

Challenges of very large searches

MASCOT

{MATRIX}
{SCIENCE}

What is 'very large'

- Number of spectra?
- Size of peak lists?
- Size of raw files?
- Size of fasta database?
- Mass of proteins?
- Size of organism
(Blue Whale vs Thailand's bumblebee bat)

MASCOT : Very large searches

© 2008 Matrix Science

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.**

MASCOT : Very large searches

© 2008 Matrix Science



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

Group Assignment	
Minimum precursor mass (Mr)	700
Maximum precursor mass (Mr)	16,000
Precursor m/z tolerance	1
Maximum intermediate scan count	5

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

MASCOT : Very large searches

© 2008 Matrix Science

MATRIX
SCIENCE

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.

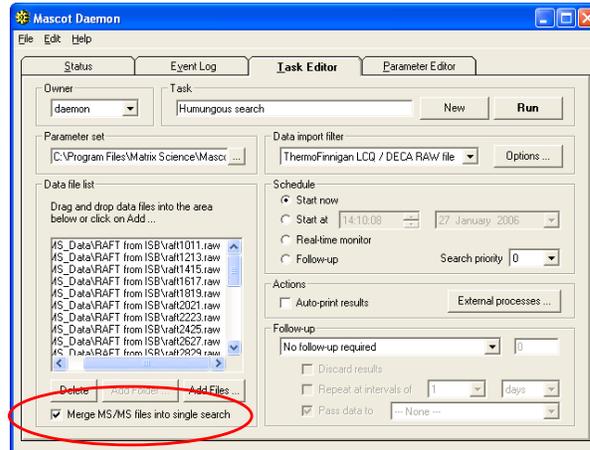
Most data systems have the option to merge ms-ms spectra with similar precursors that have eluted within a specified time window. The advantages of this are improved signal to noise, a smaller peak list and fewer duplicate peptides so reports are less cluttered.

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



MASCOT : Very large searches

© 2008 Matrix Science



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

© 2008 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

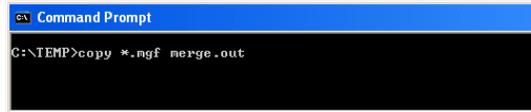
You can't merge mzData files because XML doesn't work like that.

Combining data files

Concatenating peak lists:

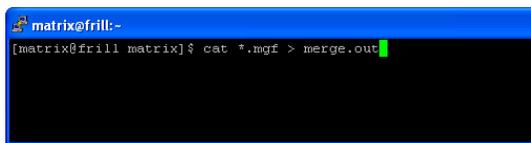
- MGF

Windows: copy



```
Command Prompt
C:\TEMP>copy *.mgf merge.out
```

Unix: cat



```
matrix@frill:~$ cat *.mgf > merge.out
```

MASCOT : Very large searches

© 2008 Matrix Science



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!

MASCOT : Very large searches

© 2008 Matrix Science



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

Submitting large searches - Web server

IIS 5 (Windows 2000) up to 2GB

IIS 5.1 (Windows XP) up to 4GB ??

IIS 6 (Windows 2003) up to 4GB

IIS 7 (Windows Vista / 2008) up to 4GB

Apache 2.0 and earlier up to 2 GB

Apache 2.2 and later - no limit

MASCOT : Very large searches

© 2008 Matrix Science



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.

Submitting large searches

Internet Explorer unable to upload > 2GB

Firefox - unable to upload > 2GB

Third party applications - varies.

MASCOT : Very large searches

© 2008 Matrix Science

MATRIX
SCIENCE

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:

64 Bit Internet Explorer - able to upload > 2GB?

The screenshot shows the Windows XP Administrator interface. The Start menu is open, and 'Internet Explorer (64-bit)' is selected. In the background, a web browser window is open to a page titled 'Cannot find server - Microsoft Internet Explorer'. The address bar shows 'http://frill/mascot/cgi/nph-mascot.exe?1'. The error message reads: 'The page cannot be displayed. The page you are looking for is currently unavailable. The Web site might be experiencing technical difficulties, or you may need to adjust your browser settings.' In the top right corner, there is a form with fields for 'Data file' (set to '\\D_Drive\\large_mgf\\2.3GB.mgf'), 'Data format' (Mascot generic), 'Instrument' (Default), and 'Decoy' (unchecked). A 'Start Search ...' button is visible below these fields.

MASCOT : Very large searches

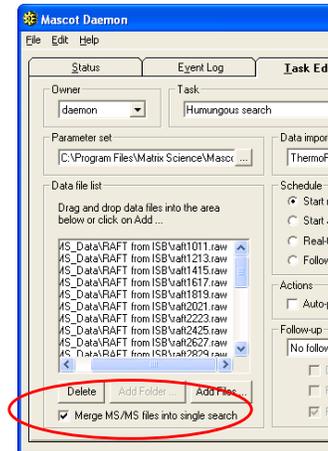
© 2008 Matrix Science



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

© 2008 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.

Performing large searches

32 bit platforms: maximum process size 2GB

Mascot divides large searches into chunks

- mascot.dat:

```
SplitNumberOfQueries 1000
```

```
SplitDataFileSize 10000000
```

Consequences:

- Search size is “unlimited” (except by disk space)
- No protein summary section in result file

Web server timeout normally 1 day (Windows)

MASCOT : Very large searches

© 2008 Matrix Science



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 their 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.

MASCOT : Very large searches

© 2008 Matrix Science



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

Reporting large search results

Essential

Simplifies & reduces memory

Select Summary Report

Format As	Select Summary (protein hits)	Help			
Significance threshold p<	0.05	Max. number of hits	AUTO		
Standard scoring	<input type="radio"/> MudPIT scoring	<input checked="" type="radio"/> Ions score cut-off	0.5	Show sub-sets	<input type="checkbox"/>
Show pop-ups	<input type="radio"/> Suppress pop-ups	<input checked="" type="radio"/> Sort unassigned	Decreasing Score	Require bold red	<input checked="" type="checkbox"/>

Reduces memory

Simplifies

**`http://.../master_results.pl?file=../data/20060202/F123.dat &REPTYPE=select
&_showpopups=FALSE &_requireboldred=1 &_ignoreionsscorebelow=0.5`**

MASCOT : Very large searches

© 2008 Matrix Science

**MATRIX
SCIENCE**

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.

Matrix Science - Help - Results Format - Microsoft Internet Explorer

Address: http://h41-dmc/mascot/help/results_help.html#FORMAT

master_results.pl

URL	mascot.dat	Value	Description
reptype		peptide	Peptide Summary
		archive	Archive Report
		concise	Concise Protein Summary
		protein	Full Protein Summary
		select	Select Summary (hits)
		unassigned	Select Summary (unassigned)
report		auto	Report all significant hits
		N	Report N hits
_showsubsets	ShowSubSets	1	Set value to 1 to report Peptide Summary hits that match a subset of peptides. Default is 0.
_requireboldred	RequireBoldRed	1	Set value to 1 to report Peptide Summary hits only if they contain at least one "bold red" peptide. Default is 0.
_showallfromerrortolerant	ShowAllFromErrorTolerant	1	Set value to 1 to report all hits from an error tolerant search, including the garbage. Default is 0.
_sigthreshold	SigThreshold	N	Probability to use for the significance threshold. Range is 0.1 to 1E-18. Default is 0.05.
_sortunassigned	SortUnassigned	scoredown	Sort unassigned matches by descending score, (default)
		queryup	Sort unassigned matches by ascending query number
		intdown	Sort unassigned matches by descending intensity
_ignoreionscorebelow	IgnoreIonsScoreBelow	N	Any ions scores below this value are set to 0. Floating point number, default 0.0.
_showpopups		true	Show top 10 peptide matches from each query in JavaScript pop-up, (default)
		false	Suppress JavaScript pop-ups.
_alwaysgettitle		1	Set to 1 to force reports to fetch Fasta titles from database when they are not included in the result file. Default is 0.
_mudpit	Mudpit	N	Number of queries at which protein score calculation switches to large search mode. Default 1000

Local intranet

MASCOT : Very large searches © 2008 Matrix Science **MATRIX SCIENCE**

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

Select Summary Report (a8) - Microsoft Internet Explorer

Address: http://t41-jsc/mascot/cgi/master_results.pl?file=R%3A%2Fbrf%2Fdat_MGF%2FA8%2FPII-HUMAN_2_18%2F005625.dat&REPTYPE=select&sigthreshold=0.05&REPORT=AU

Mass values : Monoisotopic
 Protein Mass : Unrestricted
 Peptide Mass Tolerance : ± 4 Da
 Fragment Mass Tolerance : ± 0.6 Da
 Max Missed Cleavages : 2
 Instrument type : ESI-TRAP
 Number of queries : 83316

Select Summary Report

Format As: Select Summary (protein hits) [Help](#)

Significance threshold p < 0.05 Max. number of hits: AUTO

Standard scoring: MudPIT scoring Ions score cut-off: 0.5 Show sub-sets:

Show pop-ups: Suppress pop-ups Sort unassigned: Decreasing Score Require bold red:

1. [IPI00027749](#) Mass: 83423 Score: 3048 Queries matched: 81
 Tax_Id=9606 Heat shock protein HSP 90-beta
 Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
308	362.1600	722.3054	721.4850	0.8204	0	43	0.048	1	VILHLK
341	366.1527	730.2909	729.4384	0.8524	0	44	0.054	1	LSELLR
363	367.7464	733.4782	732.4381	1.0401	0	46	0.034	1	SLVSVTK
4710	520.5880	1039.1615	1038.4869	0.6746	0	60	0.001	1	YESLIDPSK 4712 4719 4770
7141	571.5048	1140.9951	1140.5523	0.4428	0	56	0.0031	1	LGIHEDTNR 7140 7156 7159
7330	576.5980	1151.1814	1150.5506	0.6309	0	61	0.00094	1	YIQEELNK 7345 7361
7544	582.0571	1162.0997	1159.5760	2.5237	0	68	0.00017	1	SIYYITGESK 7528
8145	598.2438	1194.4730	1193.6404	0.8326	0	50	0.01	1	IDIIPHQER 8154
9184	619.1655	1236.3164	1235.6298	0.6866	1	61	0.00078	1	RAPFDLFENK 9175
9337	622.5440	1243.0269	1241.6970	1.3274	0	74	4.3e-005	1	...

MASCOT : Very large searches © 2008 Matrix Science **MATRIX SCIENCE**

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.

Internet Explorer - Does 64 bit help?

MATRIX
SCIENCE

Mascot Search Results

User :
Email :
Search title :
MS data file : D:\data\extract_msn.mgf
Database : SwissProt 51.6 (257964 sequences; 93947433 residues)
Timestamp : 25 Apr 2008 at 02:39:43 GMT
Enzyme : Trypsin
Fixed modifications : Carbamidomethyl (C)
Variable modifications : Oxidation (M)
Mass values : Monoisotopic
Protein Mass : Unrestricted
Peptide Mass Tolerance : ± 4.5 Da
Fragment Mass Tolerance : ± 0.5 Da
Max Missed Cleavages : 1
Instrument type : **ESI-TRAP**
Number of queries : **282651**
Protein hits : [ALB](#)

Browser uses
approx. 4 GB
(2000 proteins)

Image Name	PID	CPU	CPU Time	Mem Usage	Peak Mem Usage	VM Size
ieplora.exe	2076	00	0:35:13	1,663,832 K	1,912,744 K	3,994,820 K
firefox.exe *32	2220	00	0:00:04	12,648 K	37,080 K	27,068 K
vmware.exe *32	2548	00	0:01:48	716 K	46,576 K	24,360 K

MASCOT : Very large searches

© 2008 Matrix Science

MATRIX
SCIENCE

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

???

Select Summary Report			
Format As	Select Summary (protein hits) <input type="button" value="v"/>		Help
Significance threshold p<	<input type="text" value="0.05"/>	Max. number of hits	<input type="text" value="AUTO"/>
Standard scoring	<input type="radio"/> MudPIT scoring <input checked="" type="radio"/>	Ions score cut-off	<input type="text" value="0.5"/> Show sub-sets <input type="checkbox"/>
Show pop-ups	<input type="radio"/> Suppress pop-ups <input checked="" type="radio"/>	Sort unassigned	<input type="text" value="Decreasing Score"/> <input type="button" value="v"/> Require bold red <input checked="" type="checkbox"/>

MASCOT : Very large searches

© 2008 Matrix Science

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. [IP100141647](#) 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	GGIQDIAK
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

© 2008 Matrix Science

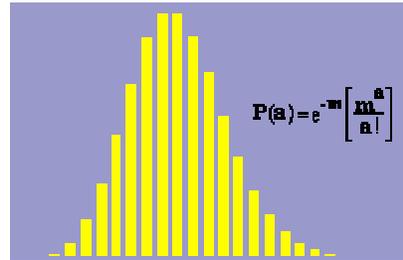


With standard peptide summary scoring, the protein score is essentially the sum of the ions scores of all the non-duplicate peptides. Where there are duplicate peptides, the highest scoring peptide is used. A correction is applied based on the number of candidate peptides that were tested. This correction 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

© 2008 Matrix Science

MATRIX
SCIENCE

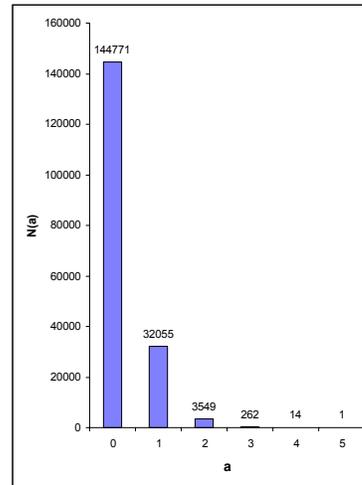
Even if peptide matches are random, you can still get multiple matches to a single protein. How likely this is depends on the ratio between the number of spectra and the number of entries in the database. We can predict whether this will be a serious problem or not using a function called a Poisson distribution.

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

© 2008 Matrix Science

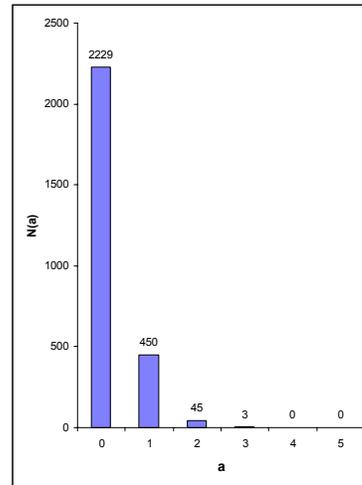


For this MudPIT search, 262 proteins are expected to pick up 3 random matches by chance. 1 protein will pick up 5

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

© 2008 Matrix Science

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. [IP100001639](#) **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	TVSPDRLEAAQK
1359	918.3068	1834.5991	1832.8839	1.7152	0	46	0.024	1	GDQENVHPDVMLVQPR

MASCOT : Very large searches

© 2008 Matrix Science

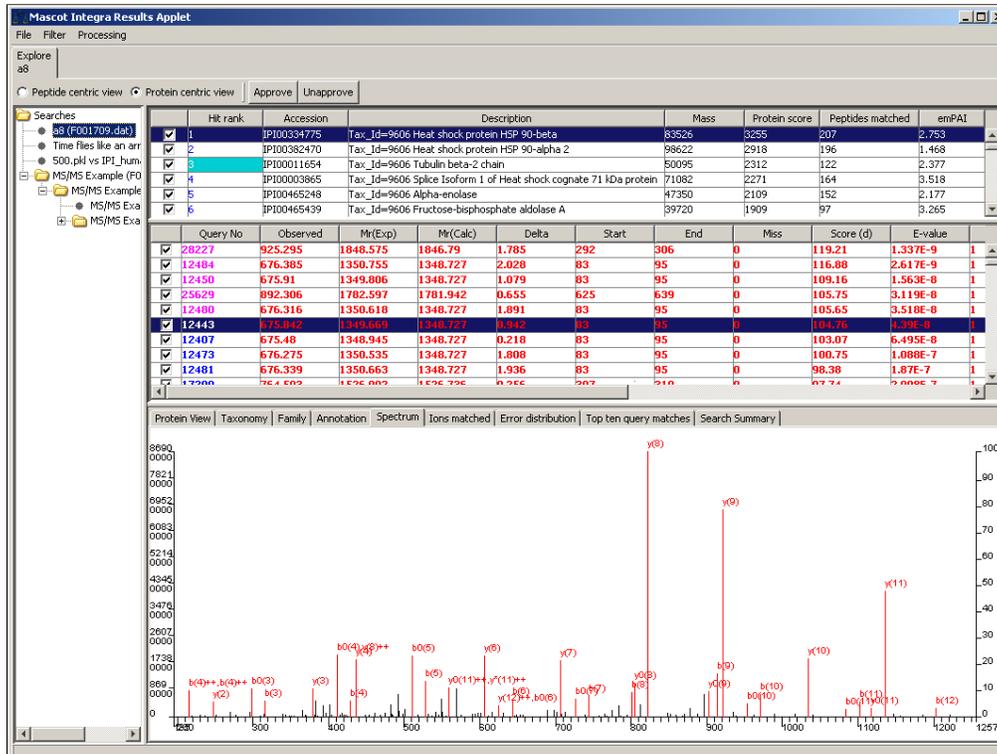


For MudPIT scoring, the score for each peptide is not its absolute score, but the amount that it is above the threshold. Therefore, peptides with a score below the threshold do not contribute to the score. Finally, the average of the thresholds used is added to the score. For each peptide, the "threshold" 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.

Scaffold

- **Version 2.0 - 5 to 10 million LTQ spectra on desktop computers with 2GB of RAM**
- **GUI slows down much beyond 8-10 million spectra**
- **Limit is 1 to 2 million Q-ToF spectra because the identification rate is substantially higher**
- **Could possibly load 5 to 10 6GB result files**

MASCOT : Very large searches

© 2008 Matrix Science

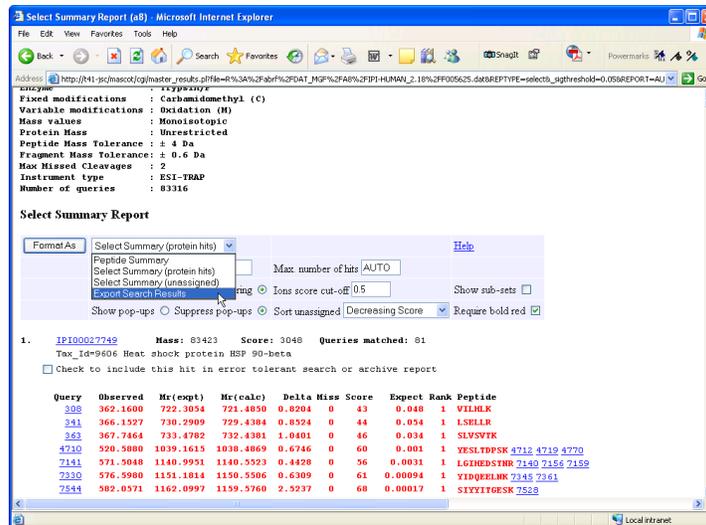


Another alternative is to use Scaffold. The new version 2 copes with much larger results file and can load 5 to 10 million LTQ spectra.

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

© 2008 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

Matrix Science - Mascot - Export search results - Microsoft Internet Explorer

Address: http://www.matrixscience.com/cgi/export_data.pl?file=...%2Fdata%2F981123.dat&REPORT=export_0.05&REPORT=AUTO&_mudpit=99999999_ignorescon

Mascot > Export search results

Export search results [Help](#)

Export format: XML (dropdown menu open showing XML, CSV, pepXML, DTASelect)

Significance threshold p<

Max. number of hits

Protein scoring: Standard (selected), MudPIT

Ions score cut-off: 0

Include same-set protein hits (additional proteins that span the same set of peptides)

Include sub-set protein hits (additional proteins that span a sub-set of peptides)

Require bold red

Optional Search Information

Header

Search parameters

Format parameters

Residue masses

Done Internet

MASCOT : Very large searches

© 2008 Matrix Science



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

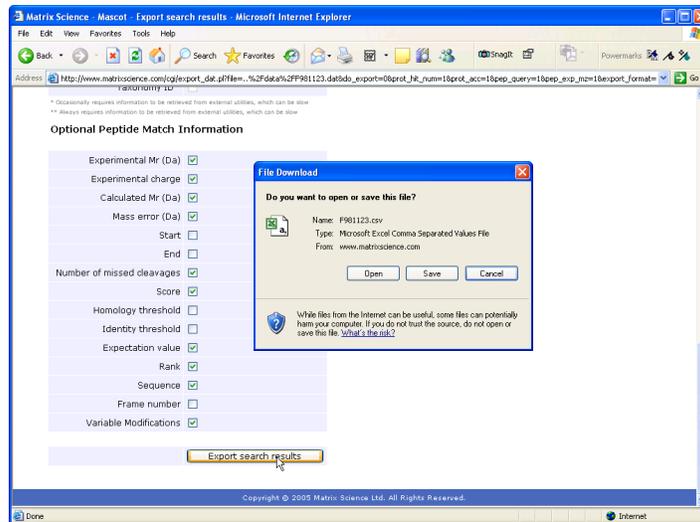
If you want the XML format, you probably know that this is what you want. If you've no idea what XML is, chances are you don't want it.

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

© 2008 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

47	prot_hit	prot_acc	prot_desc	prot_score	prot_mass	prot_match	pep_query	pep_exp_r	pep_exp_r	pep_exp_r	zpep_calc	pep_delta	pep
48	1	A32600	chaperonin	1195	61016	31	11	417.1822	832.3498	2	832.3827	-0.0329	
49	1						12	422.7433	843.472	2	843.5065	-0.0345	
50	1						13	430.7328	859.451	2	859.4837	-0.0327	
51	1						15	451.2499	900.4853	2	900.528	-0.0427	
52	1						16	456.7806	911.5467	2	911.5803	-0.0337	
53	1						21	480.7447	959.4748	2	959.5036	-0.0288	
54	1						24	595.7855	1189.557	2	1189.601	-0.0447	
55	1						25	603.772	1205.529	2	1205.596	-0.0668	
56	1						26	608.3999	1214.605	2	1214.651	-0.0454	

MASCOT : Very large searches

© 2008 Matrix Science



Much easier and safer than “screen scraping”

Search result export

```
39 <showsubsets=0</showsubsets>
40 <show_lines=ped>
41 <requireboldred=0</requireboldred>
42 </format_parameters>
43 <file>
44 <hit number="1">
45 <protein accession="A32800">
46 <prot_desc=chaperonin GroEL precursor - human</prot_desc>
47 <prot_access=1189</prot_access>
48 <prot_mass=61016</prot_mass>
49 <prot_matches=31</prot_matches>
50 <peptide query="11">
51 <pep_exp_mz=417.1822</pep_exp_mz>
52 <pep_exp_mr=832.3498</pep_exp_mr>
53 <pep_exp_z=2</pep_exp_z>
54 <pep_calc_mr=832.3827</pep_calc_mr>
55 <pep_delta=-0.0329</pep_delta>
56 <pep_mis=0</pep_mis>
57 <pep_score=45.25</pep_score>
58 <pep_expect=0.1</pep_expect>
59 <pep_rank=1</pep_rank>
60 <pep_res_before=K</pep_res_before>
61 <pep_seq=AKGFQDNR</pep_seq>
62 <pep_res_after=K</pep_res_after>
63 <pep_var_mods>
64 </peptide>
65 <peptide query="12">
66 <pep_exp_mz=422.7433</pep_exp_mz>
67 <pep_exp_mr=843.4720</pep_exp_mr>
68 <pep_exp_z=2</pep_exp_z>
69 <pep_calc_mr=843.5065</pep_calc_mr>
70 <pep_delta=-0.0345</pep_delta>
71 <pep_mis=0</pep_mis>
72 <pep_score=45.74</pep_score>
73 <pep_expect=0.11</pep_expect>
74 </peptide>
75 </hit>
76 </file>
```

MASCOT : Very large searches

© 2008 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

pep_exp_mz	pep_exp_mr	pep_calc_mr	pep_delta	pep_score	pep_expect	pep_seq	pep
417.1822	832.3498	832.3827	-0.0329	0	45.35	0.1	1 K APGFQDNR
451.2499	900.4863	900.5280	-0.0427	0	51.95	0.025	1 K LSDGVAVLK
456.7906	911.5467	911.5803	-0.0337	0	59	0.0041	1 K VGLQVAVK
460.7447	959.4748	959.5036	-0.0289	0	45.33	0.11	1 R YTDALNATR
595.7855	1189.5565	1189.6012	-0.0447	0	56.55	0.0069	1 K EIGNIISDAMK
603.7720	1205.5294	1205.5961	-0.0668	0	50.13	0.027	1 K EIGNIISDAMK
608.3099	1214.6052	1214.6506	-0.0454	0	73.21	0.00015	1 K NAGVGGSLVEK
617.2657	1232.5569	1232.5884	-0.0315	0	80.63	2.7e-05	1 K VGGTSDVEVNEK
672.8375	1343.6605	1343.7085	-0.0480	0	64.38	0.001	1 R TVIEGQSWGSPK
714.8894	1427.7623	1427.8057	-0.0434	0	64.52	0.00086	1 R GVMLAVDAVIAELK
714.8938	1427.7730	1427.8057	-0.0327	0	72.61	0.00013	1 R GVMLAVDAVIAELK
722.8849	1443.7552	1443.8006	-0.0454	0	72.71	0.00014	1 R GVMLAVDAVIAELK
722.8934	1443.7722	1443.8006	-0.0284	0	70.08	0.00025	1 R GVMLAVDAVIAELK
752.8643	1503.7141	1503.7490	-0.0349	0	89.56	2.7e-06	1 K TLNDELEIEGMK
760.8461	1519.6777	1519.7439	-0.0662	0	84.43	8.9e-06	1 K TLNDELEIEGMK
840.3281	1817.9625	1818.0636	-0.1010	0	101.5	1.3e-07	1 K ISSIGSIVPALEIANHR
960.0327	1918.0609	1918.0636	-0.0127	0	87.34	3.2e-06	1 K ISSIGSIVPALEIANHR
1019.5106	2037.0067	2037.0163	-0.0086	0	52.42	0.01	1 R IGEIEQLDYTSEYEK
1057.0537	2112.0529	2112.1322	-0.0393	0	115.78	4.6e-09	1 R ALMLGGVLLADAVAVTMGPK
1065.0399	2128.0653	2128.1271	-0.0618	0	88.73	0.00022	1 R ALMLGGVLLADAVAVTMGPK
1073.0477	2144.0809	2144.1220	-0.0411	0	89.64	0.00018	1 R ALMLGGVLLADAVAVTMGPK
789.1052	2364.2968	2364.3263	-0.0296	0	55.53	0.0038	1 R KPLVIAEDVDGEALSTLVNLR
1183.1570	2364.2994	2364.3263	-0.0269	0	65.46	0.00038	1 R KPLVIAEDVDGEALSTLVNLR
789.1094	2364.3063	2364.3263	-0.0200	0	94.59	4.5e-07	1 R KPLVIAEDVDGEALSTLVNLR
1678.1571	3461.3248	3461.3841	-0.0103	0	47.63	0.03	1 D TALLDAQVAVSLTASAAVTEK

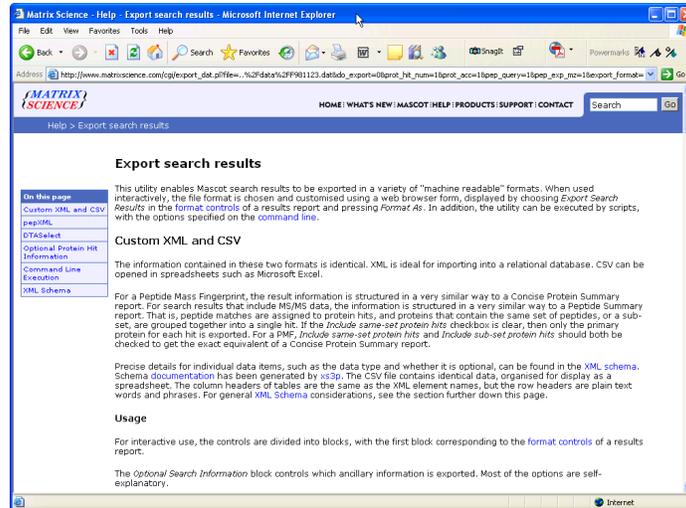
MASCOT : Very large searches

© 2008 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

© 2008 Matrix Science



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