but the amount that it is above the threshold. Therefore, peptides with a score below the threshold do not contribute to the score. Finally, the average of the thresholds used is added to the score. For each peptide, the "threshold" is the homology threshold if it exists, otherwise it is the identity threshold. Note that there will be no homology score for some peptides.

You shouldn't see proteins with a large number of weak peptide matches getting a good score. If there are no significant peptides, the protein score will be 0.

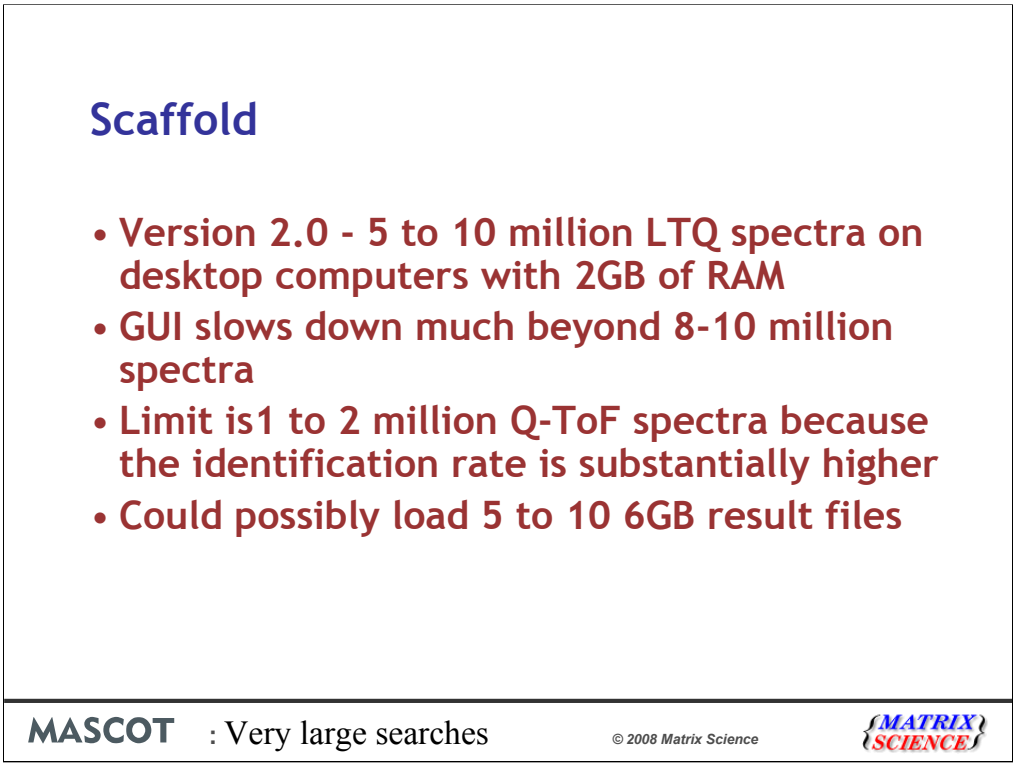
By default, MudPIT protein scoring is used when the ratio between the number of queries and the number of database entries, (after any taxonomy filter), exceeds 0.001.

This default switching point can be moved by changing the value of MudpitSwitch in mascot.dat. You can also switch between the two scoring methods by using the format controls at the top of the report.



For the rest of the talk, I'd like to return to the issue of struggling to see the results in a browser. As I've shown, you can probably see pretty much any report in 64 bit Internet Explorer, although you might wait 30 seconds for a scroll down. I can confirm that we are working hard on improving this issue for Mascot 2.3, due at the end of the year.

However, there are other superior alternatives. For example, Mascot Integra copes well with large reports because all of the results are saved in the underlying relational database. The screen shot here shows the new Mascot Integra applet that loads and allows easy browsing one of the huge results files I've shown earlier.

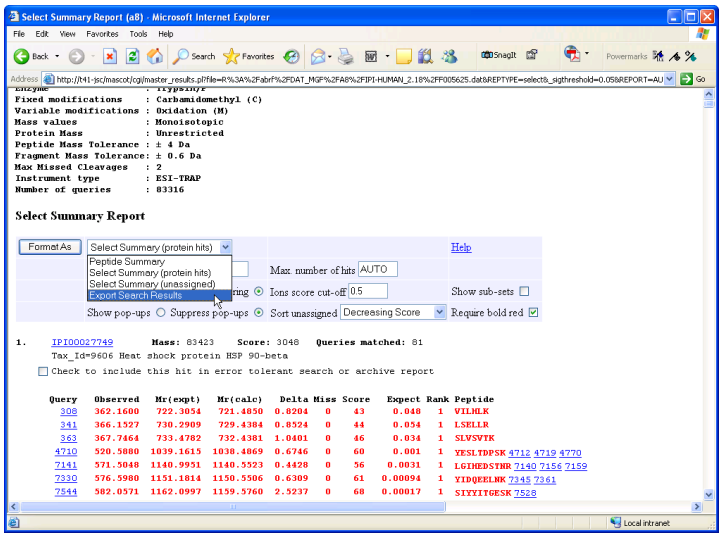


Another alternative is to use Scaffold. The new version 2 copes with much larger results file and can load 5 to 10 million LTQ spectra.

In practice, you don't want to be looking at more than 8 million spectra regardless of how much memory you have, because, like with IE and far fewer spectra, the GUI will start to slow down.

Just one point, the limits are lower with Q-TOF or Q-Star data because the identification rate is much higher. As with the standard Mascot reports, choosing to ignore junk spectra really helps with trap data.

Search result export



MASCOT : Very large searches MATRIX SCIENCE

If you don't want to consider Scaffold or Integra, perhaps because you can't afford more software, but have lots of time or cheap labour, a DIY approach is also possible, by exporting the Mascot results to a spreadsheet or database.

When a Mascot search is run, the results for the search are saved in a mime format text file on the Mascot server. A perl script reads that results file, and displays the html in a nice friendly way in your browser. The results text file itself could never be described as bedtime reading - even for me.

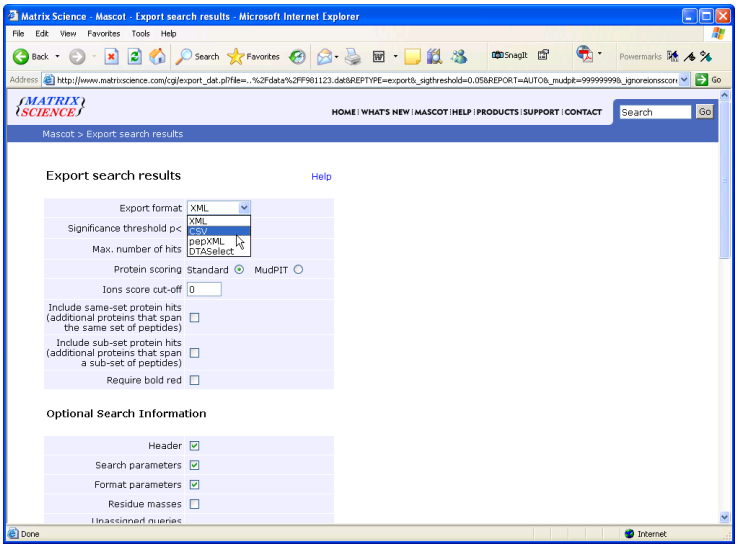
In Mascot version 2.0 and later, these perl scripts use a toolkit that we call Mascot parser. If you are developing a relational database application, you could use Mascot Parser to extract the data from the results files.

In most cases, a faster alternative is the export facility, added to Mascot in version 2.1.

It will output the results in a number of formats.

In the drop down list for the report formats, choose Export Search results, then press the "Format As" button

Search result export



MASCOT : Very large searches MATRIX SCIENCE

You now have a page with lots of formatting options - the first choice is the output format.

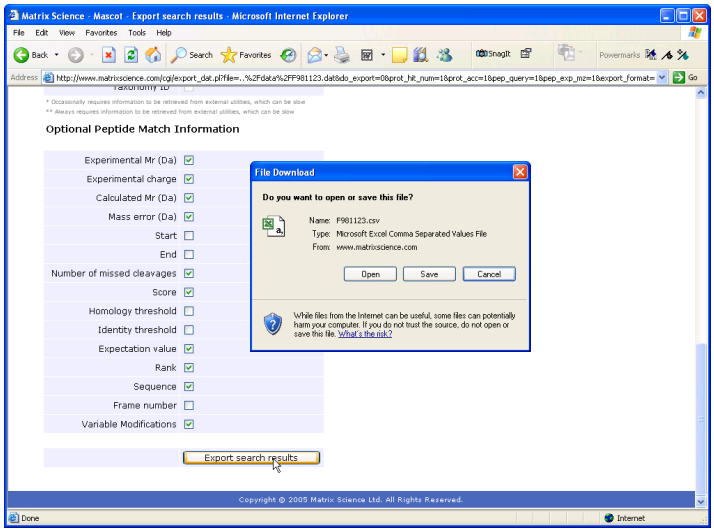
If you want the XML format, you probably know that this is what you want. If you've no idea what XML is, chances are you don't want it.

Choose CSV if you want to export to Excel - I'll show an example in a moment.

Choose pepXML if you want to export to Protein Prophet from ISB. We would recommend that you use this exporter rather than ISB's own Mascot2XML converter.

Finally, if you are using Dave Tabb's DTASelect then you will want to choose this last option.

Search result export

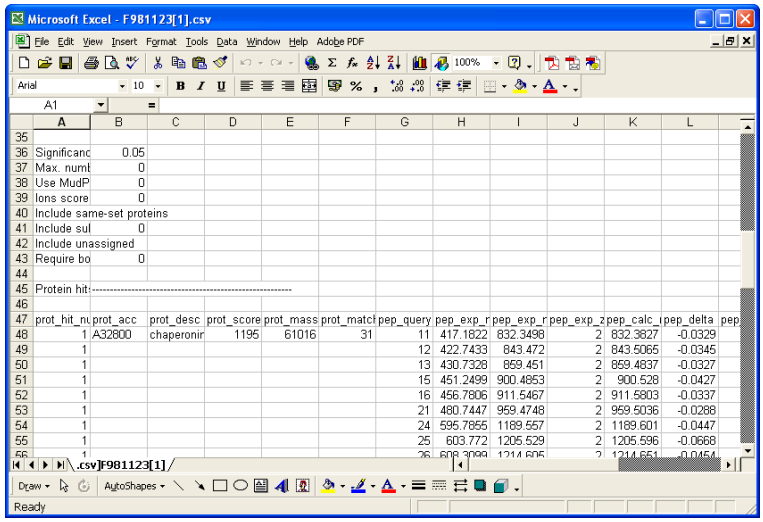


MASCOT : Very large searches MATRIX SCIENCE

To export to Excel, simply select CSV as the format, and click on the Export Search Results button at the bottom of the page.

You can then click on the Open button to open it into Excel:

Search result export

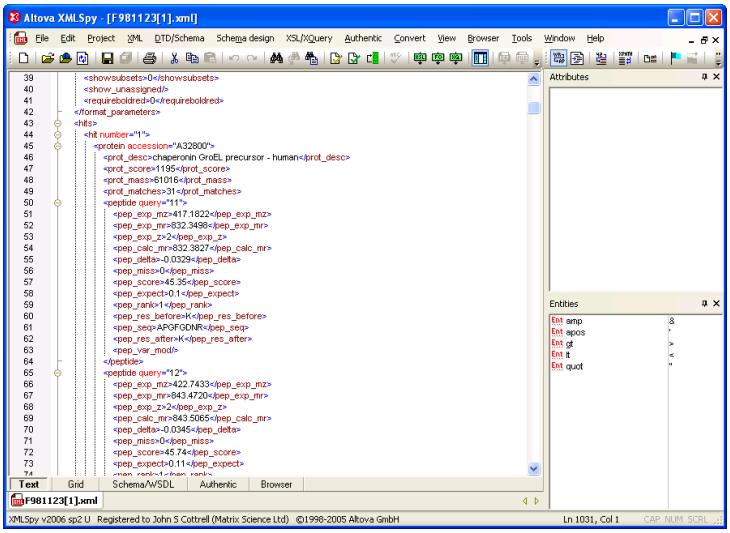


MASCOT : Very large searches

MATRIX SCIENCE

Much easier and safer than “screen scraping”

Search result export

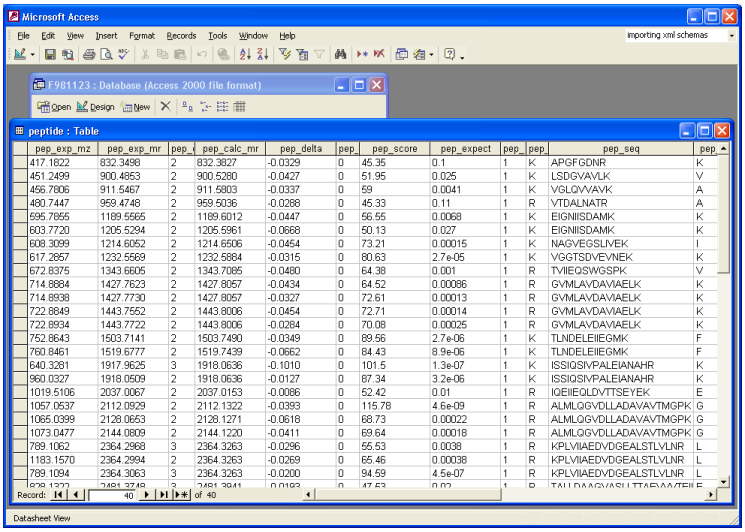


MASCOT : Very large searches MATRIX SCIENCE

For those of you into XML, here is a sample XML file. The schema is available from our web site or your local Mascot installation.

Please read the help for details.

Search result export

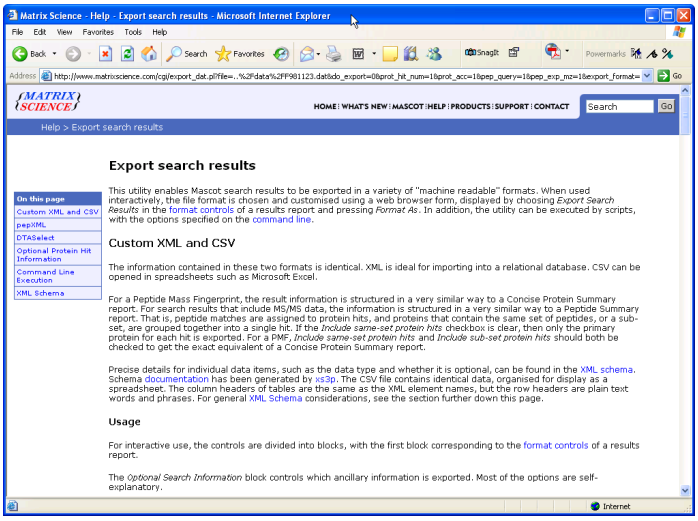


MASCOT : Very large searches

MATRIX SCIENCE

XML is ideal for transferring the results to a relational database. Even Microsoft Access can open the XML file directly into database tables

Search result export



MASCOT : Very large searches

There is a very detailed help page for all of this.

The exports can also be performed from the command line, which means that it is possible to automate the export of searches to xml for example.

Summary

- How to combine the data files
- Performing large searches
- Protein scoring – standard vs. MudPIT
- Viewing the results
- Exporting results to a relational database