

# Very Large Searches

MASCOT

*{MATRIX}*  
*{SCIENCE}*

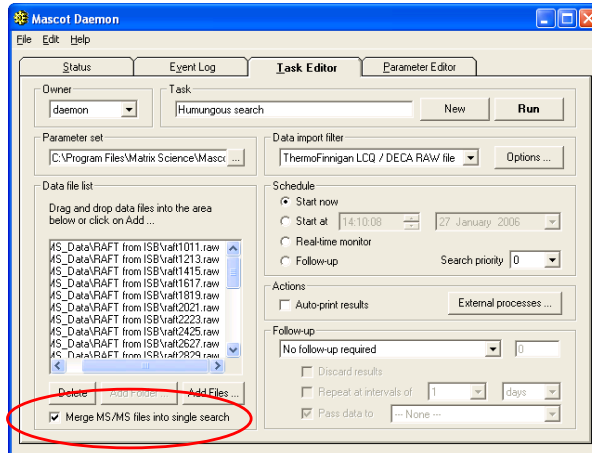
## Topics

- Combining data files
- Performing large searches
- The Protein Family summary
- Protein scoring - standard vs. MudPIT
- Exporting results to a relational database

Very large searches present a number of challenges. These are the topics we will cover during this presentation.

## 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

© 2007-2010 Matrix Science



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, although note that some web servers cannot accept uploads larger than 4 GB

## 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](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*

© 2007-2010 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.

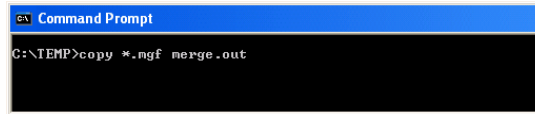
A nice feature of this script is that it puts the filename into the scan title, so you can tell which fraction a particular spectrum came from. The scan titles are displayed in the yellow pop-ups on the Mascot result report

## Data files

### Concatenating peak lists:

- MGF

Windows: copy



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

Unix: cat



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

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. Because a number of third party utilities add commands to MGF headers, and these cause a merged search to fail, Mascot Daemon 2.3 now strips out header lines when merging MGF files.

## 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

## 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

**MASCOT** : *Very Large Searches*

© 2007-2010 Matrix Science



32 bit platforms, like many 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. However, some old client software gets confused by the missing section. The work around is to increase the values so that large searches never split. Maybe setting `SplitNumberOfQueries` to 1 million spectra and `SplitDataFileSize` to 10 billion bytes.

This is OK, but remember to reset these values as soon as you are able to. Otherwise, you might find you run out of memory or address space for your large searches

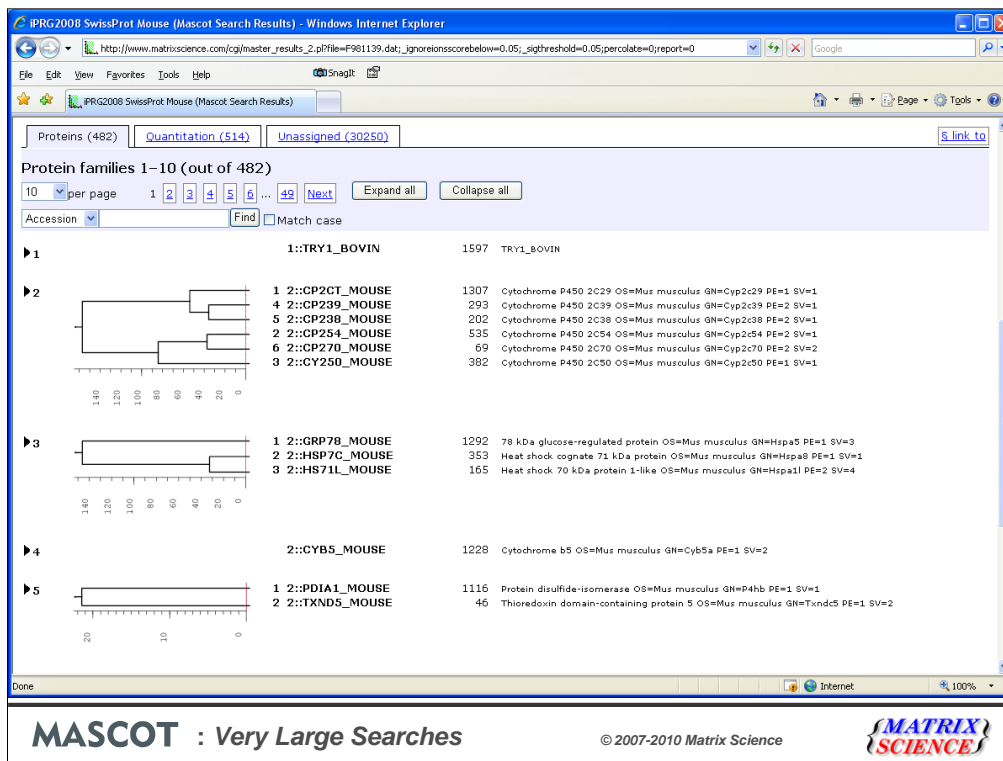
## Reporting large searches

### Protein Family Summary (new in Mascot 2.3)

- Paged report to conserve memory
- Detailed information is shown 'on demand'
- Index files are created and cached to speed loading in future
- Proteins grouped into families by means of shared peptide matches
- Hierarchical clustering within each protein family

In Mascot 2.2 and earlier, trying to display result reports for very large searches would often lead to problems with timeouts and running out of memory. To address this, we have developed a brand new report that loads most of the information 'on demand': the Protein Family Summary. This requires some index files to be created on the server, and these index files are cached, so that the report loads much faster on the second and subsequent occasions. Proteins are grouped into families by means of shared peptide matches and, within each family, hierarchical clustering is used to illustrate which proteins are closely related and which are more distant.





If there are 300 or more spectra, the Family Summary is the default. This is the appearance of a typical family report immediately after loading. The body of the report consists of three tabs, one for protein families, one for quantitation, and one for unassigned matches. The report is paged, with a default page size of 10 families. If you wish, you can choose to display a larger number of families on a single page.

Proteins are grouped into families using a novel hierarchical clustering algorithm. If the family contains a single member, the accession string, protein score and description are listed. If the family contains multiple members, the accessions, scores and descriptions are aligned with a dendrogram, which illustrates the degree of similarity between members.

The scores for the proteins in family 2 vary from 1307 down to 69. In the earlier Peptide Summary or Select Summary reports, these would have been at opposite ends of the report. It would have been difficult to recognise that these proteins belonged together, even though they have shared peptide matches and are all cytochrome P450 2C proteins.

IPRG2008 SwissProt Mouse (Mascot Search Results) - Windows Internet Explorer

http://61-hs/mascot/cgi/master\_results\_2.pl?file=P81139.dat;\_ignoreinsscorebelow=0.05;\_sigthreshold=0.05;percentat=0;report=0#3;pr.eh=%282-3%29p,2-3

File Edit View Favorites Tools Help

IPRG2008 SwissProt Mouse (Mascot Search Results)

1 2::CP2CT\_MOUSE 1307 Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=1  
 4 2::CP239\_MOUSE 293 Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=1  
 5 2::CP238\_MOUSE 202 Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=1  
 2 2::CP254\_MOUSE 535 Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1  
 6 2::CP270\_MOUSE 69 Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2  
 3 2::CY250\_MOUSE 382 Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=1

Threshold (0): 0 Cut

	Score	Mass	Matches	Sequences	emPAI	
2.1 2::CP2CT_MOUSE	1307	61433	86 (86)	13 (13)	1.47	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=1
2.2 2::CP254_MOUSE	535	60887	29 (29)	10 (10)	0.87	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
2.3 2::CY250_MOUSE	382	61037	25 (25)	10 (10)	0.87	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=1
2.4 2::CP239_MOUSE	293	60932	24 (24)	6 (6)	0.41	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=1
2.5 2::CP238_MOUSE	202	61216	20 (20)	6 (6)	0.40	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=1
2.6 2::CP270_MOUSE	69	61539	5 (5)	4 (4)	0.25	Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2

Redisplay All None

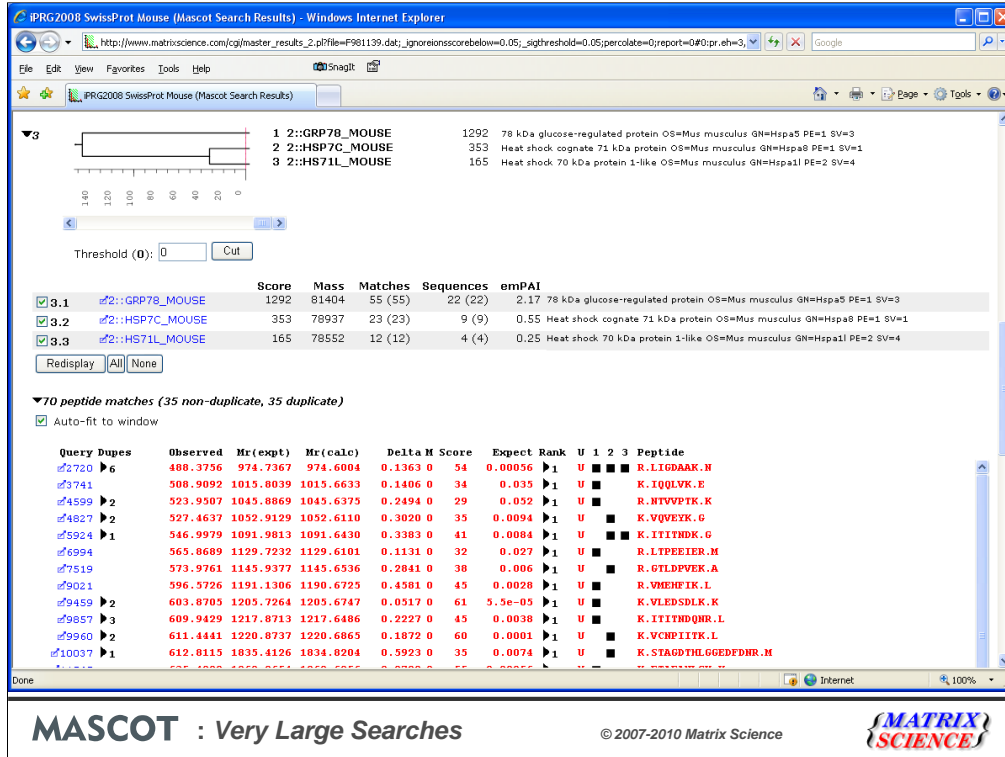
119 peptide matches (35 non-duplicate, 84 duplicate)  
 Auto-fit to window

Query	Dupes	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Exprot	Rank	U	1	2	3	4	5	6	Peptide
4431	9	503.8391	1005.6637	1005.6103	0.0533	0	35	0.831	1	U						R.LVFDK.A
4447	2	520.8626	1039.7106	1039.6157	0.0949	0	40	0.0079	1	U						SYLEK
4466		521.2416	1560.7029	1559.8187	0.8842	0	59	0.00021	1	U						K.MISQFTWFSK.A
4705		521.3753	1040.7361	1040.5810	0.1551	0	22	0.031	1	U						R.FTLMTLR.N + Oxidation (O)
4791	3	525.4566	1573.3479	1572.7634	0.5824	0	71	1.4e-005	1	U						K.EALVDHGEFAGR.G
5544	8	526.2961	1050.3776	1050.5323	0.0453	0	35	0.0083	1	U						R.CLVEELR.K
5605	5	540.3247	1078.6349	1078.5305	0.0964	0	54	0.00028	1	U						R.IGAGEGLAR.M
5708		541.3848	1080.7551	1080.6059	0.1492	0	53	0.00047	1	U						K.YPDVTAQ.V
		574.8892	1147.7638	1147.6117	0.1523	0	25	0.05	1	U						K.FPDDEK.F

Done Local intranet 100%

**MASCOT : Very Large Searches** © 2007-2010 Matrix Science **MATRIX SCIENCE**

If you are interested in family 2, then you click to expand it to show the details. Immediately under the dendrogram is a list of the proteins. The table of peptide matches is similar to that found in the other result reports. Duplicate matches to the same sequence are collapsed into a single row. The columns headed 1, 2, 3, etc. represent the proteins and contain a black square if the peptide is found in the protein. Some matches are shared, but each protein has some unique peptide matches, otherwise it would be dropped as a sub-set.



Moving down to family 3, the scale on the dendrogram is ions score, and HSP7C\_MOUSE and HS71L\_MOUSE join at a score of approximately 30. This represents the score of the significant matches that would have to be discarded in order to make one protein a sub-set of the other. These two proteins are much more similar to one other than to GRP78\_MOUSE, which has non-shared peptide matches with a total score of approximately 145. Note that, where there are multiple matches to the same peptide sequence, (ignoring charge state and modification state), it is the highest score for each sequence that is used.

Immediately under the dendrogram is a list of the proteins. In this example, because SwissProt has low redundancy, each family member is a single protein. In other cases, a family member will represent multiple same-set proteins. One of the proteins is chosen as the anchor protein, to be listed first, and the other same-set proteins are collapsed under a same-set heading. There is nothing special about the protein picked for the anchor position. You may have a preference for one according to taxonomy or description, but all proteins in a same-set group are indistinguishable on the basis of the peptide match evidence.

The table of peptide matches is similar to that found in the other result reports. Duplicate matches to the same sequence are collapsed into a single row. Click on the triangle to expand.

The black squares to the right show which peptides are found in which protein. To see the peptides that distinguish HSP7C\_MOUSE and HS71L\_MOUSE, clear the checkbox for GRP78\_MOUSE and choose Redisplay.

**IPRG2008 SwissProt Mouse (Mascot Search Results) - Windows Internet Explorer**

http://www.matrixscience.com/cgi/master\_results\_2.pl?file=F981139.dat;ignoreinscorebelow=0.05;sigthreshold=0.05;percolate=0;report=0#3;pr.eh=3;

1 2::GRP78\_MOUSE 1292 78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3  
 2 2::HSP7C\_MOUSE 353 Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1  
 3 2::HS71L\_MOUSE 165 Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa11 PE=2 SV=4

Threshold (0): 0 [Cut]

	Score	Mass	Matches	Sequences	eM-PAI
<input type="checkbox"/> 3.1	1292	81404	55 (55)	22 (22)	2.17 78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
<input checked="" type="checkbox"/> 3.2	353	78937	23 (23)	9 (9)	0.55 Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
<input checked="" type="checkbox"/> 3.3	165	78552	12 (12)	4 (4)	0.25 Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa11 PE=2 SV=4

Redisplay [All] [None]

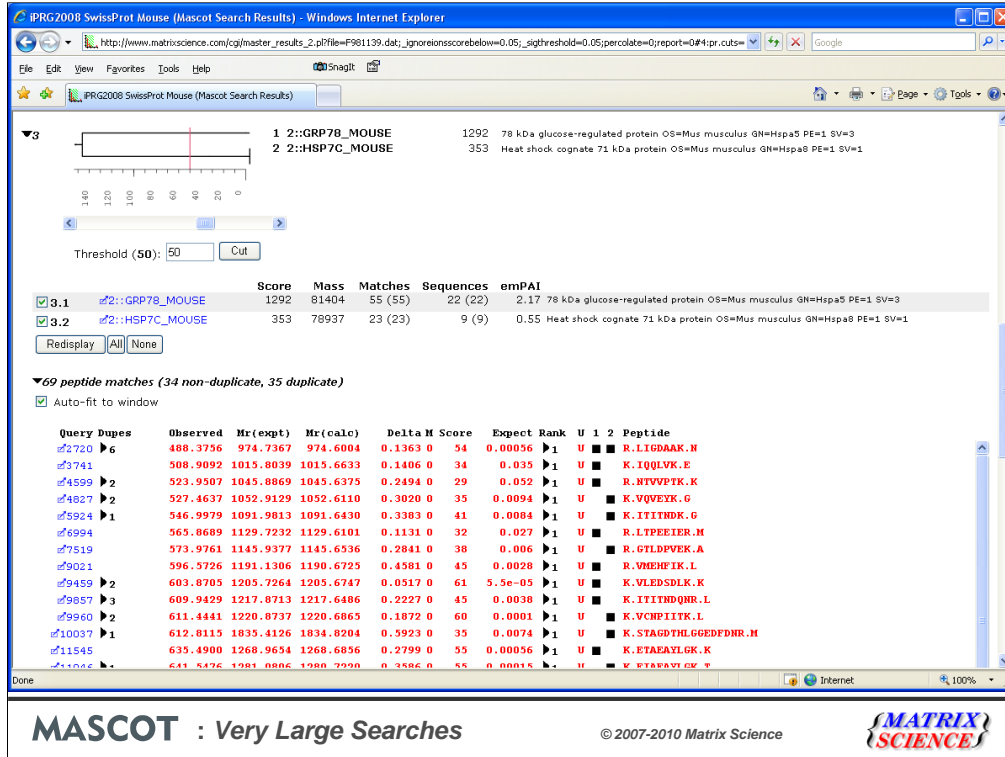
▼ 24 peptide matches (11 non-duplicate, 13 duplicate)  
 Auto-fit to window

Query Dupes	Observed	Mr(expt)	Mr(calc)	Delta M Score	Expect Rank	U	2	3	Peptide
2720 6	488.3756	974.7367	974.6004	0.1363 0	54	0.00056	1	U	R.LIGDAAK.N
4827 2	527.4637	1052.9129	1052.6110	0.3020 0	35	0.0094	1	U	K.VQVEYK.G
5924 1	546.9979	1091.9813	1091.6430	0.3383 0	41	0.0084	1	U	K.IYTRDK.G
7519	573.9761	1145.9377	1145.6536	0.2841 0	38	0.006	1	U	R.GYLDPEYK.A
9960 2	611.4441	1220.8737	1220.6865	0.1872 0	60	0.0001	1	U	K.VCHPIITK.L
10037 1	612.8115	1835.4126	1834.8204	0.5923 0	35	0.0074	1	U	K.STAGDTHLGGEDFDNR.M
11946 1	641.5476	1281.0806	1280.7220	0.3586 0	55	0.00015	1	U	K.EIAEAYLQK.T
25277	607.4422	1819.3048	1818.8255	0.4793 0	55	3.2e-05	1	U	K.ATAGDTHLGGEDFDNR.L
6376	953.0936	1904.1726	1903.9845	0.1881 0	84	1.3e-07	1	U	K.SFYPEEVSSMVLTK.M
6946	650.1325	1947.3756	1947.0920	0.2836 0	37	0.013	1	U	R.IINEPTAAAIAYGLDK
6947	974.7142	1947.4139	1947.0920	0.3218 0	43	0.00059	1	U	R.IINEPTAAAIAYGLDK

**MASCOT : Very Large Searches** © 2007-2010 Matrix Science **MATRIX SCIENCE**

It can now be seen that HS71L\_MOUSE would be a sub-set of HSP7C\_MOUSE if it was not for one match, K.ATAGDTHLGGEDFDNR.L. It is the significant score for this match that separates the two proteins in the dendrogram by a distance of 32 (score of 55 - homology threshold score of 23).

You can "cut" the dendrogram using the slider control.



If we cut the dendrogram at a score of 50, HS71L\_MOUSE will be dropped because it is now a sub-set protein. If you compare the matches to HSP7C\_MOUSE with those to GRP78\_MOUSE, it is clear that these are very different proteins. They are part of the same family because of two shared matches, but many highly significant matches would have to be discarded for either protein to become a sub-set of the other. In summary, we can quickly deduce from the Family Summary that there is abundant evidence that both GRP78\_MOUSE and HSP7C\_MOUSE were present in the sample. There is little evidence for HS71L\_MOUSE. It is more likely that the HSP7C\_MOUSE contained a SNP or two relative to the database sequence.

IPRG2008 SwissProt Mouse (Mascot Search Results) - Windows Internet Explorer

http://www.matrixscience.com/cgi/master\_results\_2.pl?file=F981139.dat;ignoreinscorebelow=0.05;sigthreshhold=0.05;percolate=0;report=0#9;pr.ed=1

Proteins (482)    Quantitation (514)    Unassigned (30250)    [S link to](#)

Protein families 41-50 (out of 482)

10 per page    [Previous](#)    [1](#)    [2](#)    [3](#)    [4](#)    [5](#)    [6](#)    [7](#)    [8](#)    [9](#)    [10](#)    ...    [49](#)    [Next](#)    [Expand all](#)    [Collapse all](#)

Sequence          Match case   

►41 2::RS3\_MOUSE 353 40S ribosomal protein S3 OS=Mus musculus GN=Rps3 PE=1 SV=1

►42 2::RL22\_MOUSE 347 60S ribosomal protein L22 OS=Mus musculus GN=Rpl22 PE=2 SV=2

▼43 2::RS15A\_MOUSE 344 40S ribosomal protein S15a OS=Mus musculus GN=Rps15a PE=2 SV=2

43.1 [11790](#) 2::RS15A\_MOUSE    **Score**    **Mass**    **Matches**    **Sequences**    **empAI**  
 344    16651    16 (16)    3 (3)    1,24 40S ribosomal protein S15a OS=Mus musculus GN=Rps15a PE=2 SV=2

▼16 peptide matches (4 non-duplicate, 12 duplicate)

Auto-fit to window

Query	Dupes	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	U	Peptide
<a href="#">11708</a>	►5	508.3777	1014.7407	1014.6308	0.1100	45	0.00053	►1	U	K.IVVHLTGR.L
<a href="#">11285</a>	▼5	631.9663	1261.9180	1261.7308	0.1872	77	2.3e-06	►1	U	R.MNVLADALK.S
<a href="#">11274</a>	▼5	631.8868	1261.7591	1261.7308	0.0284	0 (66)	1.8e-05	►1	U	R.MNVLADALK.S
<a href="#">11275</a>	▼5	631.8914	1261.7682	1261.7308	0.0375	0 (59)	9.4e-05	►1	U	R.MNVLADALK.S
<a href="#">11283</a>	▼5	631.9416	1261.8686	1261.7308	0.1379	0 (59)	0.00012	►1	U	R.MNVLADALK.S
<a href="#">11287</a>	▼5	632.0080	1262.0014	1261.7308	0.2706	0 (42)	0.0063	►1	U	R.MNVLADALK.S
<a href="#">11288</a>	▼5	632.0218	1262.0291	1261.7308	0.2983	0 (63)	6.2e-05	►1	U	R.MNVLADALK.S
<a href="#">11604</a>	►1	636.4751	1270.9355	1270.6904	0.2452	0 (28)	0.03	►1	U	K.WQHLLPSR.Q
<a href="#">11780</a>	▼2	639.8954	1277.7762	1277.7237	0.0505	0 (50)	0.00081	►1	U	R.MNVLADALK.S + Oxidation (R)
<a href="#">11790</a>	▼2	639.9899	1277.9652	1277.7257	0.2396	0 (48)	0.00054	►1	U	R.MNVLADALK.S + Oxidation (R)

Done    Internet    100%

**MASCOT** : Very Large Searches    © 2007-2010 Matrix Science    **MATRIX SCIENCE**

The family report also includes a text search facility, which is particularly important for a paged report. You can search by accession or description sub-string, or by query, mass or sequence. Here, for example, we searched for a peptide sequence. The display jumps to the first instance of the sequence, expands, and highlights (in green) the target peptides.

## Large search results in 2.2 and earlier

The screenshot shows the 'Select Summary Report' form with several callouts in yellow boxes:

- Never Peptide**: points to the 'Select Summary (protein hits)' dropdown menu.
- Important**: points to the 'Max. number of hits' input field, which is set to 'AUTO'.
- Simplifies**: points to the 'Ions score cut-off' input field, which is set to '0.5'.
- Reduces memory**: points to the 'Show pop-ups' radio button, which is set to 'Suppress pop-ups'.
- Simplifies**: points to the 'Require bold red' checkbox, which is checked.

Other visible options include: 'Format As', 'Significance threshold p < 0.05', 'Standard scoring' (radio buttons for Standard, MudPIT, Ions), 'Sort unassigned' (dropdown set to 'Decreasing Score'), and a 'Help' link circled in red.

```
http://.../master_results.pl?file=../data/20060202/F000123.dat
&REPTYPE=select &REPORT=AUTO &_showpopups=FALSE
&_ignoreionsscorebelow=0.5 &_requireboldred=1
```

**MASCOT** : *Very Large Searches*

© 2007-2010 Matrix Science

**MATRIX**  
**SCIENCE**

If you are still using Mascot 2.2 or if you have some application software that requires the results in the earlier format, and you are encountering problems with timeouts and running out of memory, here are some tips:

- 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`
- Don't specify a huge number of hits 'just in case'. Choose AUTO to display all protein hits that contain at least one significant peptide match: `&REPORT=AUTO`
- Get rid of the yellow pop-ups: `&_showpopups=FALSE`
- 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](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** © 2007-2010 Matrix Science **MATRIX SCIENCE**

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



## Reporting large search results

???

Select Summary Report			
Format As	Select Summary (protein hits) <input type="button" value="v"/>		<a href="#">Help</a>
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*

© 2007-2010 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

342. [2::IP10023283](#) Mass: 3832803 Score: 181 Matches: 51(0) Sequences: 48(0)  
 Tax\_id=9606 Gene\_Symbol=TTN Isoform 2 of Titin

Query	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Score	Expect	Rank	Unique	Peptide
<a href="#">28</a>	359.7341	717.4537	717.4537	-0.09	0	7	4.2	5	U	R.LFAIVR.G
<a href="#">209</a>	394.2371	786.4596	786.4599	-0.46	0	8	13	3	U	K.LTIADVR.A
<a href="#">334</a>	411.2073	820.4000	820.3954	5.61	0	3	15	4	U	K.TDSGLVR.C
<a href="#">357</a>	413.2642	824.5139	824.5135	0.48	1	12	1.1	5	U	K.RFLLLE.K
<a href="#">715</a>	450.7365	899.4584	899.4588	-0.38	0	10	2.9	2	U	K.IVDVSSDR.C
<a href="#">740</a>	<b>451.7681</b>	<b>901.5217</b>	<b>901.5233</b>	<b>-1.72</b>	<b>0</b>	<b>3</b>	<b>24</b>	<b>3</b>	<b>U</b>	<b>R.VTLVDVTR.N</b>
<a href="#">840</a>	459.2484	916.4821	916.4767	5.98	0	2	29	2	U	K.GVEFNPR.L
<a href="#">844</a>	459.7299	917.4452	917.4454	-0.24	0	4	15	6	U	K.ELEETAAR.M
<a href="#">1029</a>	<b>473.2757</b>	<b>944.5368</b>	<b>944.5331</b>	<b>3.97</b>	<b>1</b>	<b>3</b>	<b>21</b>	<b>3</b>	<b>U</b>	<b>R.EPPSEIKK.I</b>
<a href="#">1058</a>	475.7505	949.4864	949.4869	-0.47	0	4	22	5	U	R.SSVLSLWGR.P
<a href="#">1066</a>	<b>476.2790</b>	<b>950.5433</b>	<b>950.5425</b>	<b>0.94</b>	<b>0</b>	<b>1</b>	<b>23</b>	<b>4</b>	<b>U</b>	<b>R.PLTDLQVR.E</b>

**MASCOT** : *Very Large Searches*

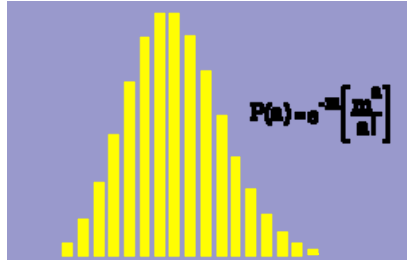
© 2007-2010 Matrix Science



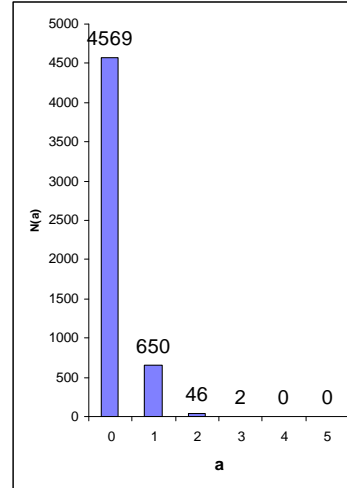
With standard peptide summary scoring, the protein score is essentially the sum of the ions scores of all the non-duplicate peptides. Where there are duplicate peptides, the highest scoring peptide is used. A correction is applied based on the number of candidate peptides that were tested. This correction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

Despite this correction, as this example shows, when we have several low scoring matches assigned to the same protein, we can still get a high protein score, even though none of the individual peptide matches are significant

## Protein Inference



- Huge MudPIT data set
- Search Swiss-Prot using drosophila taxonomy filter (5268 entries)
- 75,000 matches with 1% FDR
- i.e. 750 false matches



**MASCOT** : *Very Large Searches*

© 2007-2010 Matrix Science

**MATRIX**  
**SCIENCE**

A protein with matches to just a single peptide sequence is commonly referred to as a “one-hit wonder” and is often treated as suspect. This is actually a slight oversimplification. In a search with a large number of spectra and a small database, even though the peptide false discovery rate is low, a protein can pick up multiple false matches by chance. This is easily calculated using a Poisson Distribution, where  $m$  is the average number of false matches per protein. In this example,  $m$  is  $750/5268$ , and we would expect 650 database entries to be one-hit wonders. However, 46 entries will pick up two false matches and 2 entries will pick up three, which could mean we report 48 false proteins.

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

```
1249. 2::IPI00023283  Mass: 3832803  Score: 0  Matches: 51(0)  Sequences: 48(0)
Tax_Id=9606 Gene_Symbol=TTN Isoform 2 of Titin
Query  Observed  Mr(expt)  Mr(calc)  ppm  Miss  Score  Expect  Rank  Unique  Peptide
28  359.7341  717.4537  717.4537  -0.09  0  7  4.2  5  U  R.LFAIVR.G
209  394.2371  786.4596  786.4599  -0.46  0  8  13  3  U  K.LTIADVVR.A
334  411.2073  820.4000  820.3954  5.61  0  3  15  4  U  K.TDSGLYR.C
357  413.2642  824.5139  824.5135  0.48  1  12  1.1  5  U  K.EPLTLR.K
715  450.7365  899.4584  899.4588  -0.38  0  10  2.9  2  U  K.IVDVSSDR.C
740  451.7681  901.5217  901.5233  -1.72  0  3  24  3  U  R.VTLVDVTR.N
840  459.2484  916.4821  916.4767  5.98  0  2  29  2  U  K.GVEFNVPR.L
844  459.7299  917.4452  917.4454  -0.24  0  4  15  6  U  K.ELEETAAR.H
1029  473.2757  944.5368  944.5331  3.97  1  3  21  3  U  R.EPPSFIKK.I
1058  475.7505  949.4864  949.4869  -0.47  0  4  22  5  U  R.SSVSLSWGK.P
1066  476.2790  950.5433  950.5425  0.94  0  1  23  4  U  R.PLTDLQVR.E
```

**MASCOT** : *Very Large Searches*

© 2007-2010 Matrix Science

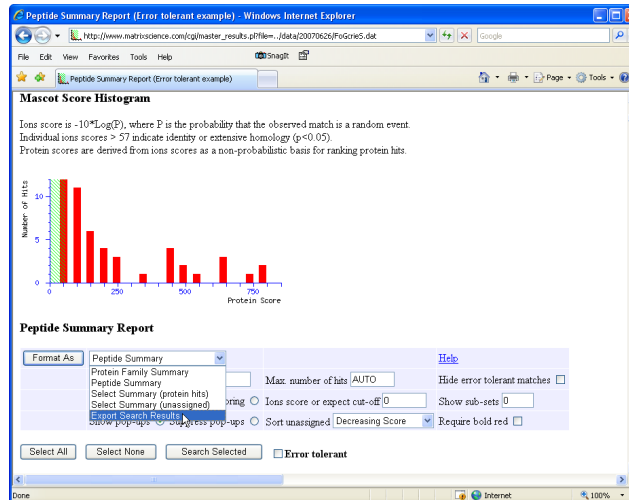


For MudPIT scoring, the score for each peptide is not its absolute score, but the amount that it is above the threshold. Therefore, peptides with a score below the threshold do not contribute to the score. The MudPIT protein score is the sum of the score excess over threshold for each of the matching peptides plus one times the average threshold. For each peptide, the "threshold" is the homology threshold if it exists, otherwise it is the identity threshold.

So, even though a large protein like titin may pick up several random matches, with MudPIT scoring, the protein score is zero, so you don't see it listed in the report unless you specify a huge number of protein hits, as here

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.

## Search result export



**MASCOT** : *Very Large Searches*

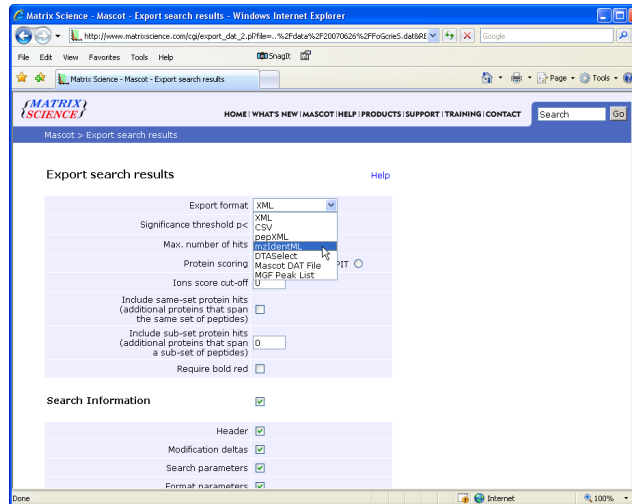
© 2007-2010 Matrix Science



At some stage, it is likely that you will want to export the search results to another application or a relational database. If you want to write your own code, we provide a free library called Mascot Parser that provides a clean, object oriented programming interface to the result file. The supported languages are C++, Java, and Perl.

Mascot also includes a flexible export utility. 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

© 2007-2010 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.

mzIdentML is the new, standard format from PSI for search result interchange.

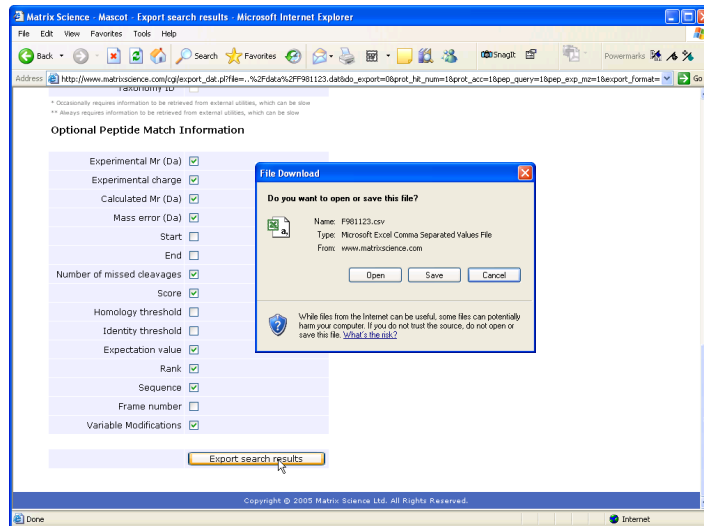
Mascot provides a very full implementation of mzIdentML and this is the one to choose if you are writing new application software that will use Mascot results

DTASelect is the tab separated format used by David Tabb's DTASelect program

The Mascot DAT file is the raw result file. If you need the result file for some reason, and don't have FTP or SCP access to your Mascot server, this is a convenient way to get the file.

MGF peak list is useful when you have the search result but can't find the peak list.

## Search result export



**MASCOT** : Very Large Searches

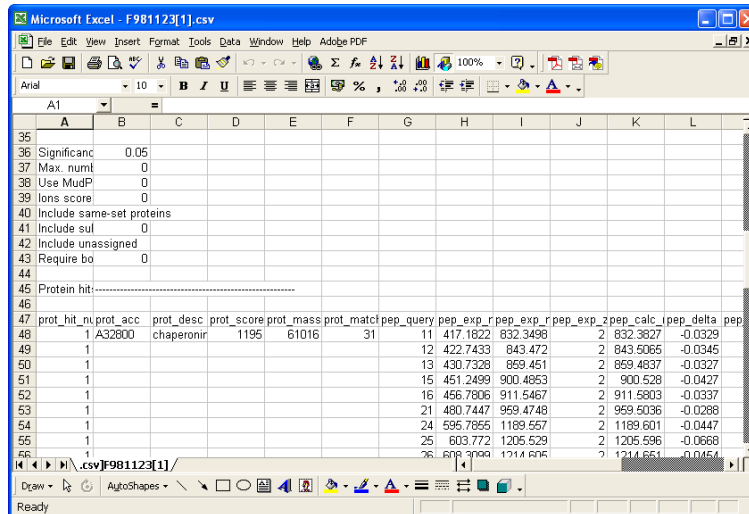
© 2007-2010 Matrix Science



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



The screenshot shows a Microsoft Excel spreadsheet titled "F981123[1].csv". The spreadsheet contains a table of search results. The first few rows are configuration options, and the rest are data rows. The data rows have the following columns: prot\_hit, prot\_acc, prot\_desc, prot\_score, prot\_mass, prot\_match, pep\_query, pep\_exp\_r, pep\_exp\_r, pep\_exp\_r, pep\_exp\_r, pep\_calc, pep\_delta, pep.

	A	B	C	D	E	F	G	H	I	J	K	L	
35													
36	Significanc	0.05											
37	Max. numt	0											
38	Use MudP	0											
39	Ions score	0											
40	Include same-set proteins												
41	Include sul	0											
42	Include unassigned												
43	Require bo	0											
44													
45	Protein hit:-----												
46													
47	prot_hit	prot_acc	prot_desc	prot_score	prot_mass	prot_match	pep_query	pep_exp_r	pep_exp_r	pep_exp_r	pep_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

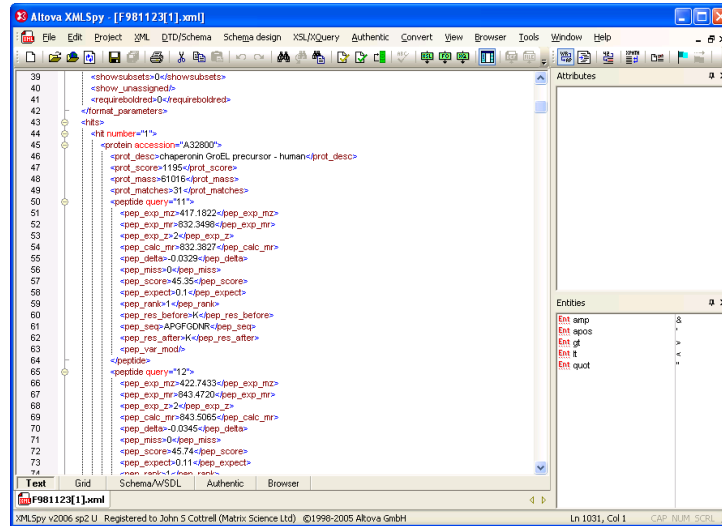
© 2007-2010 Matrix Science



Much easier and safer than “screen scraping”



## Search result export



```
<?xml version="1.0" encoding="UTF-8" ?>
<showsubsets=0</showsubsets>
<show_lines=peptide>
<requireboldred=0</requireboldred>
<format_parameters>
<file>
<id number="1">
<protein accession="A32800">
<prot_desc=chaperonin GroEL precursor - human</prot_desc>
<prot_accession=A32800</prot_accession>
<prot_mass=81016</prot_mass>
<prot_matches=31</prot_matches>
<peptide query="11">
<pep_exp_mz=417.1822</pep_exp_mz>
<pep_exp_mr=832.3498</pep_exp_mr>
<pep_exp_z=2</pep_exp_z>
<pep_calc_mr=832.3827</pep_calc_mr>
<pep_delta=-0.0329</pep_delta>
<pep_mis=0</pep_mis>
<pep_score=45.25</pep_score>
<pep_expect=0.1</pep_expect>
<pep_rank=1</pep_rank>
<pep_res_before=K</pep_res_before>
<pep_seq=AKGFQDNR</pep_seq>
<pep_res_after=K</pep_res_after>
<pep_var_mod>
<peptide>
<peptide query="12">
<pep_exp_mz=422.7433</pep_exp_mz>
<pep_exp_mr=843.4720</pep_exp_mr>
<pep_exp_z=2</pep_exp_z>
<pep_calc_mr=843.5065</pep_calc_mr>
<pep_delta=-0.0345</pep_delta>
<pep_mis=0</pep_mis>
<pep_score=45.74</pep_score>
<pep_expect=0.11</pep_expect>
</peptide>
</peptide>
</file>
</format_parameters>
</show_lines>
</showsubsets>
```

**MASCOT** : Very Large Searches

© 2007-2010 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
480.7447	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 NAGVEGSLVEK
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	84.52	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 ISSIGSVPALEIANHR
960.0327	1918.0609	1918.0636	-0.0127	0	87.34	3.2e-06	1 K ISSIGSVPALEIANHR
1019.5106	2037.0067	2037.0163	-0.0086	0	52.42	0.01	1 R IGEIEQLDVTSEYEK
1057.0537	2112.0529	2112.1322	-0.0393	0	115.78	4.6e-09	1 R ALMLGGVLLADAVAVTMGPK
1065.0399	2128.0663	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	85.46	0.00038	1 R KPLVIAEDVDGEALSTLVNLR
789.1094	2364.3063	2364.3263	-0.0200	0	94.59	4.5e-07	1 R KPLVIAEDVDGEALSTLVNLR
1678.1571	3461.3748	3461.3841	-0.0103	0	47.63	0.03	1 D TALLDAQVAVLITADAAVTEK

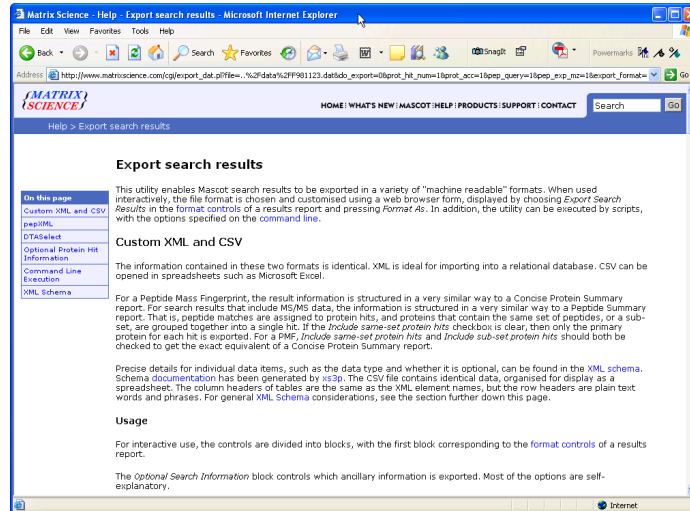
**MASCOT** : Very Large Searches

© 2007-2010 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



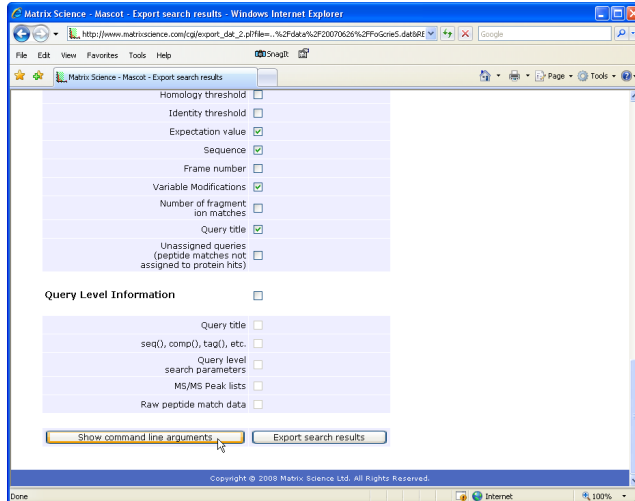
**MASCOT** : Very Large Searches

© 2007-2010 Matrix Science



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

## Search result export



**MASCOT** : *Very Large Searches*

© 2007-2010 Matrix Science

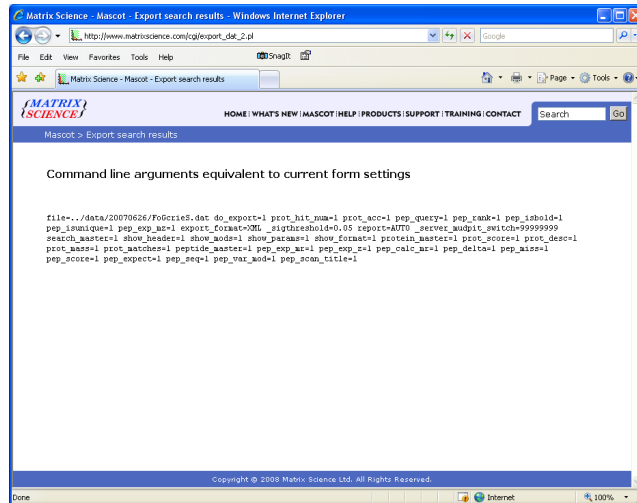


Which describes how the export script can be called from the command line or a shell prompt, as part of an automated pipeline.

I won't go into any detail here, but this means that it is possible to set up a script that will, for example, automatically convert all of your Mascot results to XML files.

Figuring out the command line arguments from the help can be tricky so, in Mascot 2.3, we added a function to display the command line corresponding to the selected options

## Search result export



```
file=../data/20070626/Fo6c1e5.dat do_export=1 prot_hit_max=1 prot_acc=1 pep_query=1 pep_rank=1 pep_sdbold=1 pep_sunique=1 pep_exp_acc=1 export_format=XML sigthreshold=0.05 report=AUTO server_audit_switch=99999999 search_wastest=1 show_headers=1 show_mods=1 show_sequences=1 show_sequences=1 protein_wastest=1 prot_desc=1 prot_name=1 prot_matches=1 peptide_wastest=1 pep_exp_acc=1 pep_exp_z=1 pep_eaic_acc=1 pep_delta=1 pep_minz=1 pep_score=1 pep_expect=1 pep_seq=1 pep_var_mod=1 pep_scan_title=1
```

**MASCOT** : Very Large Searches

© 2007-2010 Matrix Science



By the way, don't delete the original result files after exporting them or you won't be able to view the standard Mascot reports in a browser.