New

The
Protein Family
Result Report

Room for Improvement

Reporting large searches
- HTTP timeouts
- Browser hanging
- Out of memory

Protein Inference
- Need to use ions score cut-off and 'Require bold red' to create minimal list
- Proteins related by shared peptide matches could be widely separated in the report

MASCOT : *Protein Family Result Report* © 2010 Matrix Science

There were two areas that we really felt needed improvement in the result reports in Mascot 2.2 and earlier.

First, although you could run very large searches on any supported platform, even 32-bit and with limited amounts of memory, trying to display the result report would often lead to problems due to an HTTP timeout or the web browser becoming slow to respond or hanging completely or simply running out of memory on either the server or the client.

Second, protein inference was not smart enough, and left too much of the hard work to the user. To create a minimal list of proteins, you had to tweak the score cut-off and use 'Require bold red'

Of course, these two issues are related, because it is with very large search results that protein inference needs to be as automated as possible.

Protein Family Summary

Fixes both problems

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

MASCOT : *Protein Family Result Report*   *© 2010 Matrix Science*   *MATRIX SCIENCE*

To address both problems, we have developed a brand new report that loads most of the information 'on demand': the Protein Family Summary. This requires some index files to be created on the server, and these index files are cached, so that the report loads much faster on the second and subsequent occasions. It is a paged report; we don't try to display all the protein hits at once. And, detailed information is shown 'on demand'. 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.
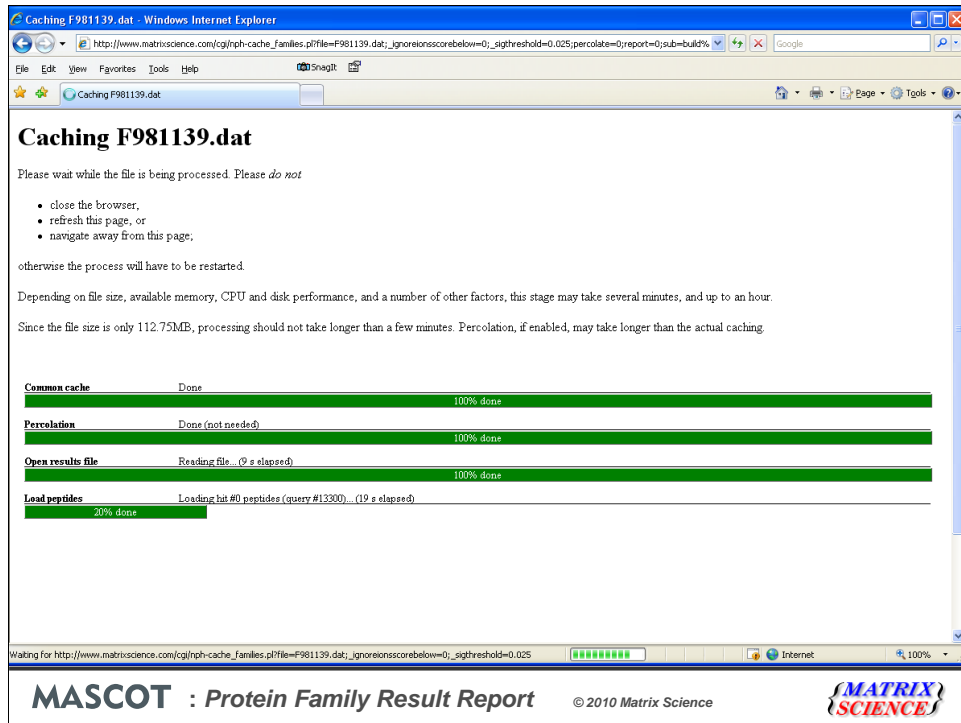
Lets begin by looking at the features that conserve memory and increase the speed of loading.

The Protein Family Summary is the default report whenever there are there are 300 or more spectra. This is controlled by a setting in mascot.dat, so you can easily change this threshold or even make the old report the default, if you prefer.

If you are running the search interactively, before the report loads for the first time, you'll see this progress page while the cache files are created. These cache files make loading much faster on subsequent occasions.

If you run searches using Mascot Daemon, the index files are created automatically and the search results should appear instantly when you click on the link in Daemon

MASCOT : *Protein Family Result Report*   © 2010 Matrix Science

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

Proteins are grouped into families using a novel hierarchical clustering algorithm. We'll return to this later. If a family contains a single member, the accession string, protein score and description are listed. If a family contains multiple members, the accessions, scores and descriptions are aligned with a dendrogram, which illustrates the degree of similarity between members. To see complete information about the proteins in the family and the peptide matches assigned to each family member, click on the family number link. We'll click on family 3
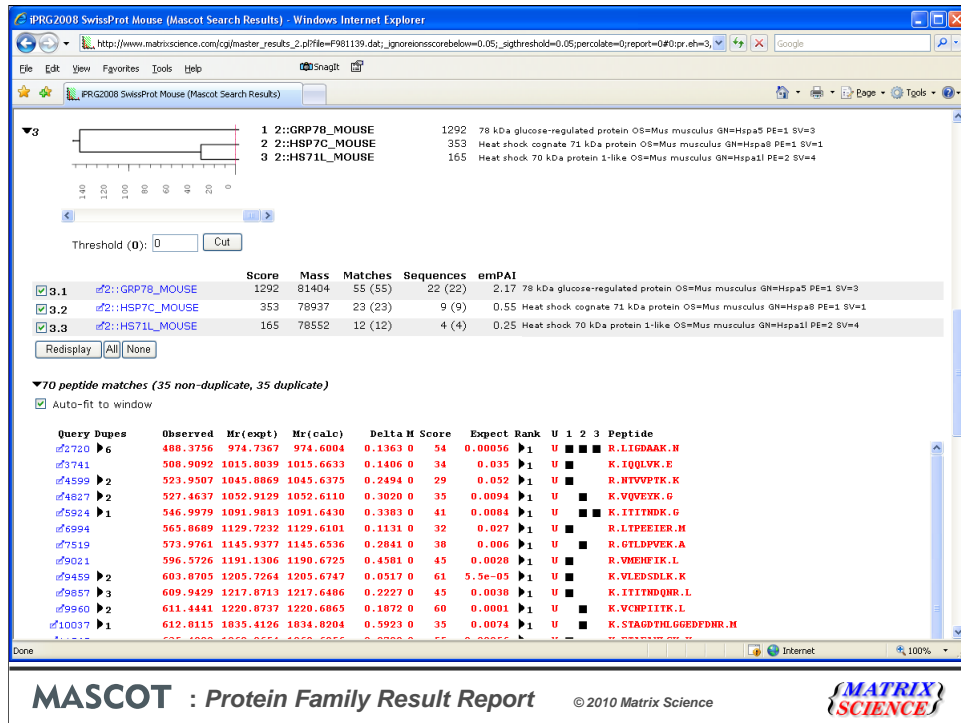
Immediately under the dendrogram is a list of the proteins. In this example, because SwissProt has low redundancy, each family member is a single protein. In other cases, a family member will represent multiple same-set proteins.

Below this is the table of peptide matches, which contains very similar information to that found in the other result reports. The black squares to the right show which peptides are found in which protein. To see only the peptides that distinguish HSP7C_MOUSE and HS71L_MOUSE, clear the checkbox for GRP78_MOUSE and choose Redisplay.

The number of columns and rows in the table is reduced to the matches to the two selected proteins.

Tooltips are used to display additional information. For example, if the mouse cursor rests over the rank column, the score thresholds are displayed.

iPRG2008 SwissProt Mouse (Mascot Search Results) - Windows Internet Explorer

http://www.matrixscience.com/cgi/master_results_2.pl?file=F981139.dat;_ignoreionsscorebelow=0.05;_sigthreshold=0.05;percolate=0;report=0#4:pr.eh=3,   Google

File   Edit   View   Favorites   Tools   Help   SnagIt

iPRG2008 SwissProt Mouse (Mascot Search Results)   Page ▾   Tools ▾

▼3   1 2::GRP78_MOUSE   1292   78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
                                            2 2::HSP7C_MOUSE   353   Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
                                            3 2::HS71L_MOUSE   165   Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2 SV=4

     140 120 100 80 60 40 20 0

     Threshold (0): 0   Cut

                          Score   Mass   Matches Sequences emPAI
☐ 3.1   2::GRP78_MOUSE   1292   81404   55 (55)   22 (22)   2.17 78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
☑ 3.2   2::HSP7C_MOUSE   353   78937   23 (23)   9 (9)   0.55 Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
☑ 3.3   2::HS71L_MOUSE   165   78552   12 (12)   4 (4)   0.25 Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2 SV=4
Redisplay   All   None

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

   Query Dupes   Observed   Mr(expt)   Mr(calc)   Delta M Score   Expect Rank U 2 3 Peptide
   2720 ▶6   488.3756   974.7367   974.6004   0.1363 0   54   0.00056 ▶1   U ■ ■ R.LIGDAAK.N
                                       Locus:19.275.3
                                       Score > 33 indicates identity
                                       Score > 27 indicates homology
   4827 ▶2   527.4637   1052.9129 1052.6110   0.3020 0   35   0.0094 ▼1   U ■   K.VQVEYK.G
                                                  -0.6643 0   12   1.6 2   MPVEYK
                                                   0.3020 0   10   3.1 3   VEVQYK
                                                   0.2907 1   9   3.2 4   ARDLYK
                                                   0.3527 0   8   5 5   VQICEK
                                                   0.3020 0   7   5.6 6   LNLDYK
                                                   0.3020 0   7   5.6 6   NIEVYK
                                                   0.3020 0   6   7.9 8   SVTAPYK
                                                   0.3570 0   5   9.3 9   MPPPAPGAR   + Oxidation (M)
                                                  -0.7036 1   4   10 10   SGGPRPALR

Done   Internet   100%

MASCOT : *Protein Family Result Report*   © 2010 Matrix Science   MATRIX SCIENCE

If you click on a marker in the rank column, it expands to show the 'yellow pop-up', containing the top 10 matches for this particular spectrum, except that this is no longer a pop-up. It is being displayed in-line so that you can print it, if you wish. You can also expand the top 10 for more than one match if you want to compare them

Duplicate matches to the same sequence are collapsed into a single row.

Usually, we are not very interested in lower scoring duplicate matches. But, if you want to see them, click on the marker in the duplicates column to expand. Obviously, clicking on either of these markers a second time causes the additional rows to collapse back to a single row

Scrolling back to the top of the report, we can see the header information and format controls. This is mostly similar to the existing reports except that sections can be expanded or hidden, as required. Note that, if you want the older reports, the Select Summary and the Protein Summary, these are still available by clicking on this link.
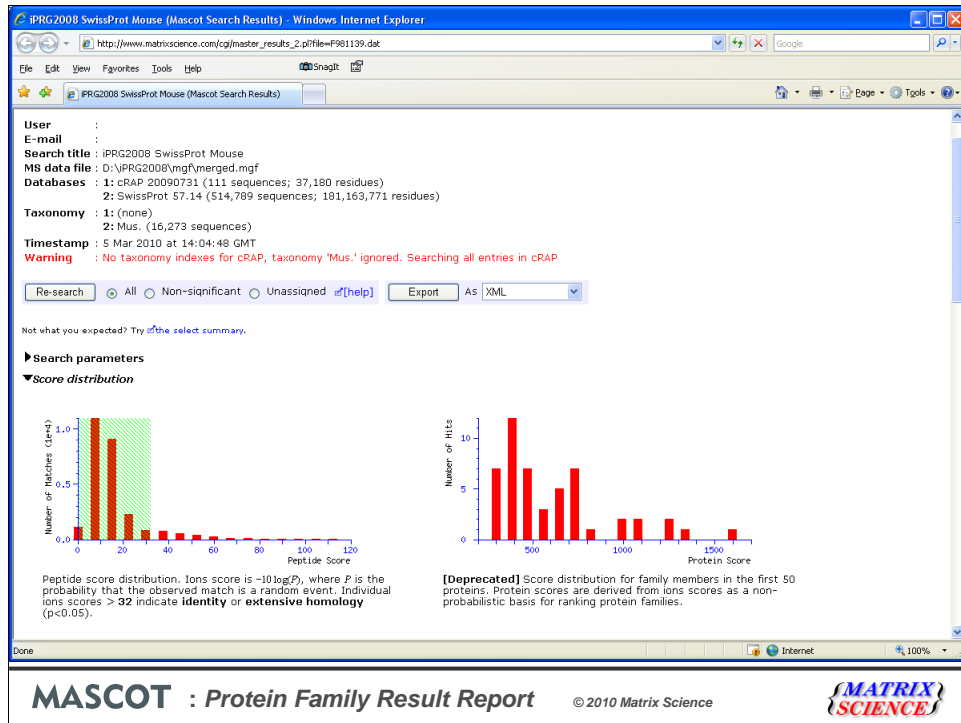
For example, we can expand to show two score distributions: one for the peptide match scores and one for the protein scores The old reports just have the protein score distribution, which isn't very meaningful, and is now marked deprecated (a software term meaning obsolete).

Lets take a quick look at the other two tabs. First, quantitation:

Protein Family Summary

Filter | Significance threshold p< 0.05 | Max. number of families AUTO | [help]
Ions score or expect cut-off 0

Quantitate | Protein ratio type Weighted | Normalisation None | [help]
Min. precursor charge 1 | Outlier removal Automatic
Min. # peptides 2 | Peptide threshold Above homology | 0.05

Display | Report peptide ratios ☑ Report protein ratios ☑

Proteins (29) | Quantitation (29) | § link to

Quantitation overview (29 proteins)

		Score	Mass	Matches	Sequences	emPAI	115/114	116/114	117/114	
1	VIT3_DROME	212	50436	19 (12)	13 (11)	1.13	1.748	2.241	3.202	Vitellogenin-3 precursor (Vitellogenin III) (Yol...
2	PDI_DROME	136	63614	16 (7)	14 (7)	0.46	1.883	3.203	1.712	Protein disulfide-isomerase precursor (EC 5.3...
3	UGGG_DROME	132	187242	11 (7)	11 (7)	0.14	1.017	1.184	0.680	UDP-glucose:glycoprotein glucosyltransferase ...
4	RL4_DROME	118	51286	9 (4)	6 (4)	0.31	0.965	1.219	2.152	60S ribosomal protein L4 (L1) - Drosophila me...
5	VIT1_DROME	115	52618	12 (6)	4 (3)	0.22	1.169	1.157	1.544	Vitellogenin-1 precursor (Vitellogenin I) (Yolk ...
6	YL_DROME	111	235777	7 (6)	6 (5)	0.08	1.088	1.829	1.178	Putative vitellogenin receptor precursor (Protei...
7	RL13_DROME	97	29639	5 (3)	4 (2)	0.26	1.569	3.054	4.981	60S ribosomal protein L13 (BBC1 protein hom...
8	VDAC_DROME	97	34278	11 (5)	6 (3)	0.35	1.039	2.001	3.410	Voltage-dependent anion-selective channel (P...
9	KPYK_DROME	87	63726	5 (3)	4 (2)	0.18	0.853	1.916	3.531	Pyruvate kinase (EC 2.7.1.40) (PK) - Drosophil...
10	ATPB_DROME	87	57821	8 (4)	7 (4)	0.27	0.359	0.828	1.724	ATP synthase subunit beta, mitochondrial prec...
11	VIT2_DROME	84	52892	4 (2)	2 (1)	0.07	1.349	0.829	1.470	Vitellogenin-2 precursor (Vitellogenin II) (Yolk...
12	RS15A_DROME	83	16629	2 (2)	2 (2)	0.50	0.833	1.009	2.162	40S ribosomal protein S15Aa - Drosophila mel...
13	PERC_DROME	81	98282	2 (2)	2 (2)	0.07	1.260	1.487	1.096	Chorion peroxidase precursor (EC 1.11.1.7) (P...
14	EF11_DROME	77	57566	10 (6)	8 (5)	0.35	0.843	1.653	3.216	Elongation factor 1-alpha (EF-1-alpha) (50 kD...
15	APLP_DROME	74	417818	28 (5)	26 (5)	0.04	0.911	1.432	1.799	Apolipophorins precursor (Retinoid- and fatty ...
16	GBLP_DROME	73	38892	5 (4)	5 (4)	0.43	1.049	2.164	2.490	Guanine nucleotide-binding protein subunit be...

MASCOT : *Protein Family Result Report* © 2010 Matrix Science | MATRIX SCIENCE

The quantitation tab is always populated for a large result report, even if you didn't choose a quantitation method, because emPAI values are calculated automatically. This tab is also useful if you want a simple, single level table of the protein hits. In Mascot 2.3, all MS/MS reports give counts of both the number of peptide matches and the number of distinct sequences. The numbers in parentheses are the counts above the specified significance threshold. If you wanted a table of proteins that conformed to the MCP guideline concerning one-hit wonders, you would exclude any protein that didn't have a count of at least two significant sequences, so VIT2_DROME would to be dropped

13

Back to the original report, this is the unassigned peptide match tab. Again, it is paged, with a default of 100 peptide matches per page. As you can see, we have over 30,000 unassigned matches in this particular search, so displaying them all in a single list would use a large amount of memory. Up at the top of both the proteins and unassigned tab, you'll find a text search control

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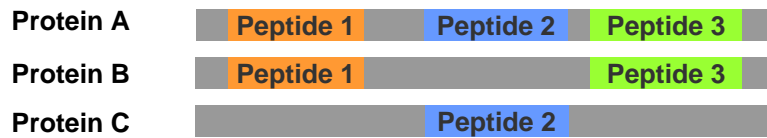
Here, for example, we searched for a peptide sequence. The display jumps to the first instance of the sequence, expands, and highlights (in green) the target peptides.

I hope you'll agree that this new report is a step in the right direction. Using cache files and displaying information 'on demand' allows even the largest search result to be displayed without memory or time-out problems

Protein Inference

General approach is to create a minimal list of proteins.

"Principal of parsimony" or "Occam's razor"

Protein A	Peptide 1 · Peptide 2 · Peptide 3	
Protein B	Peptide 1 · · Peptide 3	
Protein C	· Peptide 2 ·	

MASCOT *: Protein Family Result Report*   © 2010 Matrix Science

---

Now let's turn to the other aspect of the new report: improved protein inference.

As we all know, database matching of MS/MS spectra identifies peptides. Using these peptide sequences to deduce which proteins were present in the original sample is surprisingly difficult because many of the peptide sequences in a typical search result can be assigned to more than one protein.

The goal of protein inference is to create a minimal list of proteins. That is, the minimum number of proteins that can account for the observed peptides

Some people call this approach the principal of parsimony, others call it Occam's razor.

Imagine the very simple case we have three peptide matches, which can be assigned to three proteins, as illustrated here. Do we have evidence for all three proteins, or just one?

Mascot will report that the sample contained protein A. Proteins B and C are defined as sub-set proteins, and given an inferior status.

This is certainly a reasonable decision, but there is no guarantee that it is correct. It is possible that the sample actually did contain a mixture of proteins B and C, but not protein A. Another thing to watch for is the possibility that peptide 2 is a very weak match, maybe a random match. If so, then there is nothing to choose between Proteins A and B.

This ambiguity is made worse in a shotgun proteomics or MudPIT experiment, where the proteins from a cell lysate are digested to peptides without any prior fractionation or separation.

Protein Inference

Red indicates the top scoring peptide match for this spectrum

Bold indicates that this is the highest scoring protein in which this peptide is found

Look for protein hits with at least one bold red match

MASCOT : *Protein Family Result Report* © 2010 Matrix Science

In the Peptide Summary and Select Summary reports, the interpretation of the results from a large LC-MS/MS search can be tricky, because it is not always clear which proteins are "real" and which are just alternative groupings of peptide matches that are better assigned to other proteins.

We use red and bold typeface to highlight the most logical assignment of peptides to proteins. Whenever the top scoring peptide match for an MS/MS spectrum appears, it is shown in red. Note that this does not mean that the match is necessarily significant. It is just the best match. A peptide match is shown in bold face when this is the highest scoring protein in which a match to this spectrum has been found. This means that peptide matches that are both bold and red are the most likely assignments of the best matches. Conversely, if a protein hit doesn't have any bold red matches, either the assigned spectra have better scoring matches to other sequences or matches to the same spectra are assigned to higher scoring proteins elsewhere.

Usually, a protein hit without any bold red matches would collapse into a higher scoring protein hit were it not for the presence of one or more weak, random matches. Such hits can be filtered out of the report by ticking the 'require bold red' checkbox.

It is also a good idea to set a score cut-off, because this will make eliminate the weak, random matches that prevent proteins collapsing into a minimum number of hits. If the number entered into the Ions score cut-off is a number between 0 and 1 it is treated as a cut-off on the expect value. So, entering 0.05 will remove all non-significant matches.

Protein Inference

20: Q3UWB9	34: Q91WH2	51: Q3UEP4
GAAVALNIR		
	GAAVTLNIR	GAAVTLNIR
	AEMWLIR	AEMWLIR
IILDELVQR	IILDELVQR	
FSPGYQIEK	FSPGYQIEK	
DNLENFFIK	DNLENFFIK	
FETFPTSVSK	FETFPTSVSK	
FVDVWTYEMPR	FVDVWTYEMPR	
		WTYEVPR
		IILDELK
		TPATLGPNTR
		FSPGYYLEK
WLPQNDLLGHPK	WLPQNDLLGHPK	WLPQNDLLGHPK
ANAIAWALAQIPQK		ANAIAWALAQIPQK
GHEVTVLRPSAYYVLDPK	GHEVTVLRPSAYYVLDPK	

MASCOT : *Protein Family Result Report*   *© 2010 Matrix Science*   MATRIX SCIENCE

Using bold red and a score cut-off to filter the protein list works reasonably well, but it is far from ideal. In particular, the Peptide Summary and Select Summary reports fail to remove intersection proteins. A good example is shown here. Three proteins appear in the report at positions 20, 34, and 51. All have some bold red matches. And yet, when placed side by side, we can see that all the matches in 34 are found in 20 and 51. In other words, there is no independent evidence for protein 34. It's a sub-set of the combined matches to proteins 20 and 51.

A second shortcoming is that related proteins, that we might want to consider side by side, become separated in the report because it is sorted by protein score. This is the result report we were looking at earlier, but now displayed as a select summary. Hit 2 is one of the highest scoring proteins: a Cytochrome P450 from mouse

http://www.matrixscience.com/cgi/master_results.pl?file=..%2Fdata%2FF981139.dat&_querylist=all&REPTYPE=select&_sigthreshold=0.05&REPORT=AUTO& ✓ ✗ ✗ Google

File Edit View Favorites Tools Help   SnagIt

Select Summary Report (iPRG2008 SwissProt Mouse)   Page ▾ Tools ▾

Query		Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
17987		737.4600	2209.3582	2210.1817	-0.8235	1	3	67	5	U	M.RYVASYLLAALGGHSSPSAK.D
27962		1021.6837	3062.0292	3061.6287	0.4005	0	97	1.9e-08	1	U	K.LASVPAGGAVAVSAAPGSAAPAAGSAPAAAEEK.K 32832

28. 2::CP254_MOUSE   Mass: 60887   Score: 535   Matches: 76(25)   Sequences: 26(8)   emPAI: 0.87
Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
1373	454.3348	906.6551	906.5200	0.1350	0	26	0.33	1		R.NLGMGK.R
1663	462.4237	922.8329	922.5150	0.3180	0	(8)	21	7		R.NLGMGK.R 1644 1646
1953	470.2918	938.5691	938.3894	0.1797	0	11	4.8	1	U	R.PCMQDR.S
3499	503.8391	1005.6637	1005.6103	0.0535	0	35	0.031	1		R.LPVFDK.A 3509
4731	526.2961	1050.5776	1050.5323	0.0453	0	35	0.021	1		R.CLVEELR.K 4728 4741 4748 4758
4766	526.4296	1050.8446	1050.6317	0.2129	0	31	0.087	1		K.NFLLEK.I 4752 4773 4775
5303	535.4618	1068.9091	1068.6211	0.2880	0	31	0.082	1		R.NYFIPK.G 5279 5294 5304
5544	540.3247	1078.6349	1078.5385	0.0964	0	54	0.00028	1		R.ICAGEGLAR.M 5530 5539 5540 5542 5548 5549 5550 5566 5573
6164	551.9603	1101.9061	1102.6379	-0.7318	0	5	38	2		K.YPHVTAK.V
7014	566.3561	1130.6977	1130.5964	0.1013	1	4	26	5		R.FDYKDR.D
7576	574.8879	1147.7612	1147.6117	0.1495	0	26	0.43	1		K.EFPNPEK.F 7534 7538 7568 7573 7578
9880	610.3896	1218.7646	1218.7328	0.0318	1	(8)	25	3		R.GRLPVFDK.A 9900
9906	407.3264	1218.9573	1218.7328	0.2245	1	12	8.3	1		R.GRLPVFDK.A
10395	618.0392	1234.0638	1233.8304	0.2335	0	43	0.0053	1		R.YALLLLLK.Y
17673	731.0732	1460.1318	1460.8159	-0.6841	0	17	2.6	2		R.DFIDYFLIK.G
2918	492.4046	1474.1920	1474.7497	-0.5577	1	(7)	41	10		R.SIEDRVQEEAR.C
18043	738.5613	1475.1080	1474.7497	0.3582	1	8	14	2		R.SIEDRVQEEAR.C
18114	739.5340	1477.0534	1476.8634	0.1900	0	82	6.3e-07	1	U	K.GTTVITSLSSVLR.D 18115 18118 18122 18125 18136 18139 18142 1814
3609	507.4343	1519.2810	1518.8211	0.4599	1	1	1.4e+02	9	U	K.NTRHFSLMTLR.N
19465	510.6492	1528.9257	1528.7756	0.1500	0	(59)	9.2e-05	1	U	K.EALVDHGDVFAGR.G 19486 19488
19473	765.5187	1529.0228	1528.7756	0.2472	0	93	4.4e-08	1	U	K.EALVDHGDVFAGR.G 19468
19851	773.1436	1544.2725	1543.8521	0.4205	0	61	6.6e-05	1		R.VYGPVYTLYLGR.K 19848 19853
20447	784.5878	1567.1610	1566.7632	0.3978	0	10	8.4	3		K.SDYFHPFSTGK.R
20458	523.6711	1567.9915	1567.8813	0.1102	0	27	0.15	1		K.VQEEIEHVIGK.H
20822	528.3086	1581.9040	1582.7581	-0.8541	0	(3)	27	3		K.SDYFHPFSTGK.R
20849	792.3833	1582.7520	1582.7581	-0.0061	0	(7)	20	4		K.SDYFHPFSTGK.R 20877
23961	580.8898	1739.6475	1738.8592	0.7882	1	4	27	7		K.SDYFHPFSTGKR.I
23988	581.4760	1741.4063	1741.8982	-0.4919	1	6	18	2	U	K.DSYFHPFSTGKR.I 22002

Internet   🔍 100%

MASCOT : *Protein Family Result Report*   © *2010 Matrix Science*   {MATRIX}{SCIENCE}

If you were to browse down, you'd find another Cytochrome P450 at hit 28

And another at 44

And at 69

http://www.matrixscience.com/cgi/master_results.pl?file=..%2Fdata%2FF981139.dat&_querylist=all&REPTYPE=select&_sigthreshold=0.05&REPORT=AUTO& | Google

File   Edit   View   Favorites   Tools   Help   SnagIt

Select Summary Report (iPRG2008 SwissProt Mouse)   Page ▾ Tools ▾

| 17179 | 482.6943 | 1445.0610 | 1445.7708 | -0.7098 | 1 | (4) | 38 | 10 | U | R.SLQSVAEERAGR.H |
| 26956 | 650.4782 | 1948.4126 | 1947.9822 | 0.4304 | 0 | 36 | 0.018 | 1 | U | R.VLNSYWVGEDSTYK.F |

115.   2::CP238_MOUSE   **Mass:** 61216   **Score:** 202   **Matches:** 75(18)   **Sequences:** 19(4)   **emPAI:** 0.40
Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=1

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
1373	454.3348	906.6551	906.5200	0.1350	0	26	0.33	1		R.NLGMGK.R
1663	462.4237	922.8329	922.5150	0.3180	0	(8)	21	7		R.NLGMGK.R 1644 1646
3438	502.8658	1003.7170	1003.5906	0.1264	0	28	0.2	1		K.LNENVK.I 3441 3443 3445 3446 3457
4047	513.8493	1025.6841	1025.4214	0.2627	0	24	0.29	1		R.SPCMQDR.S 4051 4056 4057
4431	520.8626	1039.7106	1039.6157	0.0949	0	40	0.0079	1		R.SYLLEK.V 4409 4412 4420 4421 4423 4425 4429 4433 4439 4441 4442
4467	521.3901	1040.7657	1041.4164	-0.6506	0	(1)	62	10		R.SPCMQDR.S
5605	541.3848	1080.7551	1080.6059	0.1492	0	53	0.00051	1		K.YPDVTAK.V 5594 5603 5604 5608 5614 5615 5616
8679	590.5231	1179.0315	1178.6209	0.4107	0	5	37	2	U	R.GNIPMSEK.I
11467	634.4501	1266.8855	1267.7817	-0.8962	0	9	21	1		R.YALLLLMK.Y 11517
16093	706.4691	1410.9236	1410.8811	0.0425	1	2	54	8	U	R.SYLLEKVK.E
16519	713.0426	1424.0705	1423.7541	0.3164	0	8	15	4	U	K.VQEEIDHVVGR.H 16490
17673	731.0732	1460.1318	1460.8159	-0.6841	0	17	2.6	2		R.DFIDYFLIK.Q
17976	491.4253	1471.2541	1471.8642	-0.6101	1	11	6.4	5		K.YIRSYLLEK.V
18913	755.1673	1508.3200	1507.7132	0.6068	0	2	92	5	U	R.EEAQCLVEELR.K
19737	771.0876	1540.1606	1540.7476	-0.5869	0	7	18	3	U	K.SDYFMTFSAGK.R
20190	520.1012	1557.2819	1556.7425	0.5394	0	(1)	65	5	U	K.SDYFMTFSAGK.R
21325	802.5887	1603.1628	1602.7760	0.3868	0	48	0.0012	2		K.EALIDHGEEFSGR.G 21316
21547	538.4897	1612.4474	1611.9592	0.4882	0	29	0.11	1	U	R.PIVVLHGYEAVK.E
7037	566.5113	1696.5121	1696.8487	-0.3366	1	3	1e+02	6	U	K.SDYFMTFSAGKR.V
26729	967.1937	1932.3729	1932.0772	0.2957	0	(11)	5.5	2		K.GTTVVTSLTSVLHDSK.E
26736	645.1888	1932.5446	1932.0772	0.4674	0	31	0.054	2		K.GTTVVTSLTSVLHDSK.E 26724 26732 26735 12189 26737 26738 26740 2
12778	653.4007	1957.1801	1957.9260	-0.7459	0	(9)	15	2		R.SHMPYTDAMIHEVQR.F
27266	659.1885	1974.5436	1973.9209	0.6226	0	30	0.072	1		R.SHMPYTDAMIHEVQR.F
27449	664.4707	1990.3904	1989.9159	0.4745	0	(4)	31	8		R.SHMPYTDAMIHEVQR.F
31882	874.3484	2620.0233	2619.2842	0.7391	1	1	40	9	U	K.EALIDHGEEFSGRGNIPMSEK.I

116.   2::PI3_MOUSE   **Mass:** 54530   **Score:** 198   **Matches:** 34(13)   **Sequences:** 16(5)   **emPAI:** 0.37

Internet   100%

MASCOT : *Protein Family Result Report*   © *2010 Matrix Science*   **MATRIX SCIENCE**

And 115

And 254. Apart from the descriptions all containing the words Cytochrome P450, did you spot any other relationships? If you did, you're very smart and have a photographic memory. These 6 proteins are linked by a complex web of shared peptide matches. But, spotting this in the current reports is far from easy. This report, which is not particularly large, runs to 691 proteins. Think how many possible families there are in such a set of proteins.

Protein Inference: Family Report

1. From the search results, create an initial list of proteins, ordered by protein score
2. Take the highest scoring protein on the list
3. Find all other proteins in the same family:
   - select all peptide matches with homology score or better
   - for each peptide match, select all the proteins that contain this match and remove from the initial list
   - for each new protein, select all new peptide matches with homology score or better
   - loop until no further proteins or peptide matches remain
4. For each protein in the family, make a list of the distinct peptide sequences. That is, ignoring differences in modification state and precursor charge. Where there are duplicate matches to a sequence, the representative score for the sequence is the highest one
5. Using this set of distinct peptide sequences, divide and group the proteins into same-set proteins and sub-set proteins, which includes intersections.
   - Same-set proteins are collapsed into a single family member
   - Proteins that are sub-sets, including intersections, are relegated to secondary status
   - Perform hierarchical clustering of the family members
6. Loop from step 2 until no more peptide matches remain with homology score or better

MASCOT : *Protein Family Result Report*   *© 2010 Matrix Science*   **MATRIX SCIENCE**

The grouping algorithm in the new, protein family report clusters proteins into families on the basis of shared peptide matches. Only matches with scores above the homology threshold are used, non significant matches are ignored. An iterative method is used to remove intersection proteins as well as sub-set proteins.

iPRG2008 SwissProt Mouse (Mascot Search Results) - Windows Internet Explorer

http://www.matrixscience.com/cgi/master_results_2.pl?file=F981139.dat;_ignoreionsscorebelow=0.05;_sigthreshold=0.05;percolate=0;report=0#10:pr.ed= | Google

File Edit View Favorites Tools Help   SnagIt

iPRG2008 SwissProt Mouse (Mascot Search Results)   Page ▾ Tools ▾

| Proteins (482) | Quantitation (514) | Unassigned (30250) | § link to

Protein families 1–10 (out of 482)

10 ▾ per page   1 2 3 4 5 6 ... 49 Next   Expand all   Collapse all

Accession ▾   Find ☐ Match case

▶ 1   1::TRY1_BOVIN   1597 TRY1_BOVIN

▼ 2   1 2::CP2CT_MOUSE   1307 Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=1
                         4 2::CP239_MOUSE   293 Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=1
                         5 2::CP238_MOUSE   202 Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=1
                         2 2::CP254_MOUSE   535 Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
                         6 2::CP270_MOUSE   69 Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2
                         3 2::CY250_MOUSE   382 Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=1

         140 120 100 80 60 40 20 0

Threshold (D): 0   Cut

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

▶ 119 peptide matches (35 non-duplicate, 84 duplicate)

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

Done   Internet   100%

MASCOT : *Protein Family Result Report*   © 2010 Matrix Science   *MATRIX SCIENCE*

This is how the same 6 cytochrome P450 proteins are presented in the new report. Because they have shared, significant matches, they are grouped into a family.

The dendrogram illustrates the degree of similarity between the members. The scale on the dendrogram is ions score, and CP239_MOUSE and CP238_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. Where there are multiple matches to the same peptide sequence, it is the highest score for each sequence that is used to calculate the distance.

In the peptides table, we can see the complex relationship between the proteins. Each has at least one significant match to a unique sequence. Otherwise, it would become a sub-set of another family member and be relegated to the sub-sets list. If we want to see the relationship between any particular pair of proteins, we simply uncheck the boxes for the others and choose re-display

Please note that the use of bold face has changed in the new report. It now means that a match is significant. Not just that it's the first time we have seen a match to that particular spectrum. The meaning of red has not changed, it still indicates a top-ranked match

MASCOT : *Protein Family Result Report*   © 2010 Matrix Science
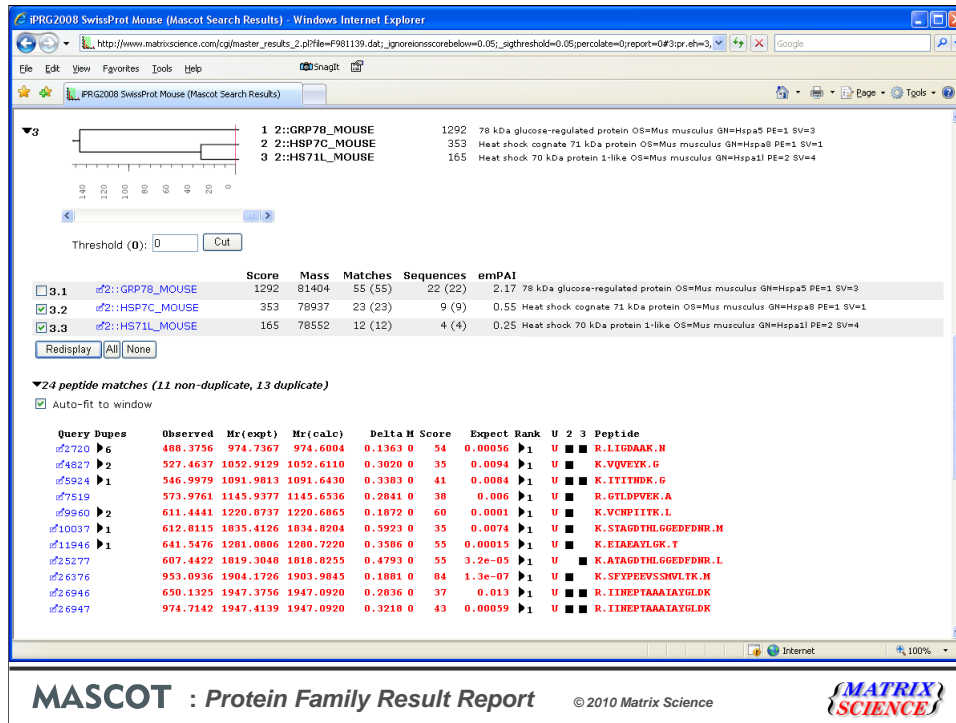
Here, for example, are a pair of proteins from family 3. You can immediately see that HS71L_MOUSE would be a sub-set of HSP7C_MOUSE if it was not for one match, K.ATAGDTHLGGEDFDNR.L. It is the significant score for this match that separates the two proteins in the dendrogram by a distance of 32 (score of 55 - homology threshold score of 23).

If you look a little more closely, you will notice that HSP7C_MOUSE has a weak match to K.STAGDTHLGGEDFDNR.M. So, the evidence for both proteins being present comes down to a single residue. If S221 in HSP7C_MOUSE was an A, or A223 in HS71L_MOUSE was an S, then HS71L_MOUSE would be a sub-set protein. Interestingly, the stronger match is for the sequence found in the protein with fewer matches. This could be chance or it could be that the analyte sequence was essentially HSP7C_MOUSE but with an A at this position.

If you decide that this is not sufficient evidence for reporting both proteins, you can simplify the report by "cutting" the dendrogram using the slider control.

iPRG2008 SwissProt Mouse (Mascot Search Results) - Windows Internet Explorer

http://www.matrixscience.com/cgi/master_results_2.pl?file=F981139.dat;_ignoreionsscorebelow=0.05;_sigthreshold=0.05;percolate=0;report=0#46:pr.cuts▼

File Edit View Favorites Tools Help

iPRG2008 SwissProt Mouse (Mascot Search Results)

Page ▼   Tools ▼

▼3   1 2::GRP78_MOUSE   1292   78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
      2 2::HSP7C_MOUSE   353   Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1

140 120 100 80 60 40 20 0

Threshold (50): 50   Cut

		Score	Mass	Matches	Sequences	emPAI	
☑ 3.1	2::GRP78_MOUSE	1292	81404	55 (55)	22 (22)	2.17	78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
☑ 3.2	2::HSP7C_MOUSE	353	78937	23 (23)	9 (9)	0.55	Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1

Redisplay All None

▼69 peptide matches (34 non-duplicate, 35 duplicate)
☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta M	Score	Expect	Rank	U	1	2	Peptide
2720	▶6	488.3756	974.7367	974.6004	0.1363	0	54	0.00056	▶1	U	■	■	R.LIGDAAK.N
3741		508.9092	1015.8039	1015.6633	0.1406	0	34	0.035	▶1	U	■		K.IQQLVK.E
4599	▶2	523.9507	1045.8869	1045.6375	0.2494	0	29	0.052	▶1	U	■		R.NTVVPTK.K
4827	▶2	527.4637	1052.9129	1052.6110	0.3020	0	35	0.0094	▶1	U		■	K.VQVEYK.G
5924	▶1	546.9979	1091.9813	1091.6430	0.3383	0	41	0.0084	▶1	U	■		K.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	0.006	▶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	0.0074	▶1	U		■	K.STAGDTHLGGEDFDNR.M
11545		635.4900	1268.9654	1268.6856	0.2799	0	55	0.00056	▶1	U	■		K.ETAEAYLGK.K
11946	▶1	641.5476	1281.0806	1280.7220	0.3586	0	55	0.00015	▶1	U		■	K.EIAEAYLGK.T

Done   Internet   100%

MASCOT : *Protein Family Result Report*   © 2010 Matrix Science   MATRIX SCIENCE

By cutting the dendrogram at a score of 50, HS71L_MOUSE becomes a sub-set of HSP7C_MOUSE, and we reduce the family to two members, each of which has a substantial number of unique sequences as evidence. These are different proteins that have sufficient homology to give some shared peptide matches. The strength of the new report is that it gives us a clear and simple overview, which allows us to browse for proteins of interest, as well as a tool for drilling down to the level of detail necessary to make decisions about whether there is evidence for any individual protein.

Protein Family Summary

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

MASCOT : *Protein Family Result Report*   *© 2010 Matrix Science*   *MATRIX SCIENCE*

To summarise