

Very Large Searches

MASCOT

 MATRIX
SCIENCE

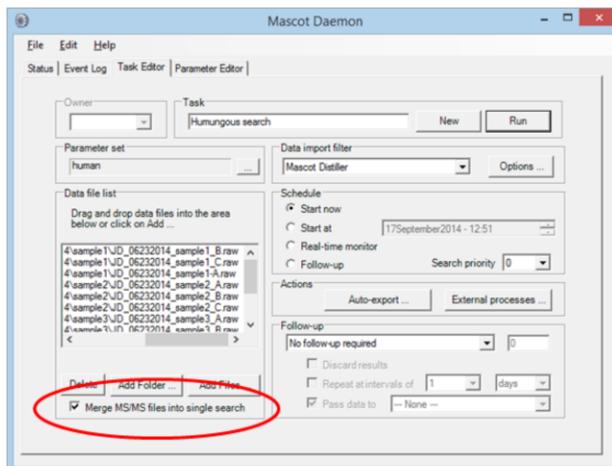
Topics

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

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 fractions
- Use Distiller or a file specific data import filter



MASCOT : *Very Large Searches*

© 2007-2022 Matrix Science



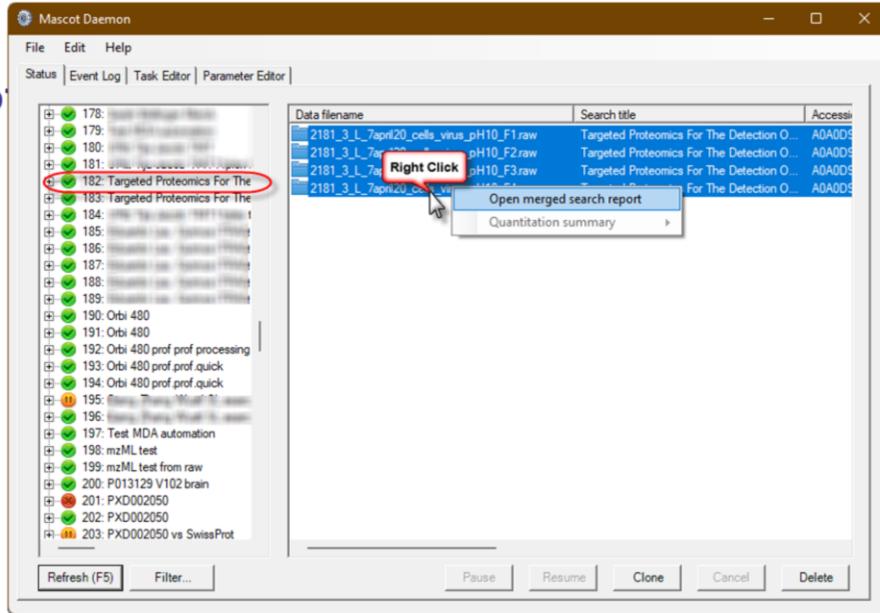
The smartest way to merge files, like fractions from a fractionated run, is using Mascot Daemon. Just tick the box at the bottom left.

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 on a Linux system.

For windows web servers the upload limit is 4 GB. In Mascot Daemon 2.6 we introduced an option to run searches from the command line if Mascot Daemon and Mascot Server are installed on the same computer. This bypasses any web server file limit and search sizes are effectively unlimited.

Comb



MASCOT : Very Large Searches

© 2007-2022 Matrix Science



Now Mascot Daemon 2.7 gives you another way to merge searches.

Instead you can select multiple searches in a Mascot Daemon task by CTRL+click individually searches or shift+click a range then right click and choose combined report.

MASCOT Search Results

Results collection set of 4 files
 ▶ ../data/20200513/F015421.dat
 ▶ ../data/20200513/F015422.dat
 ▶ ../data/20200513/F015423.dat
 ▶ ../data/20200513/F015424.dat

Results collection set limitations
 Re-search: All Non-significant Unassigned [? \[help\]](#) Export As CSV

Search parameters
 Score distribution
 Modification statistics for all protein families
 Legend

Protein Family Summary

Format Significance threshold p< 0.05 Max. number of families AUTO [? \[help\]](#)
 Display non-sig. matches Min. number of sig. unique sequences 1
 Dendrograms cut at 0
 Preferred taxonomy All entries

Sensitivity
 Proteins (4687) [Report Builder](#) [Unassigned \(147681\)](#)

Protein families 1-10 (out of 4358)

10 per page 1 2 3 4 5 6 -- 436 [Next](#) [Expand all](#) [Collapse all](#)

Accession contains Find Clear

▶ 1

1	2::A0A0D9R924	10970	Flacin OS=Chlorocephus sabaeus OX=60711 OX=
2	2::A0A0D9S7P7	856	Microtubule actin crosslinking factor 1 OS=Chloro-
3	2::A0A0D9RLP2	362	Dystonin OS=Chlorocephus sabaeus OX=60711 G-

▶ 2

1	2::A0A0D9S9M0	9915	Uncharacterized protein OS=Chlorocephus sabaeu...
3	2::A0A0D9R4B0	5241	Actin alpha cardiac muscle 1 OS=Chlorocephus sa...
2	2::A0A0D9RYK7	6588	Uncharacterized protein OS=Chlorocephus sabaeu...

MASCOT : Very Large Searches © 2007-2022 Matrix Science **MATRIX SCIENCE**

The combined search will open in a web page and list the results files that have been merged at the top of the report.

This will work with searches that have been processed by any peak picking software including Mascot Distiller.

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
```

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:

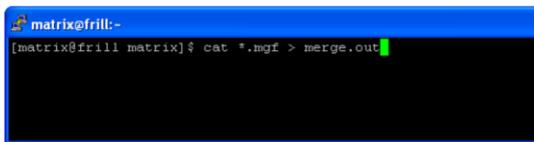
- MGF

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 and later 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

32 bit platforms 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

Mascot Server is now fully 64 bit. **If you have enough RAM you could avoid splitting the search into chunks, however we still do by default because there is no performance penalty in doing so.**

Reporting large searches

Protein Family Summary

- 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 early versions of Mascot, trying to display result reports for very large searches would often lead to problems with timeouts and running out of memory. To address this, the Protein Family Summary loads most of the information 'on demand'. 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.

MASCOT : Very Large Searches © 2007-2022 Matrix Science

If there are 300 or more spectra, the Protein 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 Report Builder, 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 1337 down to 73. 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.

The screenshot displays the Mascot search results for a specific search. At the top, a dendrogram shows the hierarchical clustering of proteins. Below it, a list of proteins is shown with their accession numbers and descriptions. A table of peptide matches is displayed, showing the query, observed and expected matches, scores, and the peptide sequence. The interface includes a threshold setting and a 'Cut' button.

Query Dupes	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	1	2	3	4	5	6	7	Peptide
Q5466	503.3162	1004.6178	1004.5083	0.1095	0	31	0.015	1	0						R.MPTLEDR.T
Q5505	503.8846	1005.7547	1005.6093	0.1454	0	36	0.016	1	0						R.FSVQILR.N
Q5193	516.8977	1031.7808	1031.5369	0.2439	0	32	0.049	1							VQREIDR.V
Q5447	521.2416	1040.7029	1040.5817	0.1212	0	59	1.2e-05	1	0						K.NISQSFTRFSK.A
Q5466	521.3753	1040.7361	1040.5810	0.1551	0	22	0.031	1							R.FFLMILR.N + Oxidation (M)
Q5705	525.4566	1073.3479	1072.7654	0.5824	0	71	1.3e-05	1							R.EALVDRSEFAGR.G
Q5731	526.2961	1050.5776	1050.5323	0.0453	0	35	0.0084	1							R.CLVSELR.R
Q5541	540.3247	1078.6349	1078.5385	0.0964	0	54	0.00019	1							R.TCAQGLR.M
Q5605	541.3848	1080.7551	1080.6059	0.1492	0	53	0.00034	1							K.YFDVFAK.V
Q5793	577.9297	1153.8449	1153.6045	0.2404	0	49	0.00049	1	0						R.GSFMAEK.I
Q5013	581.2500	1160.4854	1160.6167	-0.1313	0	43	0.0013	1	0						R.LCLSEFLR.M

MASCOT : Very Large Searches

© 2007-2022 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. We only reports the significant peptide matches less than the p value for the search that defaults to $p < 0.05$. 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.

The screenshot displays the MASCOT search results for a protein family. At the top, a dendrogram shows the hierarchical clustering of three proteins: GRP78_MOUSE (1), HSP7C_MOUSE (2), and HS71L_MOUSE (3). The HSP7C_MOUSE and HS71L_MOUSE proteins are clustered together, indicating they are more similar to each other than to GRP78_MOUSE. Below the dendrogram, a table lists the proteins with their scores and descriptions:

Protein	Score	Mass	Matches	Sequences	empAI	Description
1 2::GRP78_MOUSE	1308	81404	55 (55)	21 (21)	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
2 2::HSP7C_MOUSE	362	78937	21 (21)	8 (8)	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa6 PE=1 SV=1
3 2::HS71L_MOUSE	188	78552	13 (13)	4 (4)	0.28	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1 PE=2 SV=4

Below the protein list, a table of peptide matches is shown. The table includes columns for Query Dupes, Observed, Mr (expt), Mr (calc), Delta M, Score, Expect, Rank, and Peptide. The peptides are listed with their corresponding scores and expectations. The table is sorted by score, with the highest scores at the top.

MASCOT : Very Large Searches © 2007-2022 Matrix Science

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.

The screenshot displays the MASCOT search results for a query. At the top, a dendrogram shows the relationship between three protein entries: 1 2::GRP78_MOUSE (1308), 2 2::HSP7C_MOUSE (362), and 3 2::HS71L_MOUSE (188). Below the dendrogram is a table of protein matches:

Query	Score	Mass	Matches	Sequences	emPAI	Description
3.1 2::GRP78_MOUSE	1308	81404	55 (55)	21 (21)	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
3.2 2::HSP7C_MOUSE	362	78937	21 (21)	8 (8)	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
3.3 2::HS71L_MOUSE	188	78552	13 (13)	4 (4)	0.28	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa11 PE=2 SV=4

Below the protein matches is a table of peptide matches (22 total, 10 non-duplicate, 12 duplicate):

Query Digest	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	U	2	3	Peptide
2720 7	488.3756	974.7367	974.6004	0.1363	54	0.0024	1	U	1	1	R.LIGDAAK.N
5924 1	546.9979	1091.9813	1091.6430	0.3383	41	0.0061	1	U	1	1	K.ITLTNDK.G
7519	573.9761	1145.9377	1145.6536	0.2841	38	0.0061	1	U	1	1	R.OTLDPVEK.A
9960 2	611.4441	1220.8737	1220.6865	0.1872	60	0.0001	1	U	1	1	K.VGNPILTK.L
10037 1	612.8115	1835.4126	1834.8204	0.5923	35	0.0075	1	U	1	1	K.STAGDTLGGEDFDNR.M
11946 1	641.5476	1281.0806	1280.7220	0.3586	55	0.0015	1	U	1	1	K.EIABAYLQK.T
25277	607.4422	1819.3048	1818.8255	0.4793	55	3.2e-05	1	U	1	1	K.ATAGDTHLGGEDFDNR.L
26376	953.0936	1904.1726	1903.9845	0.1881	84	1.3e-07	1	U	1	1	K.SFYFEEVSSMYLTK.M
26946	650.1325	1947.3756	1947.0920	0.2836	37	0.013	1	U	1	1	R.IINEPTAAAIAYLDK
26947	974.7142	1947.4139	1947.0920	0.3218	43	0.0059	1	U	1	1	R.IINEPTAAAIAYLDK

At the bottom, there are additional protein matches:

4	2::CYB5_MOUSE	1217	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	2::PDIA1_MOUSE	1123	Protein disulfide-isomerase OS=Mus musculus GN=P44b PE=1 SV=2
6	2::CPIA2_MOUSE	1054	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	2::ENPL_MOUSE	1018	Endoplasmic OS=Mus musculus GN=Mp90b1 PE=1 SV=2

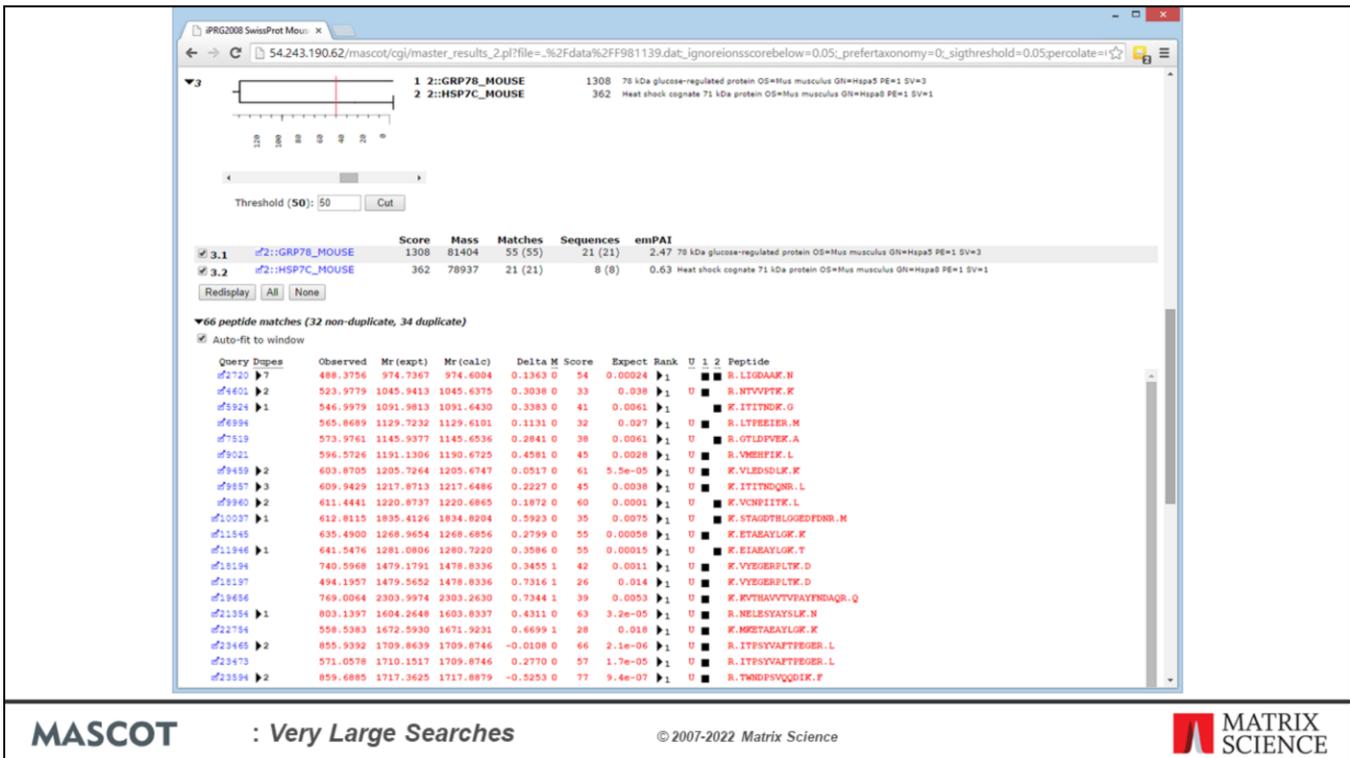
MASCOT : Very Large Searches

© 2007-2022 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.



MASCOT

: Very Large Searches

© 2007-2022 Matrix Science



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.

The screenshot displays the Mascot search results for protein families 41-50. The interface includes a search bar with the query 'MNVLADALK' and a table of protein families. The first family, 2::NBSR3_MOUSE, is expanded to show peptide matches. The table below details these matches:

Query Dups	Observed	Mr (expt)	Mr (calc)	Delta M	Score	Expect	Rank	Peptide
qf1708 5	508.3777	1014.7407	1014.6308	0.1100	45	0.00053	1	K.IVNVILTOR.L
qf11285 5	631.9663	1261.9180	1261.7308	0.1872	77	2.4e-06	1	R.MNVLADALK.S
qf11274	631.8868	1261.7591	1261.7308	0.0284	(66)	1.8e-05	1	R.MNVLADALK.S
qf11278	631.8914	1261.7682	1261.7308	0.0375	(59)	9.7e-05	1	R.MNVLADALK.S
qf11283	631.9416	1261.8686	1261.7308	0.1379	(59)	0.00013	1	R.MNVLADALK.S
qf11287	632.0080	1262.0014	1261.7308	0.2706	(42)	0.0065	1	R.MNVLADALK.S
qf11288	632.0218	1262.0291	1261.7308	0.2983	(63)	6.4e-05	1	R.MNVLADALK.S
qf11604 1	636.4751	1270.9355	1270.6904	0.2452	28	0.03	1	K.WQNNLLFSR.Q
qf11780 1	639.8954	1277.7762	1277.7257	0.0505	50	0.00084	1	R.MNVLADALK.S + Oxidation (M)
qf11790	639.9899	1277.9652	1277.7257	0.2396	(48)	0.00054	1	R.MNVLADALK.S + Oxidation (M)

MASCOT : Very Large Searches

© 2007-2022 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.

PRG2008 SwissProt Mouse x

54.243.190.62/mascot/cgi/master_results_2.pl?file=%2Fdata%2FF981139.dat_ignoreionsscorebelow=0.05_prefertaxonomy=0_sigthreshold=0.05_percolate=1

Proteins (448) Report Builder Unassigned (30397) [s.permalink](#)

Protein hits (476 proteins)

Columns: Standard (12 out of 16)

Filters: (none)

Export as CSV

Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	cRAP	#1::sp TRV1_BOVIN	1606	28266	48	48	7	7	2.86	sp TRV1_BOVIN
2	1	SwissProt	#2::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1
2	2	SwissProt	#2::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2
2	3	SwissProt	#2::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1
2	4	SwissProt	#2::CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=
2	5	SwissProt	#2::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2
2	6	SwissProt	#2::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2
2	7	SwissProt	#2::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2
3	1	SwissProt	#2::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa7
3	2	SwissProt	#2::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hsp
4	1	SwissProt	#2::CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	#2::PDIA1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1
6	1	SwissProt	#2::CP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV
7	1	SwissProt	#2::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum protein OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	#2::RDH7_MOUSE	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV
8	2	SwissProt	#2::H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus muscul
9	1	SwissProt	#2::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=
10	1	SwissProt	#2::RL7A_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2
11	1	SwissProt	#2::RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 P
12	1	SwissProt	#2::CP2AC_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1
12	2	SwissProt	#2::CP2A3_MOUSE	59	61896	5	5	2	2	0.17	Cytochrome P450 2A3 OS=Mus musculus GN=Cyp2a3 PE=2 SV
13	1	SwissProt	#2::ACSL1_MOUSE	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid-CoA ligase 1 OS=Mus musculus GN=AcS
13	2	SwissProt	#2::ACSL5_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid-CoA ligase 5 OS=Mus musculus GN=AcS
14	1	SwissProt	#2::RL13_MOUSE	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2
15	1	SwissProt	#2::PDIA3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE
16	1	SwissProt	#2::CP3A8_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1
17	1	SwissProt	#2::UDB17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2
17	2	SwissProt	#2::UD11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a
17	3	SwissProt	#2::UD16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a
18	1	SwissProt	#2::EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2

MASCOT : Very Large Searches © 2007-2022 Matrix Science

MATRIX SCIENCE

The Report Builder tab is useful when you need a table of proteins suitable for publication. Lets assume we want to drop the 'one hit wonders' and only report proteins that have significant matches to at least 2 different peptide sequences

Protein hits (476 proteins)
 Columns: Standard (12 out of 16)
 Filters: (none)

Family	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1606	28266	48	48	7	7	2.86	spiTRY1_BOVIN
2	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1
2	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2
2	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1
2	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=
2	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2
2	251	60856	13	13	4	4	0.37	Cytochrome P450 2C36 OS=Mus musculus GN=Cyp2c36 PE=2
2	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2
2	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa
2	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hsp
4	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1
6	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=
7	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum chaperone protein OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV
8	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus muscul
9	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=
10	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2
11	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 P
12	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1
12	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV
13	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acs
13	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acs
14	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2
15	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE
16	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1
17	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2

MASCOT : Very Large Searches

© 2007-2022 Matrix Science



We open up the filters section and add a suitable filter.

Protein hits (229 proteins)

Filters: "Num. of significant sequences" >= 2

Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPA1	Description
1	1	cRAP	#1::sp TRY1_BOVIN	1606	28266	48	48	7	7	2.86	sp TRY1_BOVIN
2	1	SwissProt	#2::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV
2	2	SwissProt	#2::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV
2	3	SwissProt	#2::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV
2	4	SwissProt	#2::CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1
2	5	SwissProt	#2::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV
2	6	SwissProt	#2::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV
2	7	SwissProt	#2::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV
3	1	SwissProt	#2::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 P
3	2	SwissProt	#2::HSP97_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8
4	1	SwissProt	#2::CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	#2::PDIA1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV
6	1	SwissProt	#2::CP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	1	SwissProt	#2::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum protein OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	#2::RDH7_MOUSE	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
8	2	SwissProt	#2::H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus
9	1	SwissProt	#2::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mt
10	1	SwissProt	#2::RL7A_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV
11	1	SwissProt	#2::RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=
12	1	SwissProt	#2::CP2A2_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1 SV
12	2	SwissProt	#2::CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV=1
13	1	SwissProt	#2::ACSL1_MOUSE	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid-CoA ligase 1 OS=Mus musculus GN=Acsl1
13	2	SwissProt	#2::ACSL3_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid-CoA ligase 3 OS=Mus musculus GN=Acsl3
14	1	SwissProt	#2::RL13_MOUSE	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2 SV
15	1	SwissProt	#2::PDIA3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1
16	1	SwissProt	#2::CP3A8_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1 SV
17	1	SwissProt	#2::UDB17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugt2b1
17	2	SwissProt	#2::UDH1_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 f
17	3	SwissProt	#2::UDH16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a6 f
18	1	SwissProt	#2::EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2

MASCOT : Very Large Searches

© 2007-2022 Matrix Science



Only proteins with significant matches to at least 2 sequences remain. The filtering is very flexible, with lots of useful terms.

PRG2008 SwissProt Mouse X

54.243.190.62/mascot/cgi/master_results_2.pl?file=%2Fdata%2FF981139.dat_ignoreionsscorebelow=0.05_prefertaxonomy=0_sigthreshold=0.05_percolate=1

Proteins (448) Report Builder Unassigned (30397) Permalink

Protein hits (228 proteins)

Columns: Standard (12 out of 16)

Filters: (NOT(Database is cRAP) AND "Num. of significant sequences" >= 2)

Export as CSV

Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
2	1	SwissProt	f2::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=
2	2	SwissProt	f2::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=
2	3	SwissProt	f2::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=
2	4	SwissProt	f2::CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1
2	5	SwissProt	f2::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=
2	6	SwissProt	f2::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=
2	7	SwissProt	f2::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=
2	1	SwissProt	f2::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspas78
2	2	SwissProt	f2::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa81
4	1	SwissProt	f2::CYB5_MOUSE	1217	16817	42	42	5	5	3.08	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
5	1	SwissProt	f2::PDIA1_MOUSE	1123	64694	53	53	16	16	2.54	Protein disulfide-isomerase OS=Mus musculus GN=Pdia1 PE=1 SV=
6	1	SwissProt	f2::CP1A2_MOUSE	1054	63034	38	38	10	10	1.31	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
7	1	SwissProt	f2::ENPL_MOUSE	1018	103744	63	63	19	19	1.53	Endoplasmic reticulum protein OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
8	1	SwissProt	f2::RDH7_MOUSE	1005	38455	45	45	12	12	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
8	2	SwissProt	f2::H17B6_MOUSE	597	38949	23	23	7	7	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus GN=
9	1	SwissProt	f2::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgst1
10	1	SwissProt	f2::RL7A_MOUSE	770	35860	28	28	8	8	1.91	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=
11	1	SwissProt	f2::RLA0_MOUSE	763	37215	24	24	7	7	1.47	60S acidic ribosomal protein P0 OS=Mus musculus GN=Rplp0 PE=1
12	1	SwissProt	f2::CP2A2_MOUSE	763	61325	35	35	14	14	2.25	Cytochrome P450 2A12 OS=Mus musculus GN=Cyp2a12 PE=1 SV=
12	2	SwissProt	f2::CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 PE=2 SV=1
13	1	SwissProt	f2::ACSL1_MOUSE	749	86078	38	38	18	18	1.90	Long-chain-fatty-acid--CoA ligase 1 OS=Mus musculus GN=Acsl1 P
13	2	SwissProt	f2::ACSL5_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acsl5 P
14	1	SwissProt	f2::RL13_MOUSE	748	28083	31	31	7	7	2.90	60S ribosomal protein L13 OS=Mus musculus GN=Rpl13 PE=2 SV=
15	1	SwissProt	f2::PDIA3_MOUSE	692	64504	40	40	15	15	2.06	Protein disulfide-isomerase A3 OS=Mus musculus GN=Pdia3 PE=1
16	1	SwissProt	f2::CP3A8_MOUSE	686	65154	32	32	10	10	1.25	Cytochrome P450 3A11 OS=Mus musculus GN=Cyp3a11 PE=1 SV=
17	1	SwissProt	f2::UGT17_MOUSE	677	67040	34	34	9	9	0.91	UDP-glucuronosyltransferase 2B17 OS=Mus musculus GN=Ugtb17
17	2	SwissProt	f2::UGT11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 PE
17	3	SwissProt	f2::UGT16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a6 PE
18	1	SwissProt	f2::EST3A_MOUSE	668	67490	28	28	5	5	0.43	Carboxylesterase 3A OS=Mus musculus GN=Ces3a PE=1 SV=2
19	1	SwissProt	f2::RL4_MOUSE	650	55568	34	34	11	11	1.59	60S ribosomal protein L4 OS=Mus musculus GN=Rpl4 PE=1 SV=3

MASCOT : Very Large Searches

© 2007-2022 Matrix Science



Another thing that you could easily do would be to exclude proteins from the contaminants database

Protein hits (228 proteins)

▼ Columns: Standard (12 out of 16)

Arrangement: <custom> Load Make default

Enabled

- Family
- Member
- Database
- Accession
- Score
- Mass
- Num. of matches
- Num. of significant matches
- Num. of sequences
- Num. of significant sequences
- emPAI
- Description

Available

- Protein hits
- Num. of unique sequences
- Num. of significant unique sequences
- Sequence coverage
- pl

Filters: (NOT(Database is cRAP) AND "Num. of significant sequences" >= 2)

Export as CSV

Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
2	1	SwissProt	g2::CP2CT_MOUSE	1337	61419	76	76	13	13	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=
2	2	SwissProt	g2::CP254_MOUSE	552	60887	27	27	8	8	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=
2	3	SwissProt	g2::CY250_MOUSE	489	61128	27	27	10	10	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=
2	4	SwissProt	g2::CP2F2_MOUSE	484	59267	32	32	12	12	2.11	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1
2	5	SwissProt	g2::CP237_MOUSE	339	60590	22	22	8	8	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=
2	6	SwissProt	g2::CP239_MOUSE	251	60856	13	13	4	4	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=
2	7	SwissProt	g2::CP238_MOUSE	150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=
3	1	SwissProt	g2::GRP78_MOUSE	1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE
3	2	SwissProt	g2::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock coonate 71 kDa protein OS=Mus musculus GN=Hspa8 f

MASCOT : Very Large Searches

© 2007-2022 Matrix Science



The columns section of Report Manager allows you to choose which columns to include and, if required, change their order

Family	Member	Database	Accession	Score	Mass	Num. of matches	Num. of significant matches	Num. of sequences	Num. of significant sequences	emPAI	Description	
31	1	IPRG_2012	P00925	2140	46842	148	100	53	43	44.71	Enolase 2 OS=Saccharomyces cere	
32	1	IPRG_2012	P00924	1059	46844	71	46	35	27	7.47	Enolase 1 OS=Saccharomyces cere	
33	2	IPRG_2012	P00549	1933	54909	133	87	56	43	18.28	Pyruvate kinase 1 OS=Saccharomy	
34	3	IPRG_2012	P40150	1613	62668	105	66	66	45	11.76	Heat shock protein SSB2 OS=Sacch	
35	3	IPRG_2012	P11484	1590	66732	103	65	64	44	11.12	Heat shock protein SSB1 OS=Sacch	
36	4	IPRG_2012	P10592	1591	69599	107	57	52	32	5.01	Heat shock protein SSA2 OS=Sacch	
37	4	IPRG_2012	P10591	1161	69786	85	44	48	26	3.02	Heat shock protein SSA1 OS=Sacch	
38	4	IPRG_2012	P16474	233	74479	23	8	17	6	0.32	78 kDa glucose-regulated protein hor	
39	5	IPRG_2012	P00330	1453	37282	73	51	32	25	13.48	Alcohol dehydrogenase 1 OS=Sacch	
40	5	IPRG_2012	P07246	101	40743	14	5	7	3	0.29	Alcohol dehydrogenase 3, mitochond	
41	6	IPRG_2012	P00560	1382	44768	102	58	54	33	12.75	Phosphoglycerate kinase OS=Sacch	
42	7	IPRG_2012	P00359	1361	35838	76	54	31	25	12.29	Glyceraldehyde-3-phosphate dehydro	
43	7	IPRG_2012	P00358	1242	35938	69	48	29	24	9.69	Glyceraldehyde-3-phosphate dehydro	
44	7	IPRG_2012	P00360	535	35842	30	20	14	12	2.47	Glyceraldehyde-3-phosphate dehydro	
45	7	IPRG_2012	P04406	41	36201	4	2	4	2	0.21	Glyceraldehyde-3-phosphate dehydro	
46	8	IPRG_2012	P06169	1289	61685	44	41	28	26	4.7	Pyruvate decarboxylase isozyme 1 O	
47	9	IPRG_2012	P00950	1031	27592	67	44	32	25	34.97	Phosphoglycerate mutase 1 OS=Sa	
48	10	IPRG_2012	P07281	1015	15881	51	38	16	13	22.71	40S ribosomal protein S19-B OS=Sa	
49	10	IPRG_2012	P07280	1014	15807	51	38	16	13	22.71	40S ribosomal protein S19-A OS=Sa	
50	11	1	contaminants	P00761	922	25078	37	27	7	6	2.89	SWISS-PROT P00761 TRYPL_PIG Tr
51	12	1	IPRG_2012	P32324	784	93686	49	33	33	23	1.44	Elongation factor 2 OS=Saccharomy
52	13	1	IPRG_2012	P16521	771	116727	62	33	47	30	1.52	Elongation factor 3A OS=Saccharom
53	14	1	IPRG_2012	P06319	765	10739	38	29	10	9	95.65	60S acidic ribosomal protein P2- α
54	15	1	IPRG_2012	Q03048	721	15948	28	23	17	14	17.82	Cofilin OS=Saccharomyces cerevisi
55	16	1	IPRG_2012	P0C0V8	719	9797	42	29	15	12	207.43	40S ribosomal protein S21-A OS=Sa
56	16	2	IPRG_2012	Q3E754	694	9811	41	28	15	12	148.28	40S ribosomal protein S21-B OS=Sa

MASCOT : Very Large Searches

© 2007-2022 Matrix Science



Once the list is filtered and the columns arranged as required, there is a button to export the table as CSV, which can be pasted into Excel and formatted to create a suitable figure for dropping into a publication

Large search results in 2.2 and earlier

The screenshot shows the Mascot search interface with several callouts in yellow boxes:

- Never Peptide**: Points to the 'Select Summary Report' dropdown menu.
- Important**: Points to the 'Help' link in the top right.
- Help**: Points to the 'Help' link in the top right, which is circled in red.
- Reduces memory**: Points to the 'Preferred taxonomy' dropdown menu.
- Simplifies**: Points to the 'Require bold red' checkbox.

Other visible elements include: 'Format As' dropdown, 'Significance threshold p< 0.05', 'Standard scoring' and 'MudPIT scoring' radio buttons, 'Show pop-ups' and 'Suppress pop-ups' radio buttons, 'Max. number of hits' dropdown (set to AUTO), 'Display non-significant matches' checkbox, 'Show sub-sets' dropdown (set to 0), and 'Show Percolator scores' checkbox.

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

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 leaving "Display non-significant matches" unchecked will minimize the number of hits: `&_requireboldred=1`

Mascot database search | Results X +

edectus/mascot/help/results_help.html 110%

master_results.pl and master_results_2.pl

URL	mascot.dat	master_results.pl	master_results_2.pl	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	For a Peptide Summary, set the value to 1 to report all hits that match a subset of peptides. Default is 0 for no sub-set hits. Intermediate values set a threshold on the difference in protein score between the primary hit and the sub-set hit expressed as a fraction.
_requireboldred	RequireBoldRed	✓		1	Set value to 1 to report Peptide Summary hits only if they contain at least one "bold red" peptide, (default 0).
					Set value to 1 to report all matches

MASCOT : Very Large Searches © 2007-2022 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) ▾	Help	Help
Significance threshold p<	0.05	Max. number of hits	AUTO
Standard scoring	<input type="radio"/> MudPIT scoring <input checked="" type="radio"/>	Display non-significant matches	<input type="checkbox"/>
Show pop-ups	<input checked="" type="radio"/> Suppress pop-ups <input type="radio"/>	Show sub-sets	0
Preferred taxonomy	All entries ▾	Require bold red	<input type="checkbox"/>

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 Hatches: 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.LTIADVR.A
334	411.2073	820.4000	820.3954	5.61	0	3	15	4	U	K.TDSGLVR.C
357	413.2642	824.5139	824.5135	0.48	1	12	1.1	5	U	K.RFRTLK.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.GVEFNVR.L
844	459.7299	917.4452	917.4454	-0.24	0	4	15	6	U	K.ELEETAAR.N
1029	473.2757	944.5368	944.5331	3.97	1	3	21	3	U	R.EPPSFYIKK.I
1058	475.7505	949.4864	949.4869	-0.47	0	4	22	5	U	R.SSVLSLQGR.P
1066	476.2790	950.5433	950.5425	0.94	0	1	23	4	U	R.PLTDLQVR.E

MASCOT

: Very Large Searches

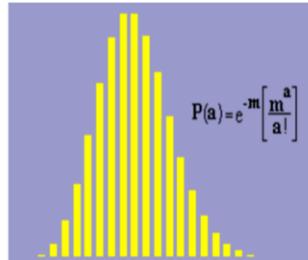
© 2007-2022 Matrix Science



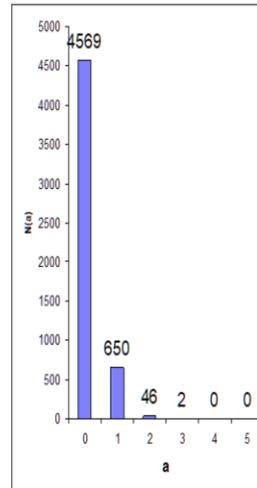
With standard peptide summary scoring, the protein score is essentially the sum of the ions scores of all the peptides assigned to the protein. Where there are duplicate matches to the same peptide, the highest scoring match is used. A correction is applied based on the number of candidate peptides that 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 many 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

© 2007-2022 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::IP100023283  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.LTIADVR.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._BFLTLR.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 K.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

© 2007-2022 Matrix Science

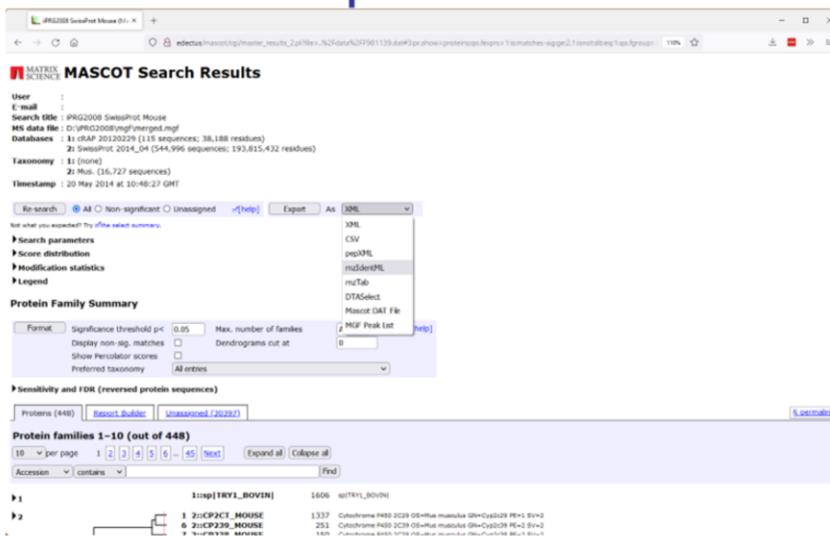


To avoid this problem, we use MudPIT protein scoring, in which the score for each peptide match is not its absolute score, but the amount that it is above the threshold. Therefore, matches with a score below the threshold do not contribute to the score. The MudPIT protein score is the sum of the score 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 was done here to capture this screen shot.

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



The screenshot shows the MASCOT Search Results page. At the top, the title is "MASCOT Search Results". Below this, there are several sections: "User", "Search ID", "MS data file", "Databases", "Taxonomy", and "Timestamp". The "Export" button is highlighted, and a dropdown menu is open, showing the following options: XML, CSV, pepXML, mzIdentML, mzTab, DTASelect, Mascot DAT file, and MGF Peak List. The "MGF Peak List" option is currently selected. Below the export menu, there are various search parameters and a "Protein Family Summary" section. The "Protein Family Summary" section includes a "Format" dropdown, a "Significance threshold p<" field, a "Max. number of families" field, and a "Dendrograms cut at" field. There are also checkboxes for "Display non sig. matches", "Show Peptide scores", and "Preferred taxonomy". The "Protein Family Summary" section also includes a "Sensitivity and FDR (reversed protein sequences)" section. At the bottom of the screenshot, there is a table of protein families with columns for "Accession", "contains", and "Find".

MASCOT : Very Large Searches

© 2007-2022 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.

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 and mzTab are the standard formats 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

The screenshot shows a web browser window displaying the Matrix Science Mascot search result export page. The page title is "Export search results". The interface includes a search bar at the top right and a navigation menu with links for "Home", "Access Mascot Server", "Database search help", and "Contact". Below the navigation menu, there is a breadcrumb trail: "Mascot database search > Access Mascot Server > Export search results". The main content area is titled "Export search results" and contains a form with the following options:

- Export format: **msl6m8L** (dropdown menu)
- Significance threshold p: **0.05** (input field) at **Identity** (radio button) **Homology** (radio button)
- Target FDR (overrides significance threshold if set): **<not set>** (dropdown menu)
- FDR type: **Detect PDMs** (dropdown menu)
- Display non-significant matches:
- Max. number of hits: **AUTO** (input field)
- Protein scoring: Standard **MuSPIT**
- 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): **1** (input field)
- Group protein families:
- Require bold red:
- Show Percolator scores:
- Preferred "taxonomy": **All entries** (dropdown menu)

Below the form, there is a note: "* Occasionally requires information to be retrieved from external utilities, which can be slow". At the bottom of the form area, it says "Optional Protein Hit Information".

MASCOT : *Very Large Searches*

© 2007-2022 Matrix Science



If you arrive here from one of the older reports, to begin with, you may need to select the required output format. Different formats have different options further down the page

Search result export



To export to Excel, simply select CSV as the format, and click on the Export Search Results button at the bottom of the page. In recent versions of Mascot, the report is prepared and then a download button is displayed. In older versions, the download would start immediately. Once the download is finished, you can open it into Excel:

Search result export

The screenshot shows an Excel spreadsheet with the following sections:

- Search Parameters (Rows 33-73):** A list of parameters for the Mascot search, such as 'Significance threshold' (0.05), 'Max. number of hits' (0), and 'Use MascOT protein score' (1).
- Protein Hits (Row 74):** A header row for the search results.
- Search Results (Rows 75-85):** A table of search results with columns for protein name, accession number, and scores. The first few rows show results for 'TRV' proteins with scores around 14000.

MASCOT : Very Large Searches

© 2007-2022 Matrix Science



Much easier and safer than “screen scraping”

Search result export

The screenshot shows a Microsoft Access window with a table named 'peptide : Table'. The table contains the following data:

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 APFGDNR
451.2499	900.4853	900.5280	-0.0427	0	51.95	0.025	1 K LSDGVAVLK
456.7906	911.5457	911.5803	-0.0337	0	59	0.0041	1 K VGLQVAVLK
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	0.0068	1 K EIGNISDAMK
603.7720	1205.5294	1205.5961	-0.0668	0	60.13	0.027	1 K EIGNISDAMK
608.3099	1214.6052	1214.6506	-0.0454	0	73.21	0.00015	1 K NAGVEGSLVEK
617.2857	1232.5569	1232.5884	-0.0315	0	80.63	2.7e-05	1 K VGGTSDVEVNEK
672.8375	1343.6605	1343.7095	-0.0480	0	64.38	0.001	1 R TVIEQSWGSPK
714.8884	1427.7623	1427.8057	-0.0434	0	64.52	0.00086	1 R GVMLAVDAVAELK
714.8898	1427.7730	1427.8057	-0.0327	0	72.61	0.00013	1 R GVMLAVDAVAELK
722.8849	1443.7552	1443.8006	-0.0454	0	72.71	0.00014	1 R GVMLAVDAVAELK
722.8934	1443.7722	1443.8006	-0.0284	0	70.08	0.00025	1 R GVMLAVDAVAELK
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	1917.9625	1918.0636	-0.1010	0	101.5	1.3e-07	1 K ISSIQSNVPALEIANHR
960.0327	1918.0599	1918.0636	-0.0127	0	87.34	3.2e-06	1 K ISSIQSNVPALEIANHR
1019.5106	2037.0067	2037.0153	-0.0086	0	52.42	0.01	1 R DQIEQLDVTTSYEK
1057.0537	2112.0529	2112.1322	-0.0393	0	115.78	4.6e-09	1 R ALMLGGVOLLADAVAVTMGPK
1065.0399	2128.0653	2128.1271	-0.0618	0	68.73	0.00022	1 R ALMLGGVOLLADAVAVTMGPK
1073.0477	2144.0809	2144.1220	-0.0411	0	69.64	0.00018	1 R ALMLGGVOLLADAVAVTMGPK
789.1062	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	2364.3263	2364.3263	-0.0200	0	94.59	4.5e-07	1 R KPLVIAEDVDGEALSTLVNLR
1076.1377	2481.3748	2481.3641	0.0103	0	47.65	0.03	1 D TALLDAQVADSI TRADAMTERE

MASCOT : Very Large Searches

© 2007-2022 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

MATRIX SCIENCE

Home Access Mascot Server Database search help Contact

Mascot database search > help > Export search results

Export search results

This utility enables Mascot search results to be exported in a variety of "machine readable" formats. When used interactively, the file format is chosen and customised using a web browser form, displayed by choosing Export Search Results in the format controls of a results report and pressing Format As. In addition, the utility can be executed by scripts, with the options specified on the command line.

Custom XML and CSV

The information contained in these two formats is identical. XML is ideal for importing into a relational database. CSV can be opened in spreadsheets such as Microsoft Excel.

For a Peptide Mass Fingerprint, the result information is structured in a very similar way to a Concise Protein Summary report. For search results that include MS/MS data, you can choose whether to structure the protein list and associated peptide matches in a similar way to a Peptide Summary report or a Protein Family report. To create an export that contains information equivalent to a particular Mascot HTML report, the settings of the format controls must match. (b/c)

Type of search	HTML Report	Threshold type	Protein Scoring	Same-sets	Sub-sets	Group proteins
PHF	Concise Protein Summary	N/A	N/A	checked	1	N/A
MS/MS	Peptide Summary	Identity	As format controls	checked	As format controls	not checked
MS/MS	Protein Family Report	Homology	MS/MS	checked	1	checked

Sections

- Custom XML and CSV
- pepXML
- msIdentML
- muTab
- DTASelect
- Mascot D&I File
- MSF Peak List
- Optional Protein Hit Information
- Command Line Execution
- XML Schema

FSB1139.csv

Show all downloads...

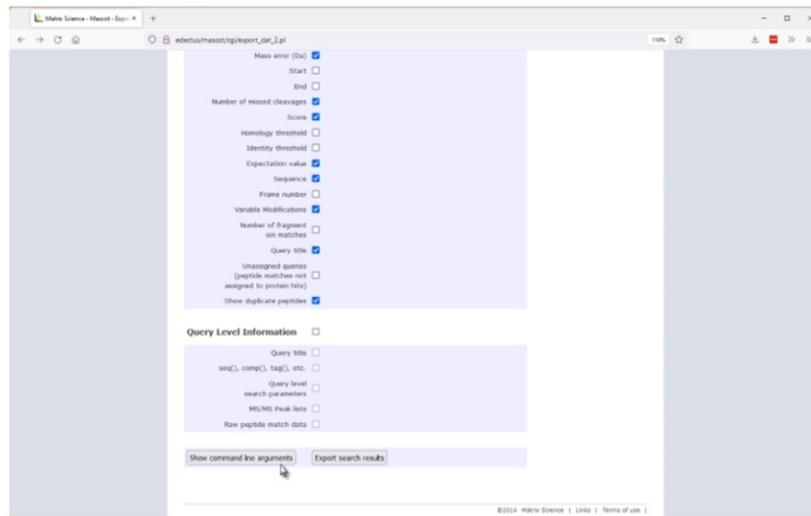
MASCOT : Very Large Searches

© 2007-2022 Matrix Science



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

Search result export



MASCOT : *Very Large Searches*

© 2007-2022 Matrix Science

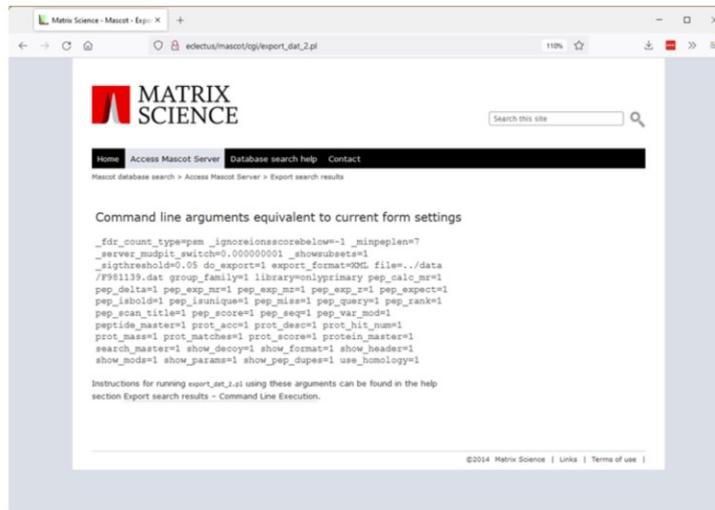


Which describes how the export script can be called from the command line or a shell prompt, as part of an automated pipeline.

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



MASCOT : *Very Large Searches*

© 2007-2022 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.