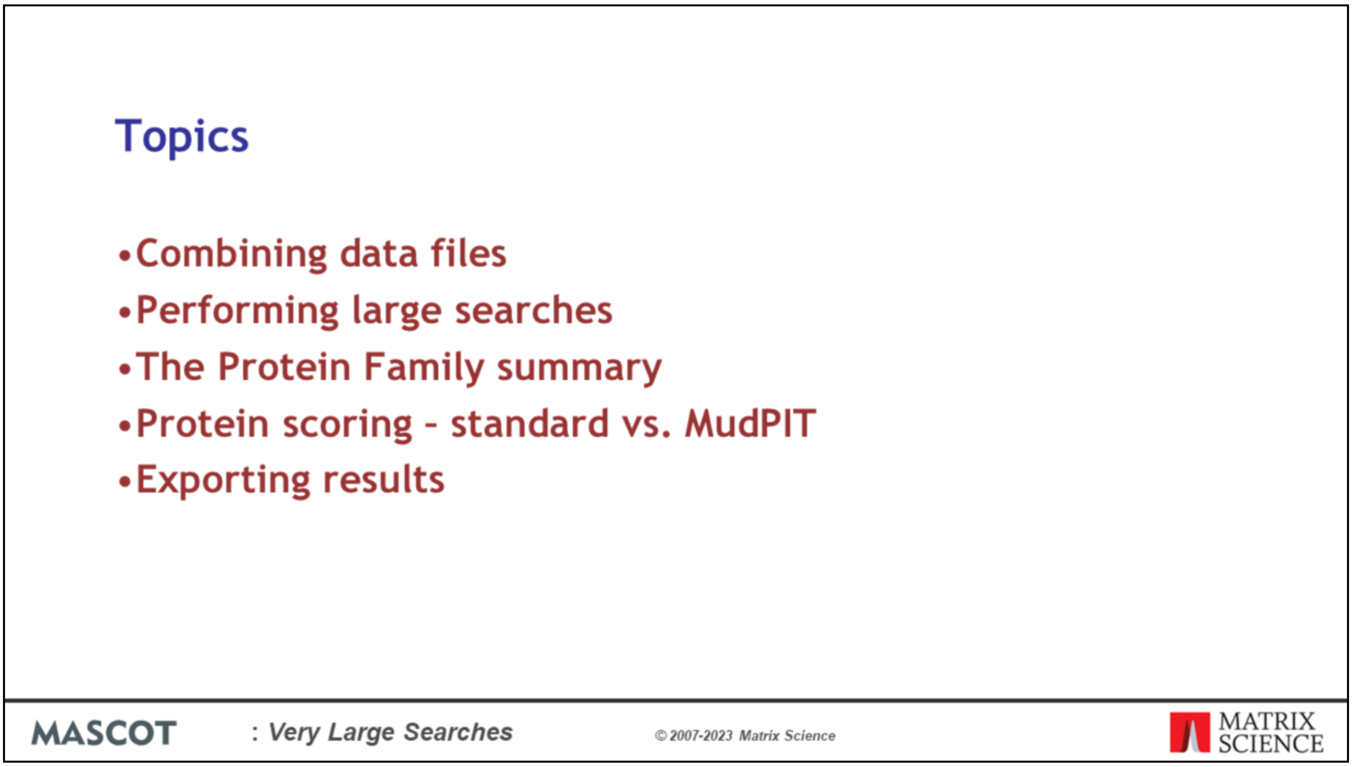
Very Large Searches



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.

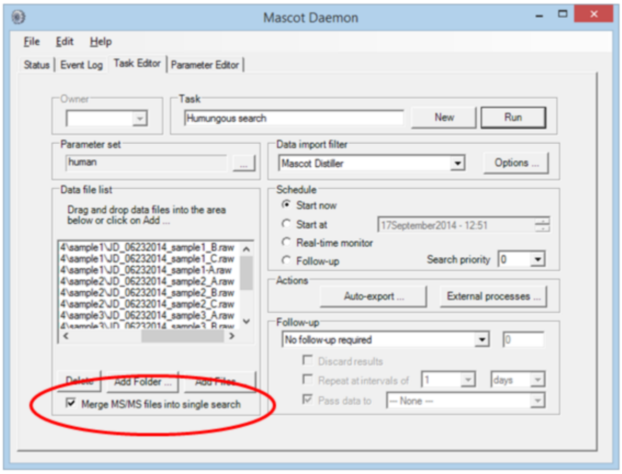
First though, what is a large search? And are there any search size limits to Mascot Server?

Data acquisition rates have sped up with each new generation of instrumentation and what was a large search 20 years ago is now a small search. For this talk we will consider an search with over 500,000 queries a large search.

There is no software limit to the maximum size of the search that Mascot Server supports, but larger searches do use more hardware resources and that is the ultimate limit. Given sufficient resources, we have had no problems with peak lists of 10GB or data sets with 50 x 1 hour Orbitrap runs.

Data files

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



Owner:

Task: Humungous search

Parameter set: human

Data import filter: Mascot Distiller

Data file list

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

```
4\sample1\JD_06232014_sample1_B.raw
4\sample1\JD_06232014_sample1_C.raw
4\sample1\JD_06232014_sample1-A.raw
4\sample2\JD_06232014_sample2_A.raw
4\sample2\JD_06232014_sample2_B.raw
4\sample2\JD_06232014_sample2_C.raw
4\sample3\JD_06232014_sample3_A.raw
4\sample3\JD_06232014_sample3_B.raw
```

☑ Merge MS/MS files into single search

Schedule

● Start now

○ Start at 17September2014 - 12:51

○ Real-time monitor

○ Follow-up

Search priority 0

Actions

Follow-up: No follow-up required 0

□ Discard results

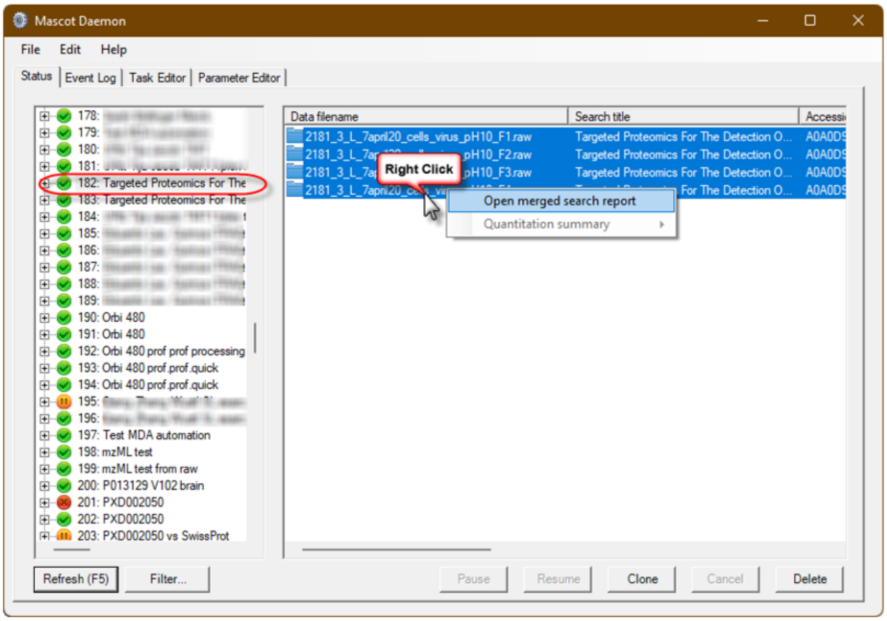
□ Repeat at intervals of 1 days

☑ Pass data to -- None --

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.

For Windows web servers, the upload limit is 4 GB. Mascot Daemon can also 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.

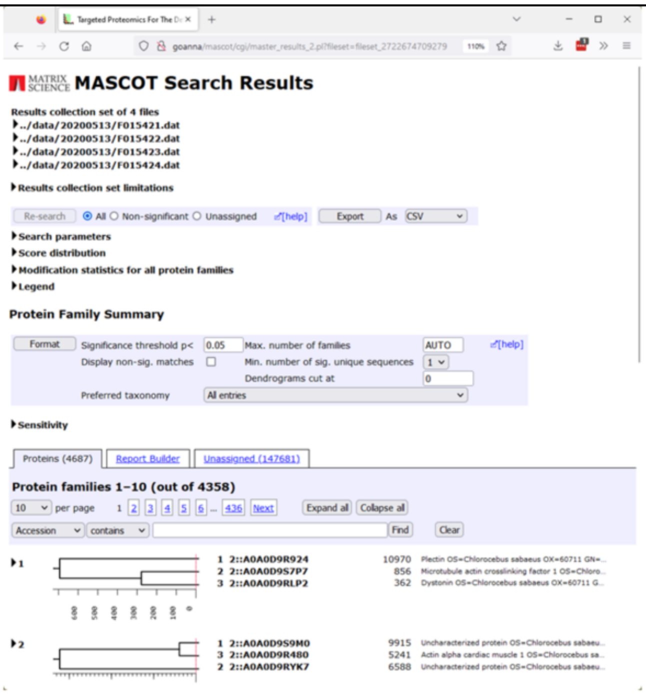


178: [redacted]
179: [redacted]
180: [redacted]
181: [redacted]
182: Targeted Proteomics For The
183: Targeted Proteomics For The
184: [redacted]
185: [redacted]
186: [redacted]
187: [redacted]
188: [redacted]
189: [redacted]
190: Orbi 480
191: Orbi 480
192: Orbi 480 prof prof processing
193: Orbi 480 prof prof quick
194: Orbi 480 prof prof quick
195: [redacted]
196: [redacted]
197: Test MDA automation
198: mzML test
199: mzML test from raw
200: P013129 V102 brain
201: PXD002050
202: PXD002050
203: PXD002050 vs SwissProt

Data filename	Search title	Accessi
2181_3_L_7april20_cells_virus_pH10_F1.raw	Targeted Proteomics For The Detection O...	A0A0D9
2181_3_L_7a... pH10_F2.raw	Targeted Proteomics For The Detection O...	A0A0D9
2181_3_L_7a... pH10_F3.raw	Targeted Proteomics For The Detection O...	A0A0D9
2181_3_L_7april20_c...	... For The Detection O...	A0A0D9

Mascot Daemon 2.7 and later give you another way to merge searches.

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.



Targeted Proteomics For The D

goanna/mascot/cgi/master_results_2.pl?fileset=fileset_2722674709279 110%

MATRIX SCIENCE 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)

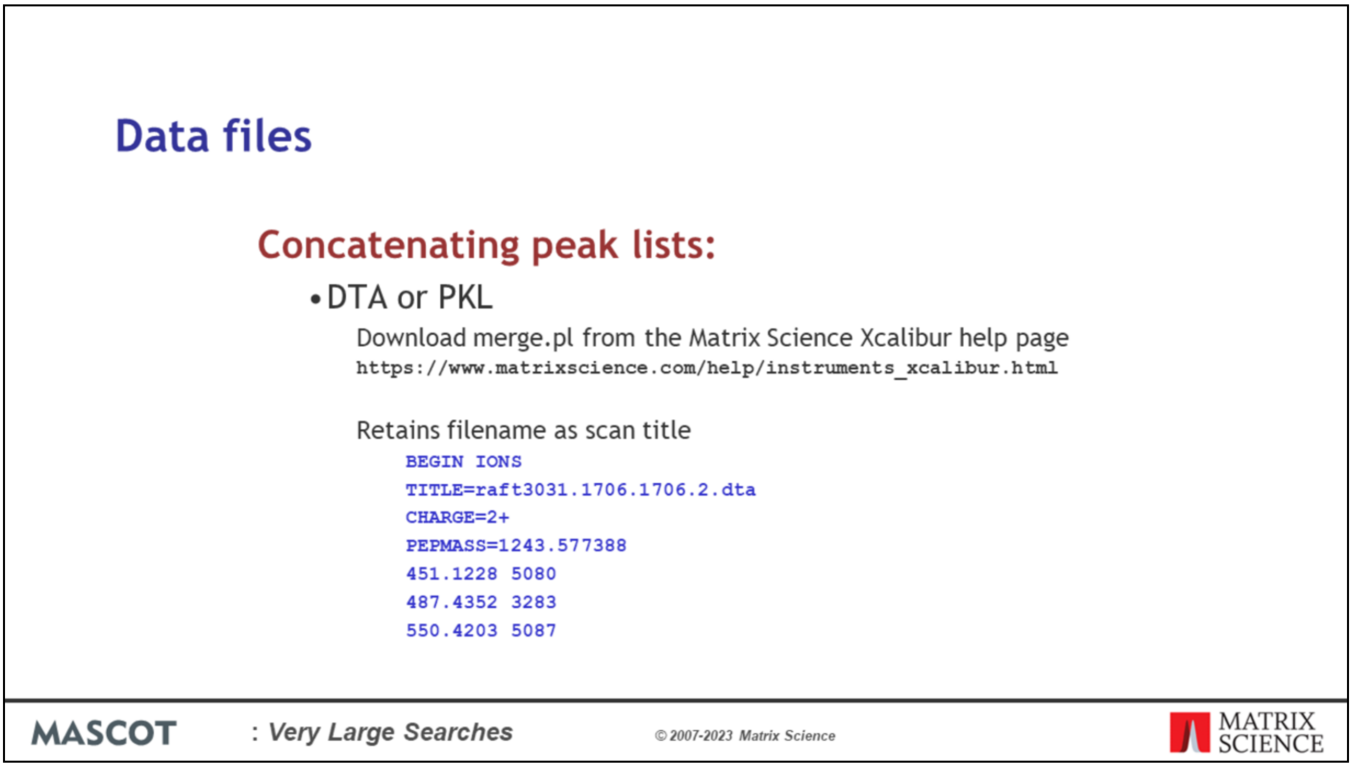
10 per page 1 2 3 4 5 6 ... 436 Next Expand all Collapse all

Accession contains Find Clear

1	1	2::A0A0D9R924	10970	Plectin OS=Chlorocebus sabaeus OX=60711 GN=...
	2	2::A0A0D9S7P7	856	Microtubule actin crosslinking factor 1 OS=Chloro...
	3	2::A0A0D9RLP2	362	Dystonin OS=Chlorocebus sabaeus OX=60711 G...
2	1	2::A0A0D9S9M0	9915	Uncharacterized protein OS=Chlorocebus sabaeu...
	3	2::A0A0D9R480	5241	Actin alpha cardiac muscle 1 OS=Chlorocebus sa...
	2	2::A0A0D9RYK7	6588	Uncharacterized protein OS=Chlorocebus sabaeu...

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
`https://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.

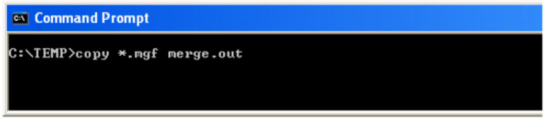
A nice feature of this script is that it puts the filename into the scan title, so you can tell which fraction a particular spectrum came from. The scan titles are displayed when you expand the rank column in the Mascot result report.

Data files

Concatenating peak lists:

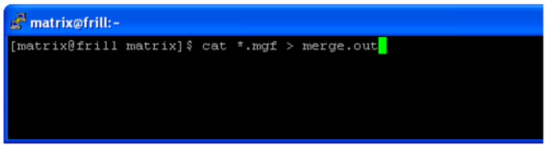
- MGF

Windows: copy



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

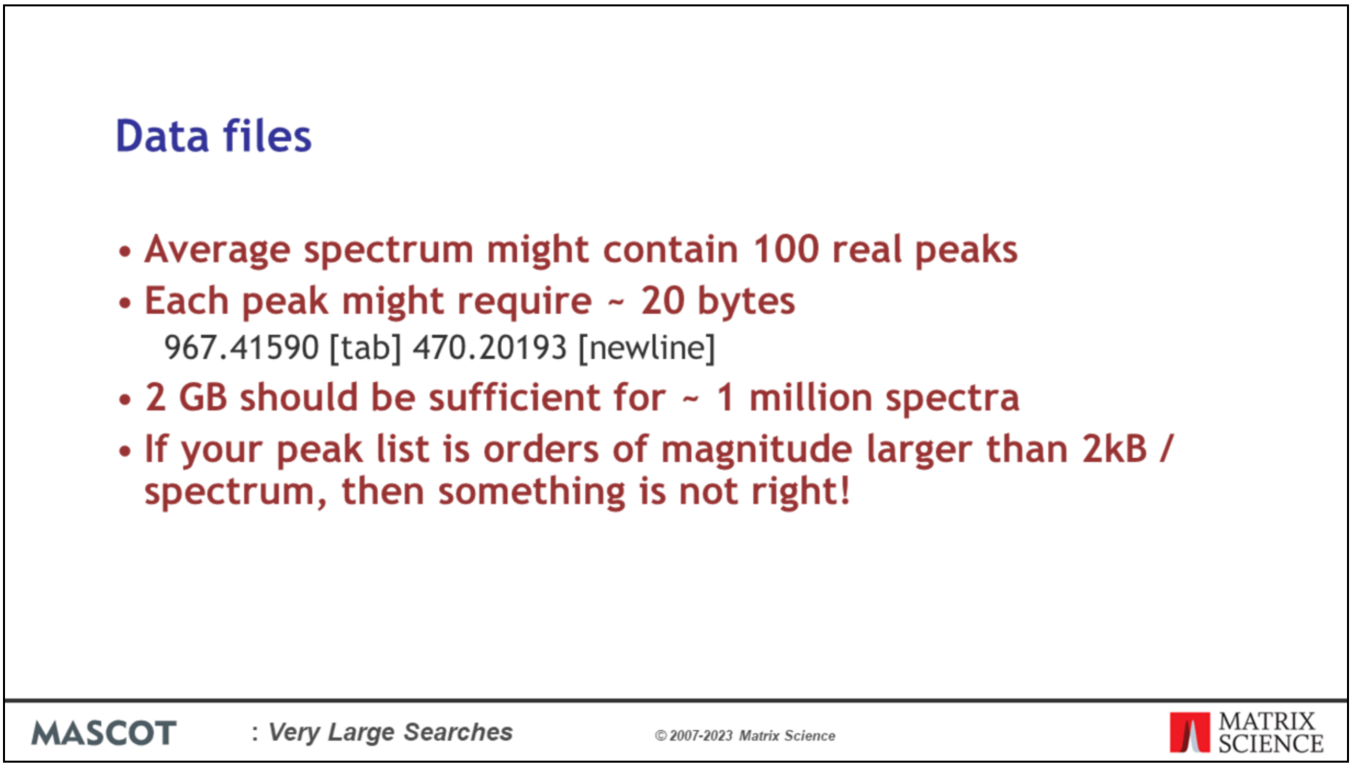
Unix: cat



```
[matrix@frill matrix]$ 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 strips out header lines when merging MGF files.



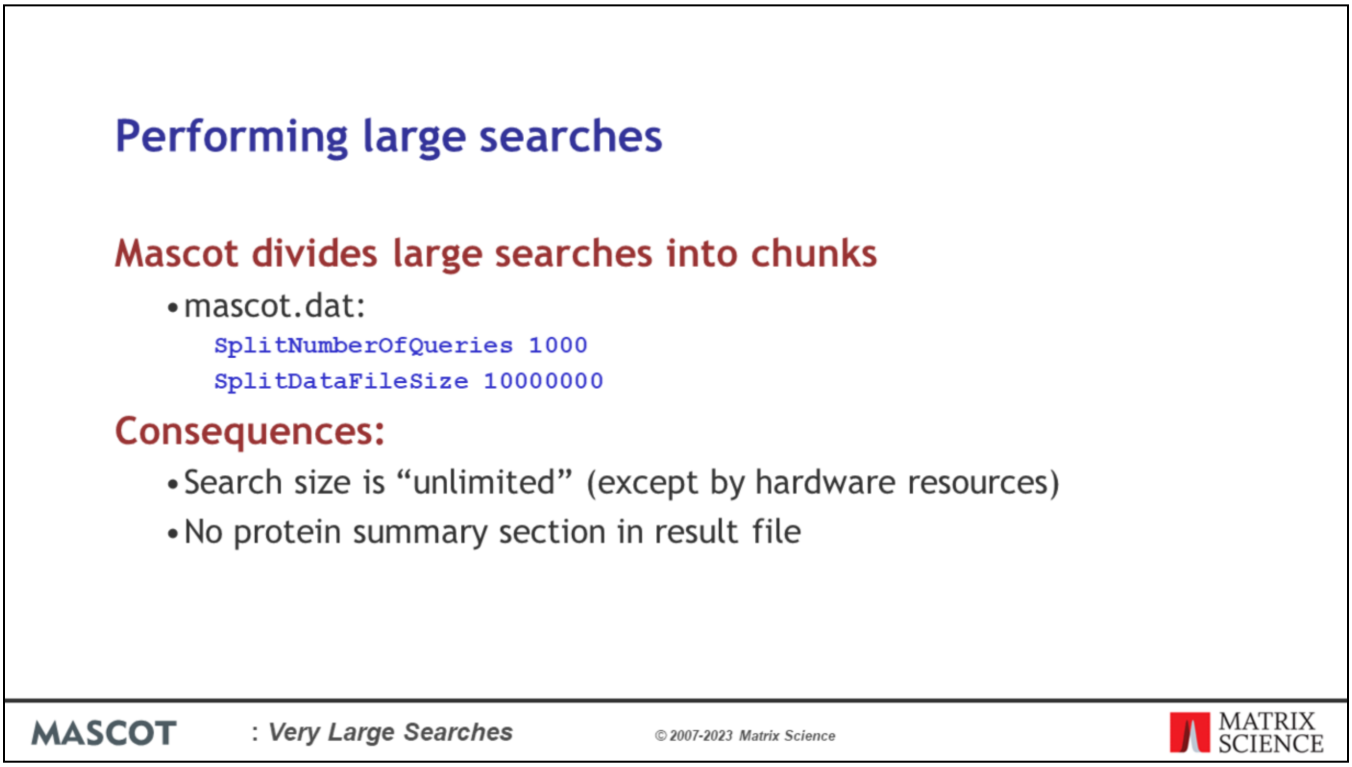
Data files

- 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

Mascot divides large searches into chunks

- mascot.dat:

```
SplitNumberOfQueries 1000
SplitDataFileSize 10000000
```

Consequences:

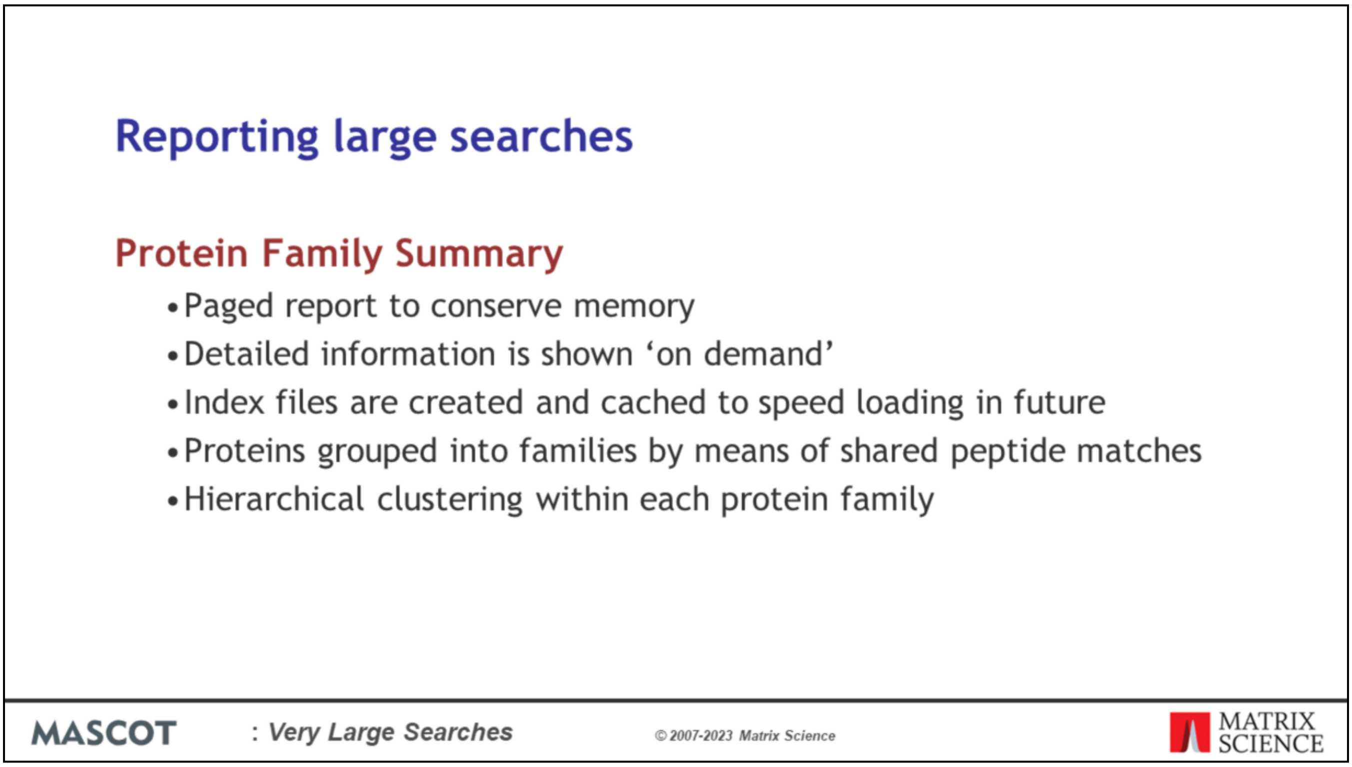
- Search size is “unlimited” (except by hardware resources)
- No protein summary section in result file

MASCOT : *Very Large Searches* © 2007-2023 Matrix Science MATRIX SCIENCE

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 often 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.



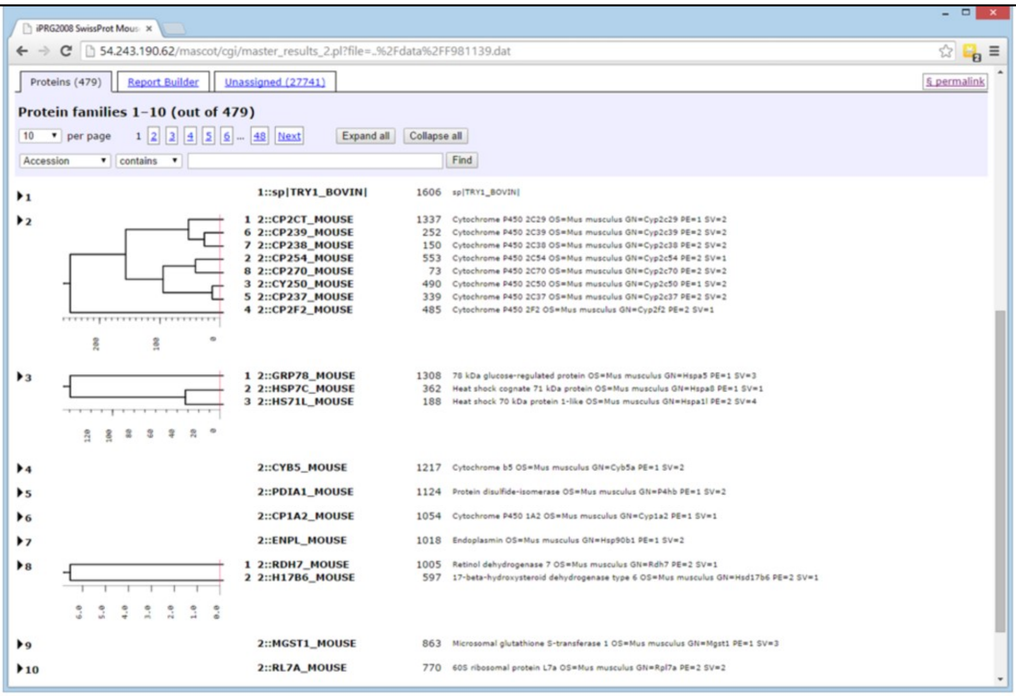
Reporting large searches

Protein Family Summary

- 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

MASCOT : *Very Large Searches* © 2007-2023 Matrix Science MATRIX SCIENCE

In *very* 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.



Protein families 1–10 (out of 479)

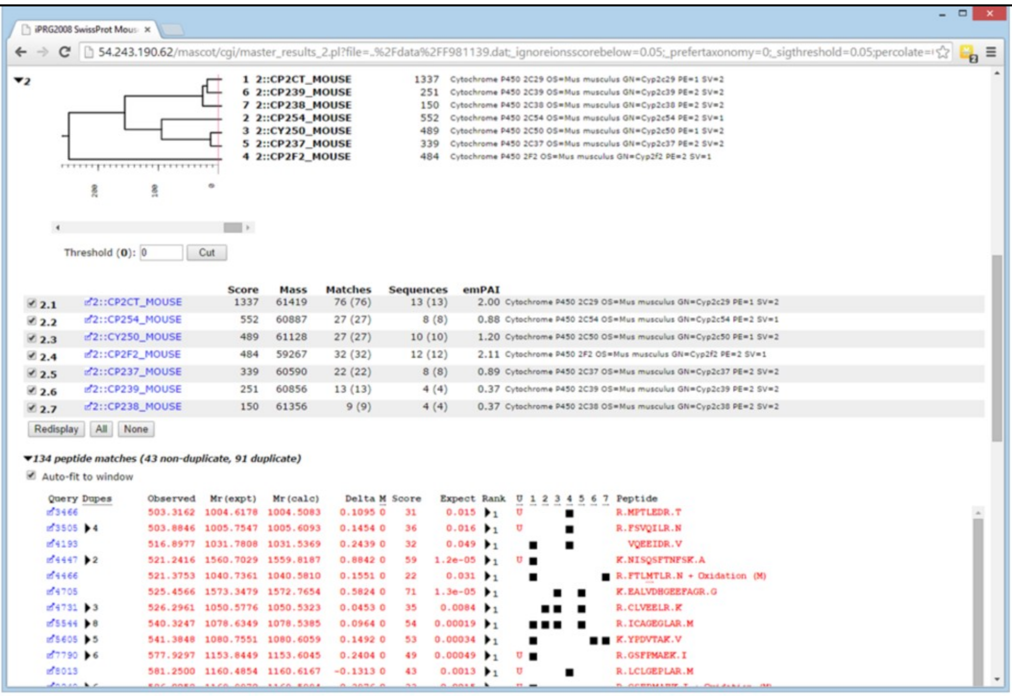
▸1		1::sp|TRY1_BOVIN|	1606	sp|TRY1_BOVIN|
▸2	1	2::CP2CT_MOUSE	1337	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=2
	6	2::CP239_MOUSE	252	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=2
	7	2::CP238_MOUSE	150	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=2
	2	2::CP254_MOUSE	553	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
	8	2::CP270_MOUSE	73	Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2
	3	2::CY250_MOUSE	490	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=2
	5	2::CP237_MOUSE	339	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=2
	4	2::CP2F2_MOUSE	485	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1
▸3	1	2::GRP78_MOUSE	1308	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
	2	2::HSP7C_MOUSE	362	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
	3	2::HS71L_MOUSE	188	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2 SV=4
▸4		2::CYB5_MOUSE	1217	Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2
▸5		2::PDIA1_MOUSE	1124	Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV=2
▸6		2::CP1A2_MOUSE	1054	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
▸7		2::ENPL_MOUSE	1018	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2
▸8	1	2::RDH7_MOUSE	1005	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
	2	2::H17B6_MOUSE	597	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus GN=Hsd17b6 PE=2 SV=1
▸9		2::MGST1_MOUSE	863	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgst1 PE=1 SV=3
▸10		2::RL7A_MOUSE	770	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=2

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.

You can also find links to older report formats, the Peptide Summary and Select Summary reports, but these are not suitable for today’s larger data sets.


iPRG2008 SwissProt Mous ×
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=
▾2
200
100
0
Threshold (0): 0
Cut
Redisplay
All
None

1	2::CP2CT_MOUSE	1337	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=2
6	2::CP239_MOUSE	251	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=2
7	2::CP238_MOUSE	150	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=2
2	2::CP254_MOUSE	552	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
3	2::CY250_MOUSE	489	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=2
5	2::CP237_MOUSE	339	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=2
4	2::CP2F2_MOUSE	484	Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1

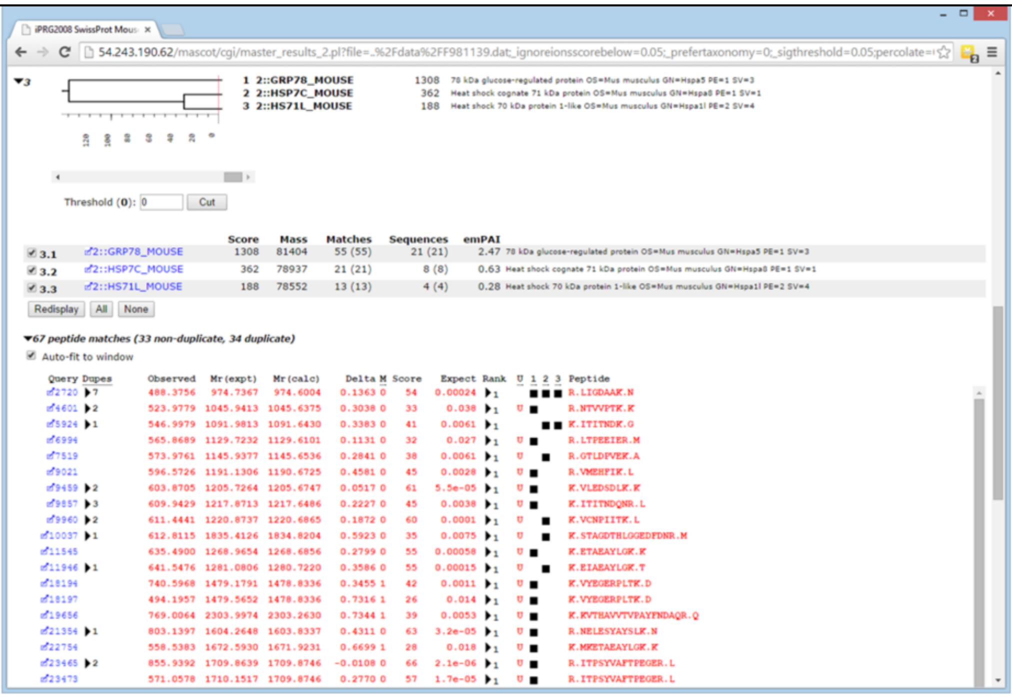
		Score	Mass	Matches	Sequences	emPAI	
☑ 2.1	2::CP2CT_MOUSE	1337	61419	76 (76)	13 (13)	2.00	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=2
☑ 2.2	2::CP254_MOUSE	552	60887	27 (27)	8 (8)	0.88	Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
☑ 2.3	2::CY250_MOUSE	489	61128	27 (27)	10 (10)	1.20	Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=2
☑ 2.4	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	2::CP237_MOUSE	339	60590	22 (22)	8 (8)	0.89	Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=2
☑ 2.6	2::CP239_MOUSE	251	60856	13 (13)	4 (4)	0.37	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=2
☑ 2.7	2::CP238_MOUSE	150	61356	9 (9)	4 (4)	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=2

▾134 peptide matches (43 non-duplicate, 91 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	3	4	5	6	7	Peptide
3466		503.3162	1004.6178	1004.5083	0.1095	0	31	0.015	▶1	U				■				R.MPTLEDR.T
3505	▶4	503.8846	1005.7547	1005.6093	0.1454	0	36	0.016	▶1	U				■				R.FSVQILR.N
4193		516.8977	1031.7808	1031.5369	0.2439	0	32	0.049	▶1		■			■				VQEEIDR.V
4447	▶2	521.2416	1560.7029	1559.8187	0.8842	0	59	1.2e-05	▶1	U	■							K.NISQSFTNFSK.A
4466		521.3753	1040.7361	1040.5810	0.1551	0	22	0.031	▶1		■						■	R.FTLMTLR.N + Oxidation (M)
4705		525.4566	1573.3479	1572.7654	0.5824	0	71	1.3e-05	▶1				■		■			K.EALVDHGEEFAGR.G
4731	▶3	526.2961	1050.5776	1050.5323	0.0453	0	35	0.0084	▶1			■	■		■			R.CLVEELR.K
5544	▶8	540.3247	1078.6349	1078.5385	0.0964	0	54	0.00019	▶1		■	■	■		■			R.ICAGEGLAR.M
5605	▶5	541.3848	1080.7551	1080.6059	0.1492	0	53	0.00034	▶1		■					■	■	K.YPDVTAK.V
7790	▶6	577.9297	1153.8449	1153.6045	0.2404	0	49	0.00049	▶1	U	■							R.GSFPMAEK.I
8013		581.2500	1160.4854	1160.6167	-0.1313	0	43	0.0013	▶1	U				■				R.LCLGEPLAR.M

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 report statistically significant peptide matches. The default significance threshold is 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.


iPRG2008 SwissProt Mous ×
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=
120
100
80
60
40
20
0

▼3

1	2::GRP78_MOUSE	1308	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
2	2::HSP7C_MOUSE	362	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
3	2::HS71L_MOUSE	188	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2 SV=4

Threshold (0): 0 Cut

		Score	Mass	Matches	Sequences	emPAI	
☑ 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=Hspa1l PE=2 SV=4

Redisplay All None

▼*67 peptide matches (33 non-duplicate, 34 duplicate)*

☑ Auto-fit to window

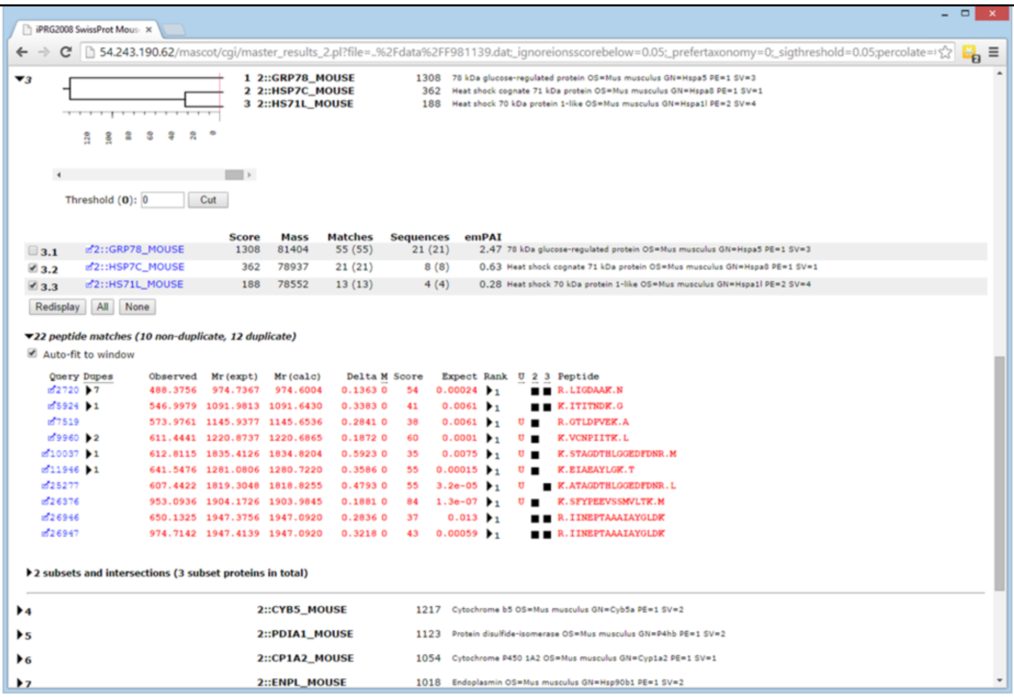
Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	3	Peptide
2720	▶7	488.3756	974.7367	974.6004	0.1363	0	54	0.00024	▶1		■	■	■	R.LIGDAAK.N
4601	▶2	523.9779	1045.9413	1045.6375	0.3038	0	33	0.038	▶1	U	■			R.NTVVPTK.K
5924	▶1	546.9979	1091.9813	1091.6430	0.3383	0	41	0.0061	▶1			■	■	K.ITITNDK.G
6994		565.8689	1129.7232	1129.6101	0.1131	0	32	0.027	▶1	U	■			R.LTPEEIER.M
7519		573.9761	1145.9377	1145.6536	0.2841	0	38	0.0061	▶1	U		■		R.GTLDPVEK.A
9021		596.5726	1191.1306	1190.6725	0.4581	0	45	0.0028	▶1	U	■			R.VMEHFIK.L
9459	▶2	603.8705	1205.7264	1205.6747	0.0517	0	61	5.5e-05	▶1	U	■			K.VLEDSDLK.K
9857	▶3	609.9429	1217.8713	1217.6486	0.2227	0	45	0.0038	▶1	U	■			K.ITITNDQNR.L
9960	▶2	611.4441	1220.8737	1220.6865	0.1872	0	60	0.0001	▶1	U		■		K.VCNPIITK.L
10037	▶1	612.8115	1835.4126	1834.8204	0.5923	0	35	0.0075	▶1	U		■		K.STAGDTHLGGEDFDNR.M
11545		635.4900	1268.9654	1268.6856	0.2799	0	55	0.00058	▶1	U	■			K.ETAEAYLGK.K
11946	▶1	641.5476	1281.0806	1280.7220	0.3586	0	55	0.00015	▶1	U		■		K.EIAEAYLGK.T
18194		740.5968	1479.1791	1478.8336	0.3455	1	42	0.0011	▶1	U	■			K.VYEGERPLTK.D
18197		494.1957	1479.5652	1478.8336	0.7316	1	26	0.014	▶1	U	■			K.VYEGERPLTK.D
19656		769.0064	2303.9974	2303.2630	0.7344	1	39	0.0053	▶1	U	■			K.KVTHAVVTVPAYFNDAQR.Q
21354	▶1	803.1397	1604.2648	1603.8337	0.4311	0	63	3.2e-05	▶1	U	■			R.NELESYAYSLK.N
22754		558.5383	1672.5930	1671.9231	0.6699	1	28	0.018	▶1	U	■			K.MKETAEAYLGK.K
23465	▶2	855.9392	1709.8639	1709.8746	-0.0108	0	66	2.1e-06	▶1	U	■			R.ITPSYVAFTPEGER.L
23473		571.0578	1710.1517	1709.8746	0.2770	0	57	1.7e-05	▶1	U	■			R.ITPSYVAFTPEGER.L

Moving down to family 3, the scale on the dendrogram is protein 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 Mous ×
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=
▼3
120
100
80
60
40
20
0

1	2::GRP78_MOUSE	1308	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
2	2::HSP7C_MOUSE	362	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
3	2::HS71L_MOUSE	188	Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2 SV=4

Threshold (0): 0 Cut

		Score	Mass	Matches	Sequences	emPAI	
□ 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=Hspa1l PE=2 SV=4

Redisplay All None

▼*22 peptide matches (10 non-duplicate, 12 duplicate)*

☑ Auto-fit to window

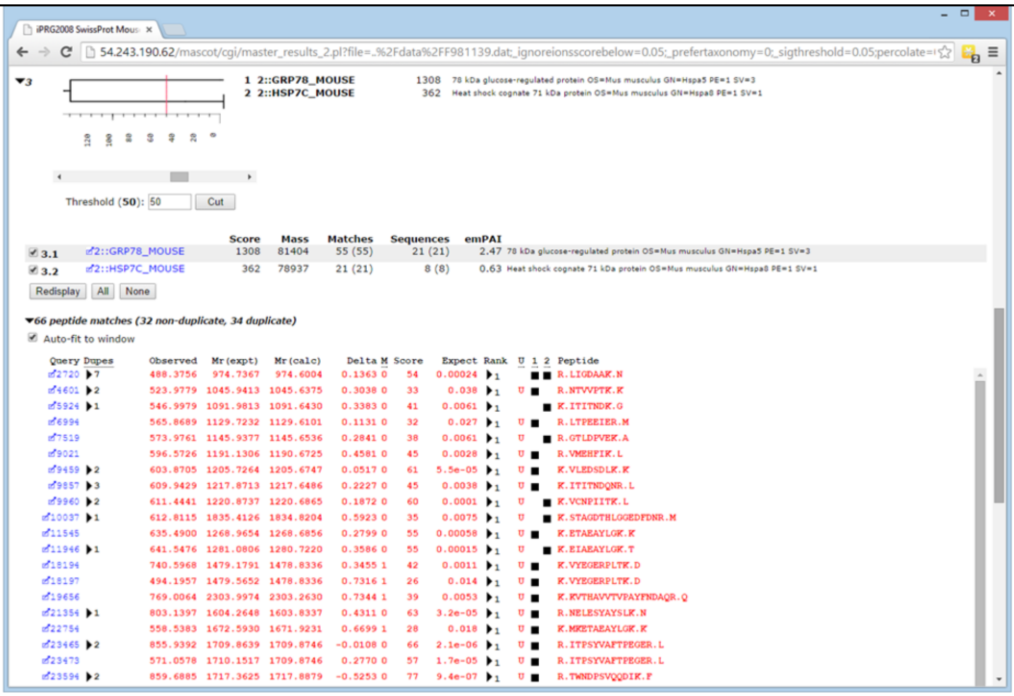
Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	2	3	Peptide
2720	▶7	488.3756	974.7367	974.6004	0.1363	0	54	0.00024	▶1		■	■	R.LIGDAAK.N
5924	▶1	546.9979	1091.9813	1091.6430	0.3383	0	41	0.0061	▶1		■	■	K.ITITNDK.G
7519		573.9761	1145.9377	1145.6536	0.2841	0	38	0.0061	▶1	U	■		R.GTLDPVEK.A
9960	▶2	611.4441	1220.8737	1220.6865	0.1872	0	60	0.0001	▶1	U	■		K.VCNPIITK.L
10037	▶1	612.8115	1835.4126	1834.8204	0.5923	0	35	0.0075	▶1	U	■		K.STAGDTHLGGEDFDNR.M
11946	▶1	641.5476	1281.0806	1280.7220	0.3586	0	55	0.00015	▶1	U	■		K.EIAEAYLGK.T
25277		607.4422	1819.3048	1818.8255	0.4793	0	55	3.2e-05	▶1	U		■	K.ATAGDTHLGGEDFDNR.L
26376		953.0936	1904.1726	1903.9845	0.1881	0	84	1.3e-07	▶1	U	■		K.SFYPEEVSSMVLTK.M
26946		650.1325	1947.3756	1947.0920	0.2836	0	37	0.013	▶1		■	■	R.IINEPTAAAIAYGLDK
26947		974.7142	1947.4139	1947.0920	0.3218	0	43	0.00059	▶1		■	■	R.IINEPTAAAIAYGLDK

▶2 subsets and intersections (3 subset proteins in total)

▶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=P4hb PE=1 SV=2
▶6	2::CP1A2_MOUSE	1054	Cytochrome P450 1A2 OS=Mus musculus GN=Cyp1a2 PE=1 SV=1
▶7	2::ENPL_MOUSE	1018	Endoplasmin OS=Mus musculus GN=Hsp90b1 PE=1 SV=2

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.


iPRG2008 SwissProt Mous ×
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=
▾3
120
100
80
60
40
20
0
Cut
Redisplay
All
None

1 2::GRP78_MOUSE 1308 78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
2 2::HSP7C_MOUSE 362 Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1

Threshold (**50**): 50

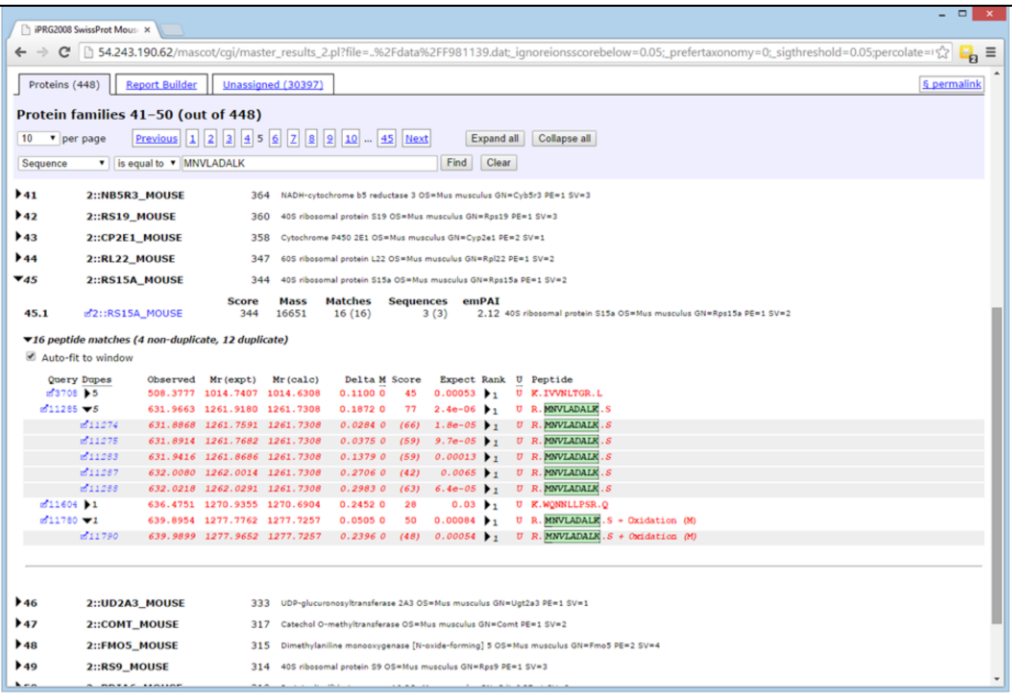
		Score	Mass	Matches	Sequences	emPAI	
☑ 3.1	2::GRP78_MOUSE	1308	81404	55 (55)	21 (21)	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
☑ 3.2	2::HSP7C_MOUSE	362	78937	21 (21)	8 (8)	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1

▾*66 peptide matches (32 non-duplicate, 34 duplicate)*

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	Peptide
2720	▶7	488.3756	974.7367	974.6004	0.1363	0	54	0.00024	▶1		■	■	R.LIGDAAK.N
4601	▶2	523.9779	1045.9413	1045.6375	0.3038	0	33	0.038	▶1	U	■		R.NTVVPTK.K
5924	▶1	546.9979	1091.9813	1091.6430	0.3383	0	41	0.0061	▶1			■	K.ITITNDK.G
6994		565.8689	1129.7232	1129.6101	0.1131	0	32	0.027	▶1	U	■		R.LTPEEIER.M
7519		573.9761	1145.9377	1145.6536	0.2841	0	38	0.0061	▶1	U		■	R.GTLDPVEK.A
9021		596.5726	1191.1306	1190.6725	0.4581	0	45	0.0028	▶1	U	■		R.VMEHFIK.L
9459	▶2	603.8705	1205.7264	1205.6747	0.0517	0	61	5.5e-05	▶1	U	■		K.VLEDSDLK.K
9857	▶3	609.9429	1217.8713	1217.6486	0.2227	0	45	0.0038	▶1	U	■		K.ITITNDQNR.L
9960	▶2	611.4441	1220.8737	1220.6865	0.1872	0	60	0.0001	▶1	U		■	K.VCNPIITK.L
10037	▶1	612.8115	1835.4126	1834.8204	0.5923	0	35	0.0075	▶1	U		■	K.STAGDTHLGGEDFDNR.M
11545		635.4900	1268.9654	1268.6856	0.2799	0	55	0.00058	▶1	U	■		K.ETAEAYLGK.K
11946	▶1	641.5476	1281.0806	1280.7220	0.3586	0	55	0.00015	▶1	U		■	K.EIAEAYLGK.T
18194		740.5968	1479.1791	1478.8336	0.3455	1	42	0.0011	▶1	U	■		K.VYEGERPLTK.D
18197		494.1957	1479.5652	1478.8336	0.7316	1	26	0.014	▶1	U	■		K.VYEGERPLTK.D
19656		769.0064	2303.9974	2303.2630	0.7344	1	39	0.0053	▶1	U	■		K.KVTHAVVTVPAYFNDAQR.Q
21354	▶1	803.1397	1604.2648	1603.8337	0.4311	0	63	3.2e-05	▶1	U	■		R.NELESYAYSLK.N
22754		558.5383	1672.5930	1671.9231	0.6699	1	28	0.018	▶1	U	■		K.MKETAEAYLGK.K
23465	▶2	855.9392	1709.8639	1709.8746	-0.0108	0	66	2.1e-06	▶1	U	■		R.ITPSYVAFTPEGER.L
23473		571.0578	1710.1517	1709.8746	0.2770	0	57	1.7e-05	▶1	U	■		R.ITPSYVAFTPEGER.L
23594	▶2	859.6885	1717.3625	1717.8879	-0.5253	0	77	9.4e-07	▶1	U	■		R.TWNDPSVQQDIK.F

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.



54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=

Protein families 41–50 (out of 448)

10 per page

Sequence ▾ is equal to ▾ MNVLADALK

▸41	2::NB5R3_MOUSE	364	NADH-cytochrome b5 reductase 3 OS=Mus musculus GN=Cyb5r3 PE=1 SV=3
▸42	2::RS19_MOUSE	360	40S ribosomal protein S19 OS=Mus musculus GN=Rps19 PE=1 SV=3
▸43	2::CP2E1_MOUSE	358	Cytochrome P450 2E1 OS=Mus musculus GN=Cyp2e1 PE=2 SV=1
▸44	2::RL22_MOUSE	347	60S ribosomal protein L22 OS=Mus musculus GN=Rpl22 PE=1 SV=2
▾45	2::RS15A_MOUSE	344	40S ribosomal protein S15a OS=Mus musculus GN=Rps15a PE=1 SV=2

		Score	Mass	Matches	Sequences	emPAI	
45.1	2::RS15A_MOUSE	344	16651	16 (16)	3 (3)	2.12	40S ribosomal protein S15a OS=Mus musculus GN=Rps15a PE=1 SV=2

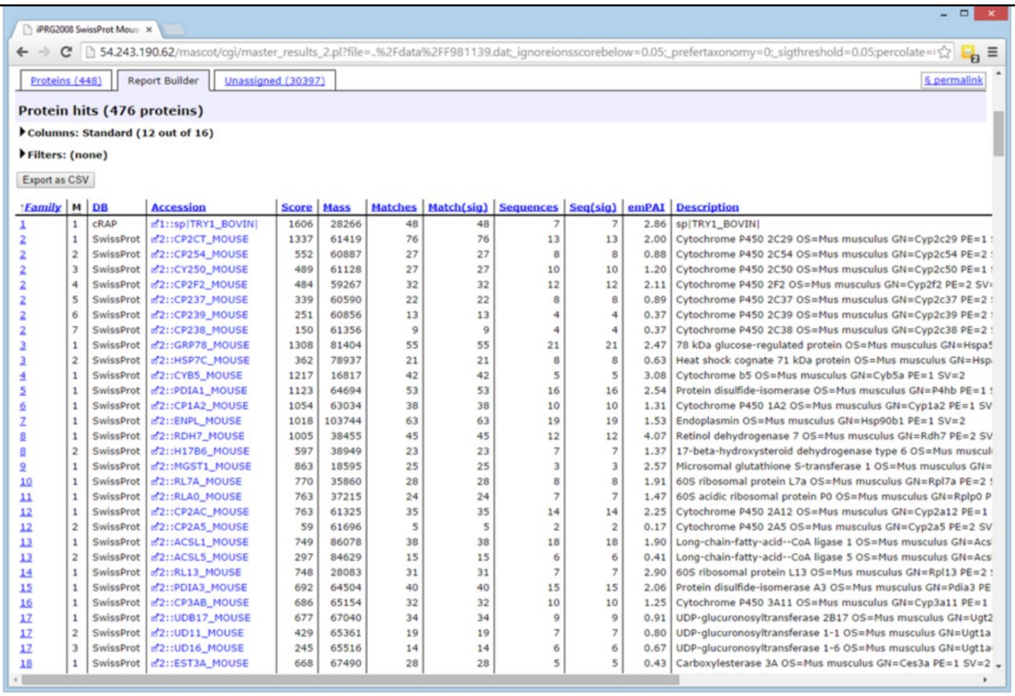
▾*16 peptide matches (4 non-duplicate, 12 duplicate)*

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	Peptide
3708	▸5	508.3777	1014.7407	1014.6308	0.1100	0	45	0.00053	▸1	U	K.IVVNLTGR.L
11285	▾5	631.9663	1261.9180	1261.7308	0.1872	0	77	2.4e-06	▸1	U	R.MNVLADALK.S
11274		631.8868	1261.7591	1261.7308	0.0284	0	(66)	1.8e-05	▸1	U	R.MNVLADALK.S
11275		631.8914	1261.7682	1261.7308	0.0375	0	(59)	9.7e-05	▸1	U	R.MNVLADALK.S
11283		631.9416	1261.8686	1261.7308	0.1379	0	(59)	0.00013	▸1	U	R.MNVLADALK.S
11287		632.0080	1262.0014	1261.7308	0.2706	0	(42)	0.0065	▸1	U	R.MNVLADALK.S
11288		632.0218	1262.0291	1261.7308	0.2983	0	(63)	6.4e-05	▸1	U	R.MNVLADALK.S
11604	▸1	636.4751	1270.9355	1270.6904	0.2452	0	28	0.03	▸1	U	K.WQNNLLPSR.Q
11780	▾1	639.8954	1277.7762	1277.7257	0.0505	0	50	0.00084	▸1	U	R.MNVLADALK.S + Oxidation (M)
11790		639.9899	1277.9652	1277.7257	0.2396	0	(48)	0.00054	▸1	U	R.MNVLADALK.S + Oxidation (M)

▸46	2::UD2A3_MOUSE	333	UDP-glucuronosyltransferase 2A3 OS=Mus musculus GN=Ugt2a3 PE=1 SV=1
▸47	2::COMT_MOUSE	317	Catechol O-methyltransferase OS=Mus musculus GN=Comt PE=1 SV=2
▸48	2::FMO5_MOUSE	315	Dimethylaniline monooxygenase [N-oxide-forming] 5 OS=Mus musculus GN=Fmo5 PE=2 SV=4
▸49	2::RS9_MOUSE	314	40S ribosomal protein S9 OS=Mus musculus GN=Rps9 PE=1 SV=3

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.



iPRG2008 SwissProt Mous
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=
Proteins (448)
Report Builder
Unassigned (30397)
§ permalink

Protein hits (476 proteins)

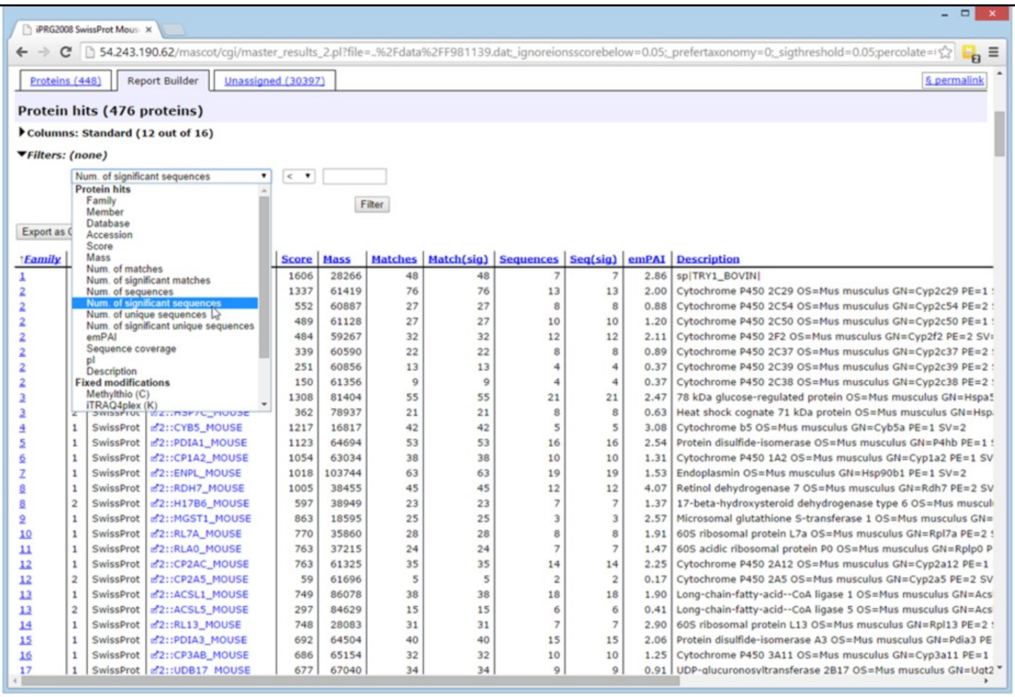
▸Columns: Standard (12 out of 16)

▸Filters: (none)

Export as CSV

↑Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	cRAP	1::sp|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
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=Hspa
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	Endoplasmin 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::CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 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::CP3AB_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

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



iPRG2008 SwissProt Mous ×

54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=

Proteins (448) | Report Builder | Unassigned (30397) | § permalink

Protein hits (476 proteins)

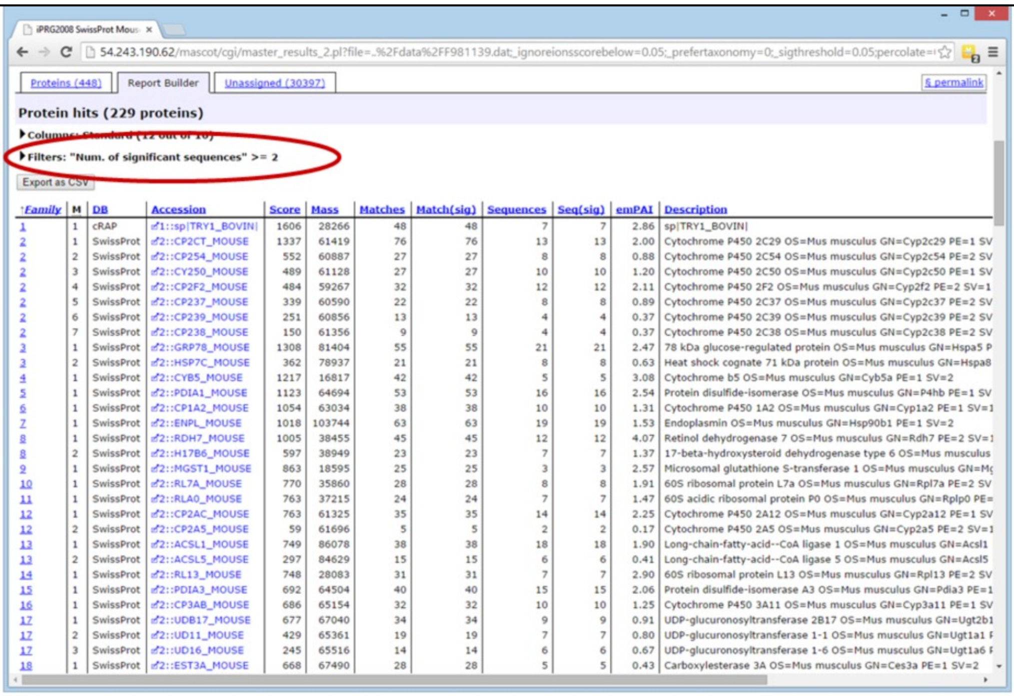
▸ **Columns: Standard (12 out of 16)**

▾ ***Filters: (none)***

Num. of significant sequences ▾ | < ▾ | Filter

↑Family				Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1				1606	28266	48	48	7	7	2.86	sp|TRY1_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 2C39 OS=Mus musculus GN=Cyp2c39 PE=2
2				150	61356	9	9	4	4	0.37	Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2
3				1308	81404	55	55	21	21	2.47	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5
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	Endoplasmin 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::CP2A5_MOUSE	59	61696	5	5	2	2	0.17	Cytochrome P450 2A5 OS=Mus musculus GN=Cyp2a5 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::CP3AB_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

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



iPRG2008 SwissProt Mous
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=
Proteins (448)
Report Builder
Unassigned (30397)
§ permalink
Export as CSV

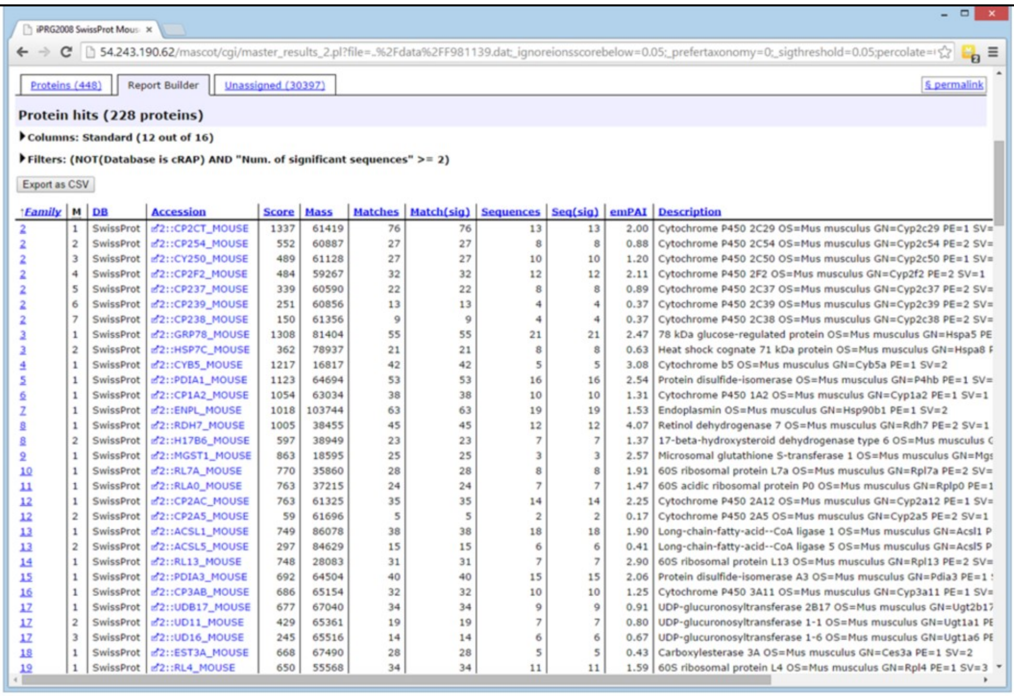
Protein hits (229 proteins)

▸ Columns: Standard (12 out of 16)

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

↑Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	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::HSP7C_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	Endoplasmin 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=Mg
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::CP2AC_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::ACSL5_MOUSE	297	84629	15	15	6	6	0.41	Long-chain-fatty-acid--CoA ligase 5 OS=Mus musculus GN=Acsl5
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::CP3AB_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::UD11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 F
17	3	SwissProt	2::UD16_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

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


iPRG2008 SwissProt Mous ×
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=
Proteins (448)
Report Builder
Unassigned (30397)
§ permalink
Export as CSV

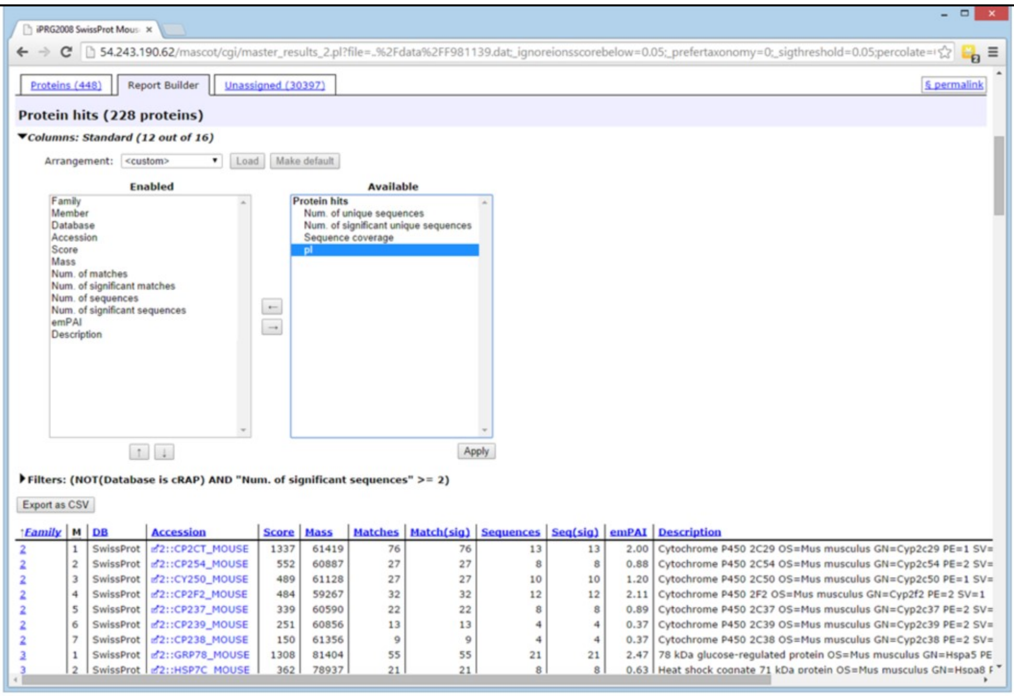
Protein hits (228 proteins)

▸Columns: Standard (12 out of 16)

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

↑Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
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 PE
3	2	SwissProt	2::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 F
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	Endoplasmin 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 C
9	1	SwissProt	2::MGST1_MOUSE	863	18595	25	25	3	3	2.57	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgs
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=1
12	1	SwissProt	2::CP2AC_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 P
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=Acsl5 P
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::CP3AB_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=Ugt2b17
17	2	SwissProt	2::UD11_MOUSE	429	65361	19	19	7	7	0.80	UDP-glucuronosyltransferase 1-1 OS=Mus musculus GN=Ugt1a1 PE
17	3	SwissProt	2::UD16_MOUSE	245	65516	14	14	6	6	0.67	UDP-glucuronosyltransferase 1-6 OS=Mus musculus GN=Ugt1a6 PE
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
19	1	SwissProt	2::RL4_MOUSE	650	55568	34	34	11	11	1.59	60S ribosomal protein L4 OS=Mus musculus GN=Rpl4 PE=1 SV=3

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



54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat;_ignoreionsscorebelow=0.05;_prefertaxonomy=0;_sigthreshold=0.05;percolate=

Protein hits (228 proteins)

▼Columns: Standard (12 out of 16)

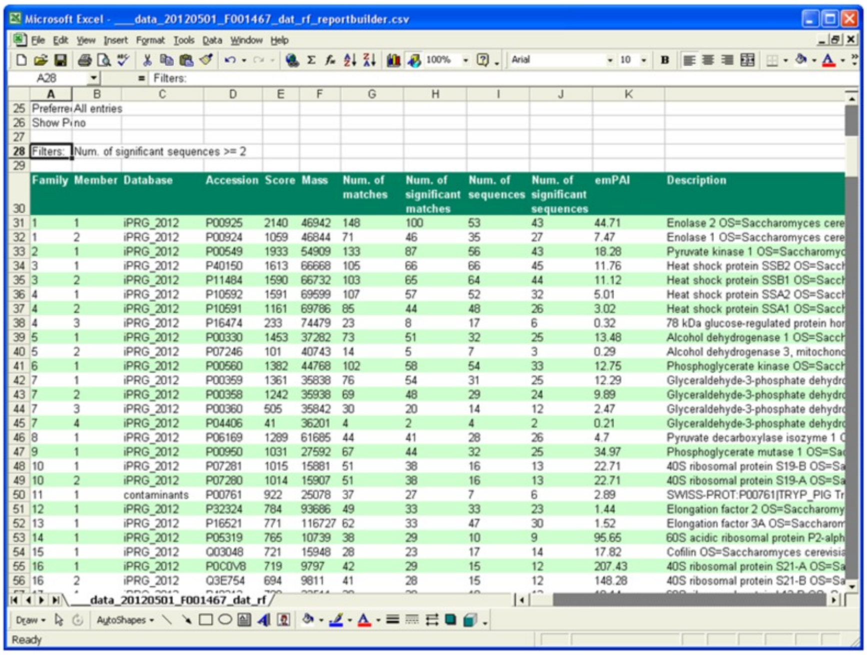
Arrangement: <custom>

Enabled	Available
Family	Protein hits
Member	Num. of unique sequences
Database	Num. of significant unique sequences
Accession	Sequence coverage
Score	pI
Mass	
Num. of matches	
Num. of significant matches	
Num. of sequences	
Num. of significant sequences	
emPAI	
Description	

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

↑Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
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 PE
3	2	SwissProt	2::HSP7C_MOUSE	362	78937	21	21	8	8	0.63	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 F

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



Microsoft Excel - ____data_20120501_F001467_dat_rf_reportbuilder.csv

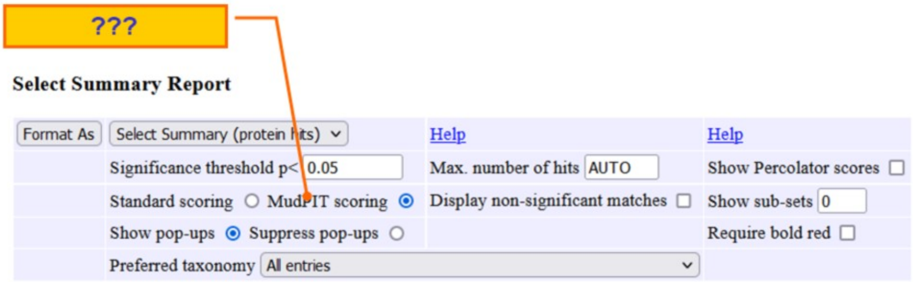
A28 = Filters:

	A	B	C	D	E	F	G	H	I	J	K	
25	Preferre	All entries										
26	Show P	no										
27												
28	Filters:	Num. of significant sequences >= 2										
29												
30	Family	Member	Database	Accession	Score	Mass	Num. of matches	Num. of significant matches	Num. of sequences	Num. of significant sequences	emPAI	Description
31	1	1	iPRG_2012	P00925	2140	46942	148	100	53	43	44.71	Enolase 2 OS=Saccharomyces cere
32	1	2	iPRG_2012	P00924	1059	46844	71	46	35	27	7.47	Enolase 1 OS=Saccharomyces cere
33	2	1	iPRG_2012	P00549	1933	54909	133	87	56	43	18.28	Pyruvate kinase 1 OS=Saccharomyc
34	3	1	iPRG_2012	P40150	1613	66668	105	66	66	45	11.76	Heat shock protein SSB2 OS=Sacch
35	3	2	iPRG_2012	P11484	1590	66732	103	65	64	44	11.12	Heat shock protein SSB1 OS=Sacch
36	4	1	iPRG_2012	P10592	1591	69599	107	57	52	32	5.01	Heat shock protein SSA2 OS=Sacch
37	4	2	iPRG_2012	P10591	1161	69786	85	44	48	26	3.02	Heat shock protein SSA1 OS=Sacch
38	4	3	iPRG_2012	P16474	233	74479	23	8	17	6	0.32	78 kDa glucose-regulated protein hor
39	5	1	iPRG_2012	P00330	1453	37282	73	51	32	25	13.48	Alcohol dehydrogenase 1 OS=Sacch
40	5	2	iPRG_2012	P07246	101	40743	14	5	7	3	0.29	Alcohol dehydrogenase 3, mitochonc
41	6	1	iPRG_2012	P00560	1382	44768	102	58	54	33	12.75	Phosphoglycerate kinase OS=Sacch
42	7	1	iPRG_2012	P00359	1361	35838	76	54	31	25	12.29	Glyceraldehyde-3-phosphate dehydro
43	7	2	iPRG_2012	P00358	1242	35938	69	48	29	24	9.89	Glyceraldehyde-3-phosphate dehydro
44	7	3	iPRG_2012	P00360	505	35842	30	20	14	12	2.47	Glyceraldehyde-3-phosphate dehydro
45	7	4	iPRG_2012	P04406	41	36201	4	2	4	2	0.21	Glyceraldehyde-3-phosphate dehydro
46	8	1	iPRG_2012	P06169	1289	61685	44	41	28	26	4.7	Pyruvate decarboxylase isozyme 1 O
47	9	1	iPRG_2012	P00950	1031	27592	67	44	32	25	34.97	Phosphoglycerate mutase 1 OS=Sac
48	10	1	iPRG_2012	P07281	1015	15881	51	38	16	13	22.71	40S ribosomal protein S19-B OS=Sa
49	10	2	iPRG_2012	P07280	1014	15907	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|TRYP_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	P05319	765	10739	38	29	10	9	95.65	60S acidic ribosomal protein P2-alph
54	15	1	iPRG_2012	Q03048	721	15948	28	23	17	14	17.82	Cofilin OS=Saccharomyces cerevisia
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

File Edit View Insert Format Tools Data Window Help
100%
Arial
10
B
____data_20120501_F001467_dat_rf
Draw
AutoShapes
Ready

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


???

Select Summary Report

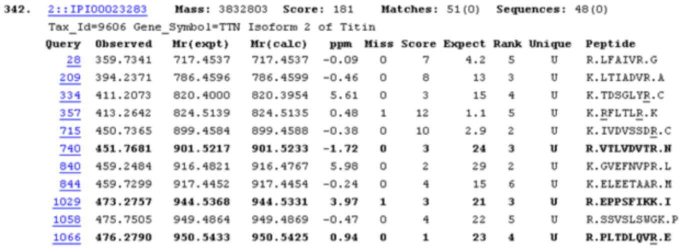
Format As	Select Summary (protein hits)	Help	Help
	Significance threshold p< 0.05	Max. number of hits AUTO	Show Percolator scores □
	Standard scoring ○ MudPIT scoring ●	Display non-significant matches □	Show sub-sets 0
	Show pop-ups ● Suppress pop-ups ○		Require bold red □
	Preferred taxonomy All entries		

The older Peptide Summary and Select summary (Proteins) reports have an options choose between Standard scoring and MudPIT scoring. The standard protein family report always uses MudPIT scoring. What do we mean by Standard scoring and MudPIT scoring?

Protein Scores for MS/MS Searches

Standard protein score

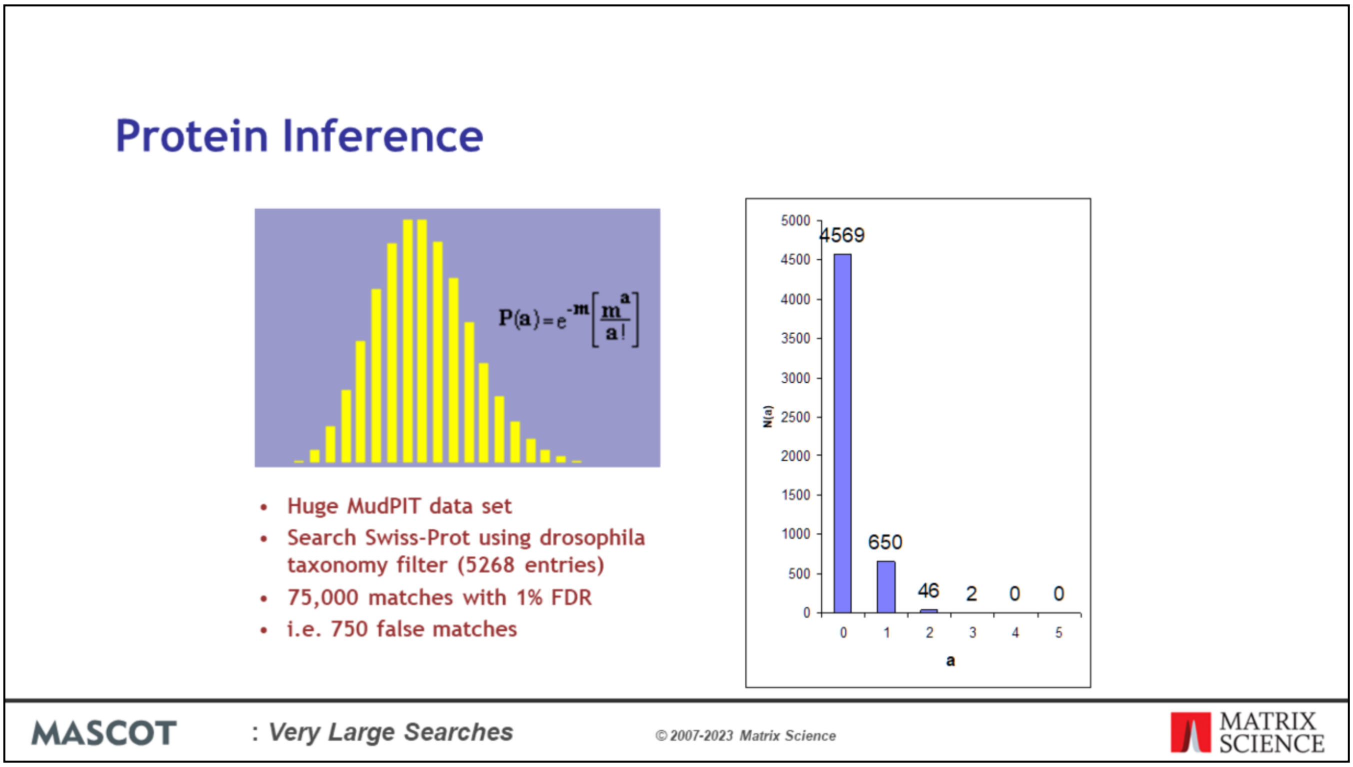
- the sum of the ions scores
- excluding the scores for duplicate matches, which are shown in parentheses
- correction to reduce the contribution of low-scoring random matches



```
342.  2::IPI00023283    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
        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.RFLTLR.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.M
      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
```

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

$$P(a)=e^{-m}\left[\frac{m^a}{a!}\right]$$

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

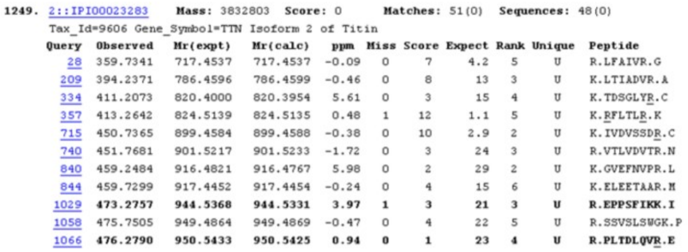
A protein with matches to just a single peptide sequence is commonly referred to as a "one-hit wonder" and is often treated as suspect. This is actually a slight over-simplification. In a search with a large number of spectra and a small database, even though the peptide false discovery rate is low, 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



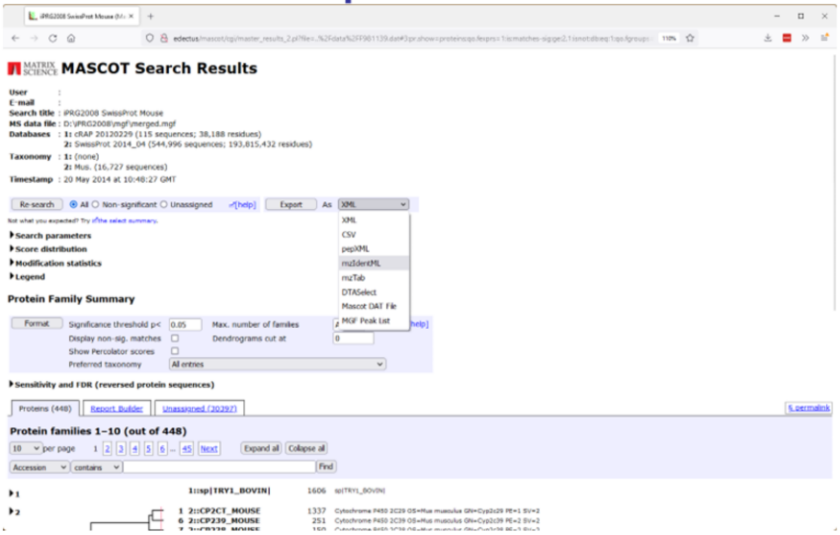
```
1249. 2::IPI00023283     Mass: 3832803  Score: 0      Matches: 51(0)  Sequences: 48(0)
      Tax_Id=9606 Gene_Symbol=TTN Isoform 2 of Titin
      Query  Observed   Mr(expt)   Mr(calc)    ppm  Miss Score Expect Rank Unique  Peptide
         28  359.7341   717.4537   717.4537  -0.09    0     7    4.2    5     U    R.LFAIVR.G
        209  394.2371   786.4596   786.4599  -0.46    0     8     13    3     U    K.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.RFLTLR.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.M
       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
```

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 and always used on the Protein Family Summary. 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



iPRG2008 SwissProt Mouse (M...
edectus/mascot/cgi/master_results_2.pl?file=../%2Fdata%2FF981139.dat#3:pr.show=proteins
MATRIX SCIENCE

MASCOT Search Results

User :
E-mail :
Search title : iPRG2008 SwissProt Mouse
MS data file : D:\iPRG2008\mgf\merged.mgf
Databases : **1:** cRAP 20120229 (115 sequences; 38,188 residues)
2: SwissProt 2014_04 (544,996 sequences; 193,815,432 residues)
Taxonomy : **1:** (none)
2: Mus. (16,727 sequences)
Timestamp : 20 May 2014 at 10:48:27 GMT

Not what you expected? Try the select summary.

Protein Family Summary

Significance threshold p< 0.05
Max. number of families
Display non-sig. matches □
Dendrograms cut at 0
Show Percolator scores □
Preferred taxonomy All entries

Protein families 1–10 (out of 448)

1		1::sp|TRY1_BOVIN|	1606	sp|TRY1_BOVIN|
2	1	2::CP2CT_MOUSE	1337	Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=2
	6	2::CP239_MOUSE	251	Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=2

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 Python, C#, 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, v1.8, if you want to export to Protein Prophet from ISB.

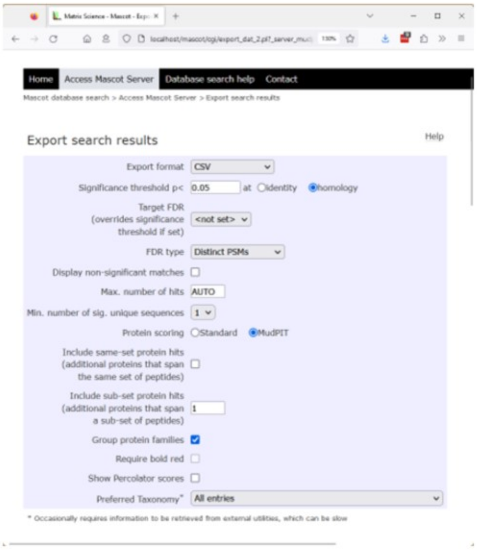
mzIdentML and mzTab are the standard formats from PSI for search result interchange. Mascot provides a very full implementation of mzIdentML and this is the one to choose if you are writing new application software that will use Mascot results.

DTASelect, v1.9, 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


Matrix Science - Mascot - Exp
localhost/mascot/cgi/export_dat_2.pl?_server_mu
130%
Home
Access Mascot Server
Database search help
Contact
Help

Mascot database search > Access Mascot Server > Export search results

Export search results

Export format: CSV

Significance threshold p< 0.05 at ○identity ●homology

Target FDR (overrides significance threshold if set): <not set>

FDR type: Distinct PSMs

Display non-significant matches □

Max. number of hits: AUTO

Min. number of sig. unique sequences: 1

Protein scoring: ○Standard ●MudPIT

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

Group protein families ☑

Require bold red □

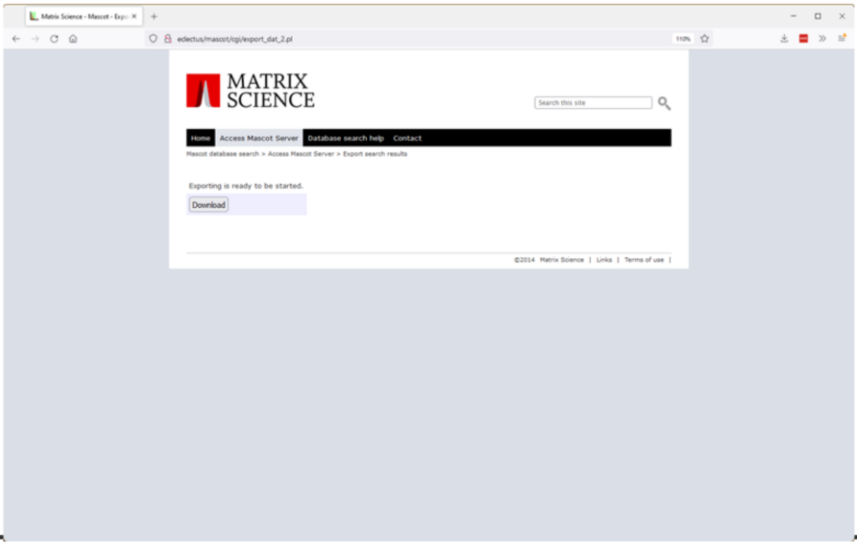
Show Percolator scores □

Preferred Taxonomy*: All entries

* Occasionally requires information to be retrieved from external utilities, which can be slow

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.

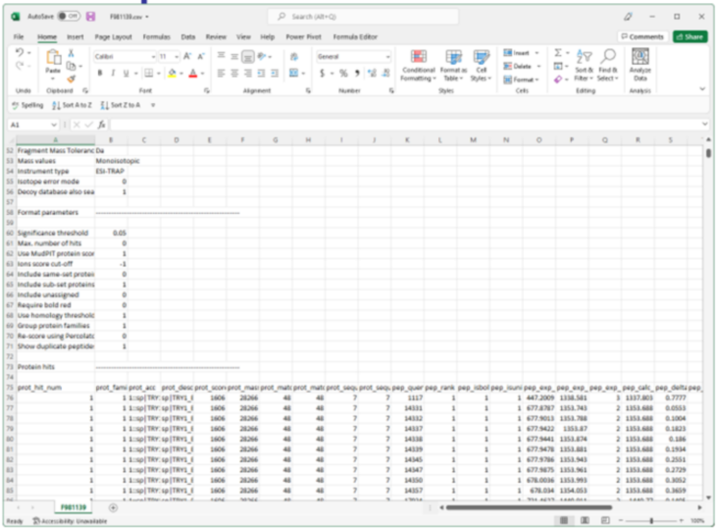
Search result export


Matrix Science - Mascot - Expo
eclectus/mascot/cgi/export_dat_2.pl
110%
MATRIX SCIENCE
Search this site
Home
Access Mascot Server
Database search help
Contact
Mascot database search > Access Mascot Server > Export search results
Download
©2014 Matrix Science | Links | Terms of use |

Exporting is ready to be started.

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. One the download is finished, you can open it into Excel.

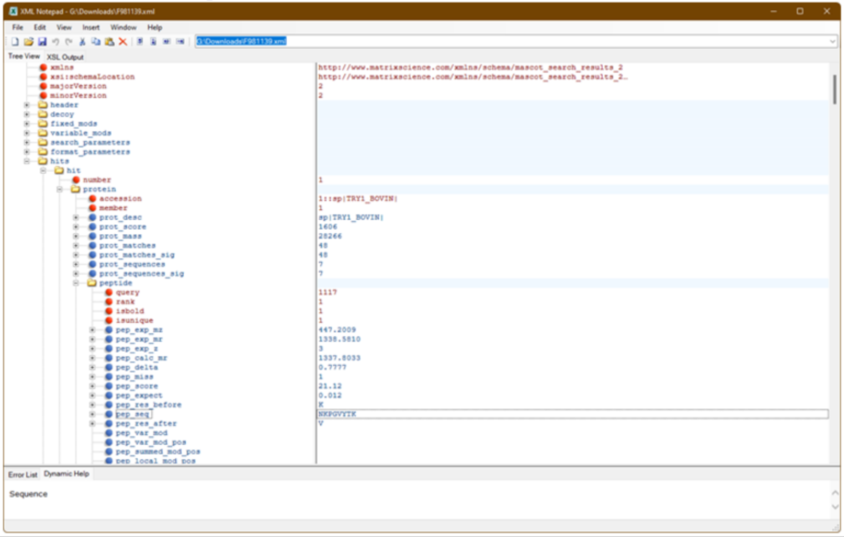
Search result export



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
52	Fragment Mass Toleranc	Da																		
53	Mass values	Monoisotopic																		
54	Instrument type	ESI-TRAP																		
55	Isotope error mode	0																		
56	Decoy database also sea	1																		
57																				
58	Format parameters	--																		
59																				
60	Significance threshold	0.05																		
61	Max. number of hits	0																		
62	Use MudPIT protein scor	1																		
63	Ions score cut-off	-1																		
64	Include same-set protei	0																		
65	Include sub-set proteins	1																		
66	Include unassigned	0																		
67	Require bold red	0																		
68	Use homology threshold	1																		
69	Group protein families	1																		
70	Re-score using Percolatc	0																		
71	Show duplicate peptide:	1																		
72																				
73	Protein hits	--																		
74																				
75	prot_hit_num	prot_fami	prot_acc	prot_desc	prot_scor	prot_mas	prot_matc	prot_matc	prot_sequ	prot_sequ	pep_quer	pep_rank	pep_isbol	pep_isuni	pep_exp_	pep_exp_	pep_exp_	pep_calc_	pep_delta	pep_
76	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	1117	1	1	1	447.2009	1338.581	3	1337.803	0.7777	
77	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14331	1	1	1	677.8787	1353.743	2	1353.688	0.0553	
78	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14332	1	1	1	677.9013	1353.788	2	1353.688	0.1004	
79	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14337	1	1	1	677.9422	1353.87	2	1353.688	0.1823	
80	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14338	1	1	1	677.9441	1353.874	2	1353.688	0.186	
81	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14339	1	1	1	677.9478	1353.881	2	1353.688	0.1934	
82	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14345	1	1	1	677.9786	1353.943	2	1353.688	0.2551	
83	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14347	1	1	1	677.9875	1353.961	2	1353.688	0.2729	
84	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14350	1	1	1	678.0036	1353.993	2	1353.688	0.3052	
85	1	1	1::sp|TRY:	sp|TRY1_I	1606	28266	48	48	7	7	14357	1	1	1	678.034	1354.053	2	1353.688	0.3659	

Much easier and safer than “screen scraping”.

Search result export



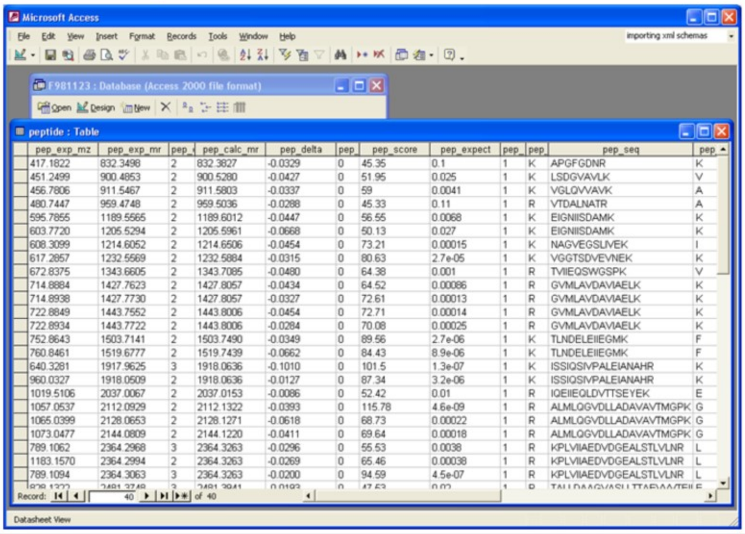
```
xmlns                    http://www.matrixscience.com/xmlns/schema/mascot_search_results_2
xsi:schemaLocation       http://www.matrixscience.com/xmlns/schema/mascot_search_results_2…
majorVersion             2
minorVersion             2
header
decoy
fixed_mods
variable_mods
search_parameters
format_parameters
hits
  hit
    number               1
    protein
      accession          1::sp|TRY1_BOVIN|
      member             1
      prot_desc          sp|TRY1_BOVIN|
      prot_score         1606
      prot_mass          28266
      prot_matches       48
      prot_matches_sig   48
      prot_sequences     7
      prot_sequences_sig 7
      peptide
        query            1117
        rank             1
        isbold           1
        isunique         1
        pep_exp_mz       447.2009
        pep_exp_mr       1338.5810
        pep_exp_z        3
        pep_calc_mr      1337.8033
        pep_delta        0.7777
        pep_miss         1
        pep_score        21.12
        pep_expect       0.012
        pep_res_before   K
        pep_seq          NKPGVYTK
        pep_res_after    V
        pep_var_mod
        pep_var_mod_pos
        pep_summed_mod_pos
        pep_local_mod_pos
```

Sequence

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.

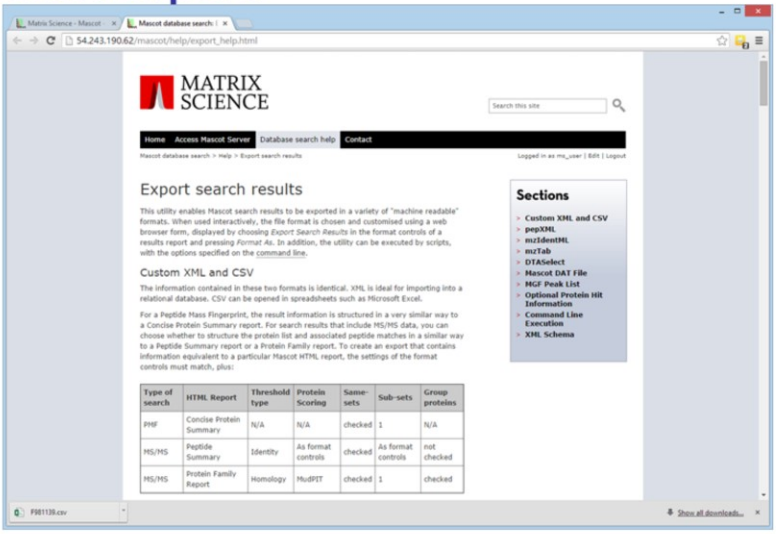
Search result export


Microsoft Access
File Edit View Insert Format Records Tools Window Help
importing xml schemas
F981123 : Database (Access 2000 file format)
Open Design New
peptide : Table
Record: 40 of 40
Datasheet View

pep_exp_mz	pep_exp_mr	pep_	pep_calc_mr	pep_delta	pep_	pep_score	pep_expect	pep_	pep_	pep_seq	pep_
417.1822	832.3498	2	832.3827	-0.0329	0	45.35	0.1	1	K	APGFGDNR	K
451.2499	900.4853	2	900.5280	-0.0427	0	51.95	0.025	1	K	LSDGVAVLK	V
456.7806	911.5467	2	911.5803	-0.0337	0	59	0.0041	1	K	VGLQVVAVK	A
480.7447	959.4748	2	959.5036	-0.0288	0	45.33	0.11	1	R	VTDALNATR	A
595.7855	1189.5565	2	1189.6012	-0.0447	0	56.55	0.0068	1	K	EIGNIISDAMK	K
603.7720	1205.5294	2	1205.5961	-0.0668	0	50.13	0.027	1	K	EIGNIISDAMK	K
608.3099	1214.6052	2	1214.6506	-0.0454	0	73.21	0.00015	1	K	NAGVEGSLIVEK	I
617.2857	1232.5569	2	1232.5884	-0.0315	0	80.63	2.7e-05	1	K	VGGTSDVEVNEK	K
672.8375	1343.6605	2	1343.7085	-0.0480	0	64.38	0.001	1	R	TVIIEQSWGSPK	V
714.8884	1427.7623	2	1427.8057	-0.0434	0	64.52	0.00086	1	R	GVMLAVDAVIAELK	K
714.8938	1427.7730	2	1427.8057	-0.0327	0	72.61	0.00013	1	R	GVMLAVDAVIAELK	K
722.8849	1443.7552	2	1443.8006	-0.0454	0	72.71	0.00014	1	R	GVMLAVDAVIAELK	K
722.8934	1443.7722	2	1443.8006	-0.0284	0	70.08	0.00025	1	R	GVMLAVDAVIAELK	K
752.8643	1503.7141	2	1503.7490	-0.0349	0	89.56	2.7e-06	1	K	TLNDELEIIEGMK	F
760.8461	1519.6777	2	1519.7439	-0.0662	0	84.43	8.9e-06	1	K	TLNDELEIIEGMK	F
640.3281	1917.9625	3	1918.0636	-0.1010	0	101.5	1.3e-07	1	K	ISSIQSIVPALEIANAHR	K
960.0327	1918.0509	2	1918.0636	-0.0127	0	87.34	3.2e-06	1	K	ISSIQSIVPALEIANAHR	K
1019.5106	2037.0067	2	2037.0153	-0.0086	0	52.42	0.01	1	R	IQEIIEQLDVTTSEYEK	E
1057.0537	2112.0929	2	2112.1322	-0.0393	0	115.78	4.6e-09	1	R	ALMLQGVDLLADAVAVTMGPK	G
1065.0399	2128.0653	2	2128.1271	-0.0618	0	68.73	0.00022	1	R	ALMLQGVDLLADAVAVTMGPK	G
1073.0477	2144.0809	2	2144.1220	-0.0411	0	69.64	0.00018	1	R	ALMLQGVDLLADAVAVTMGPK	G
789.1062	2364.2968	3	2364.3263	-0.0296	0	55.53	0.0038	1	R	KPLVIIAEDVDGEALSTLVLNR	L
1183.1570	2364.2994	2	2364.3263	-0.0269	0	65.46	0.00038	1	R	KPLVIIAEDVDGEALSTLVLNR	L
789.1094	2364.3063	3	2364.3263	-0.0200	0	94.59	4.5e-07	1	R	KPLVIIAEDVDGEALSTLVLNR	L

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



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, plus:

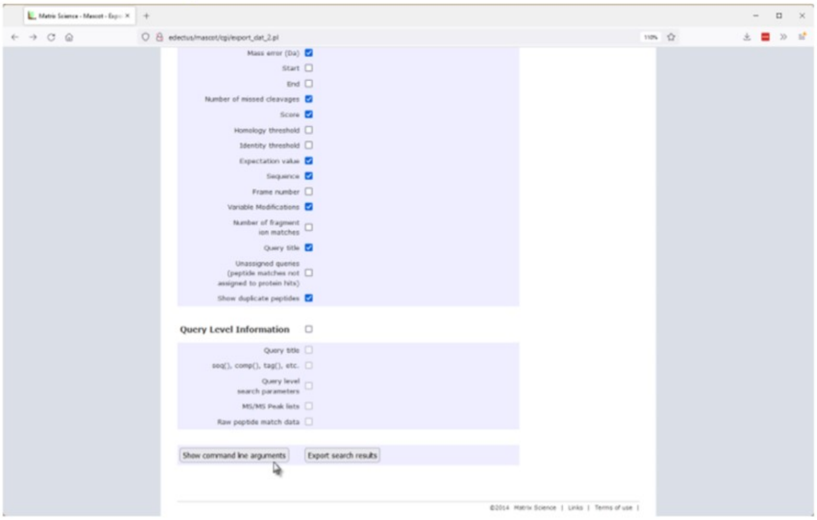
Type of search	HTML Report	Threshold type	Protein Scoring	Same-sets	Sub-sets	Group proteins
PMF	Concise Protein Summary	N/A	N/A	checked	1	N/A
MS/MS	Peptide Summary	Identity	As format controls	checked	As format controls	not checked
MS/MS	Protein Family Report	Homology	MudPIT	checked	1	checked

Sections

- Custom XML and CSV
- pepXML
- mzIdentML
- mzTab
- DTASelect
- Mascot DAT File
- MGF Peak List
- Optional Protein Hit Information
- Command Line Execution
- XML Schema

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

Search result export



Matrix Science - Mascot - Expo
edectus/mascot/cgi/export_dat_2.pl

Mass error (Da) ☑
Start □
End □
Number of missed cleavages ☑
Score ☑
Homology threshold □
Identity threshold □
Expectation value ☑
Sequence ☑
Frame number □
Variable Modifications ☑
Number of fragment ion matches □
Query title ☑
Unassigned queries (peptide matches not assigned to protein hits) □
Show duplicate peptides ☑

Query Level Information □

Query title □
seq(), comp(), tag(), etc. □
Query level search parameters □
MS/MS Peak lists □
Raw peptide match data □

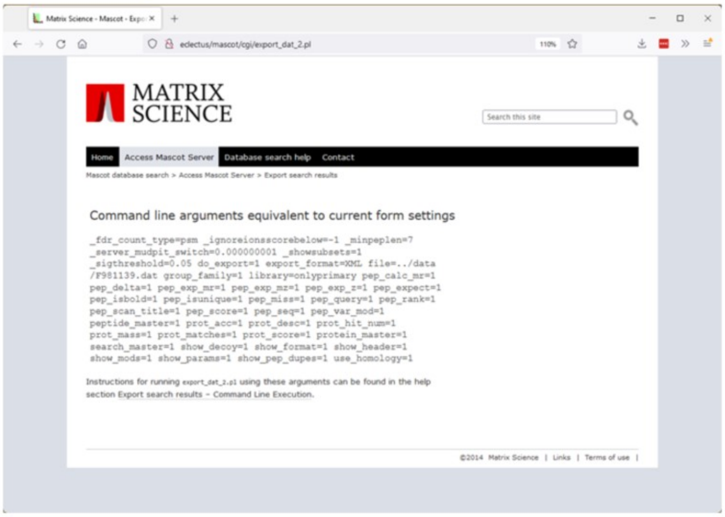
Show command line arguments
Export search results
©2014 Matrix Science | Links | Terms of use |

The help 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, there is a function to display the command line corresponding to the selected options.

Search result export



Matrix Science - Mascot - Expo ×
eclectus/mascot/cgi/export_dat_2.pl
110%
MATRIX SCIENCE
Search this site
Home
Access Mascot Server
Database search help
Contact
Mascot database search > Access Mascot Server > Export search results

Command line arguments equivalent to current form settings

```
_fdr_count_type=psm _ignoreionsscorebelow=-1 _minpeplen=7
_server_mudpit_switch=0.000000001 _showsubsets=1
_sigthreshold=0.05 do_export=1 export_format=XML file=../data
/F981139.dat group_family=1 library=onlyprimary pep_calc_mr=1
pep_delta=1 pep_exp_mr=1 pep_exp_mz=1 pep_exp_z=1 pep_expect=1
pep_isbold=1 pep_isunique=1 pep_miss=1 pep_query=1 pep_rank=1
pep_scan_title=1 pep_score=1 pep_seq=1 pep_var_mod=1
peptide_master=1 prot_acc=1 prot_desc=1 prot_hit_num=1
prot_mass=1 prot_matches=1 prot_score=1 protein_master=1
search_master=1 show_decoy=1 show_format=1 show_header=1
show_mods=1 show_params=1 show_pep_dupes=1 use_homology=1
```

Instructions for running export_dat_2.pl using these arguments can be found in the help section Export search results - Command Line Execution.

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