New Features in Mascot Server 2.4

MASCOT

MATRIX SCIENCE

Mascot Server 2.4

Improvements in reports

- Report Builder
- Preferred taxonomy
- FDR button
- Site analysis (delta score)

Database Manager

Other improvements

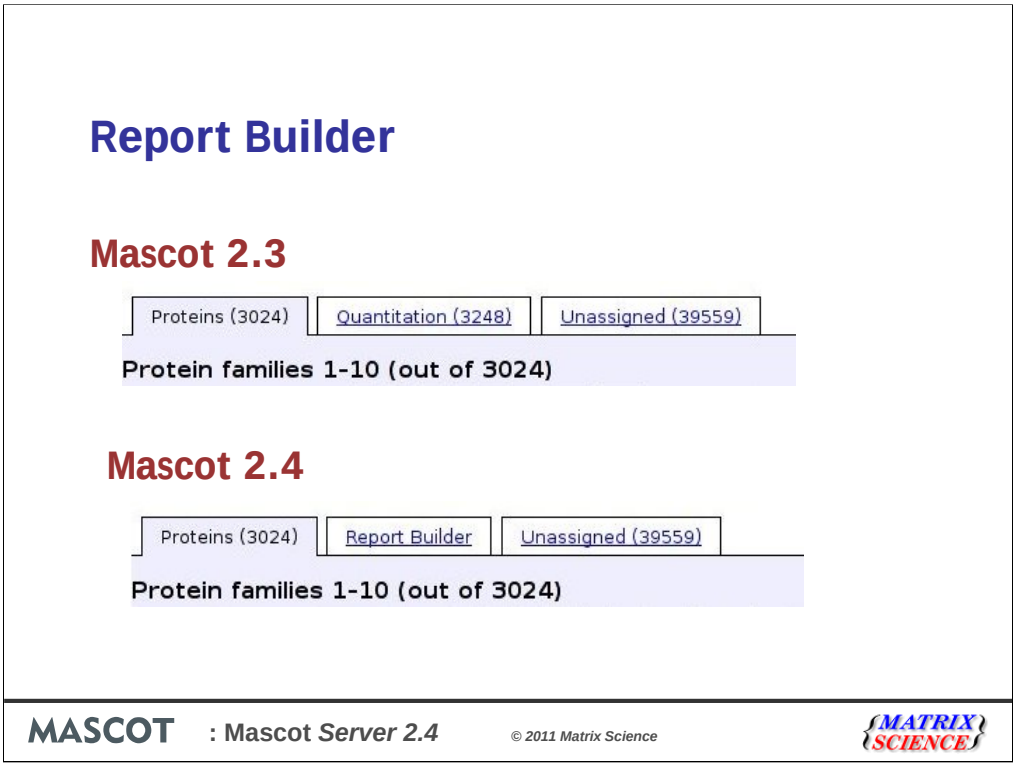
- Quantitation of modified peptides

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

I'd like to describe some of the new features in Mascot 2.4. Many of these relate to reports, such as 'Report Builder', that allows you to build a customised table of proteins and export it straight to Excel. More on this and the other new report features shortly.

One of the big stumbling blocks in Mascot administration has been automatic database updates. The functionality to automatically keep FASTA files up to date has always been there, but configuring it has required editing Perl files and having a good understanding of the operating system. In Mascot 2.4, we've replaced the old database configuration script (db_gui.pl) and the database update script (db_update.pl) with a single, browser-based Database Manager. We'll see it in action in the latter part of the talk.

Finally, if I have time, I'll mention some other improvements and bug fixes.



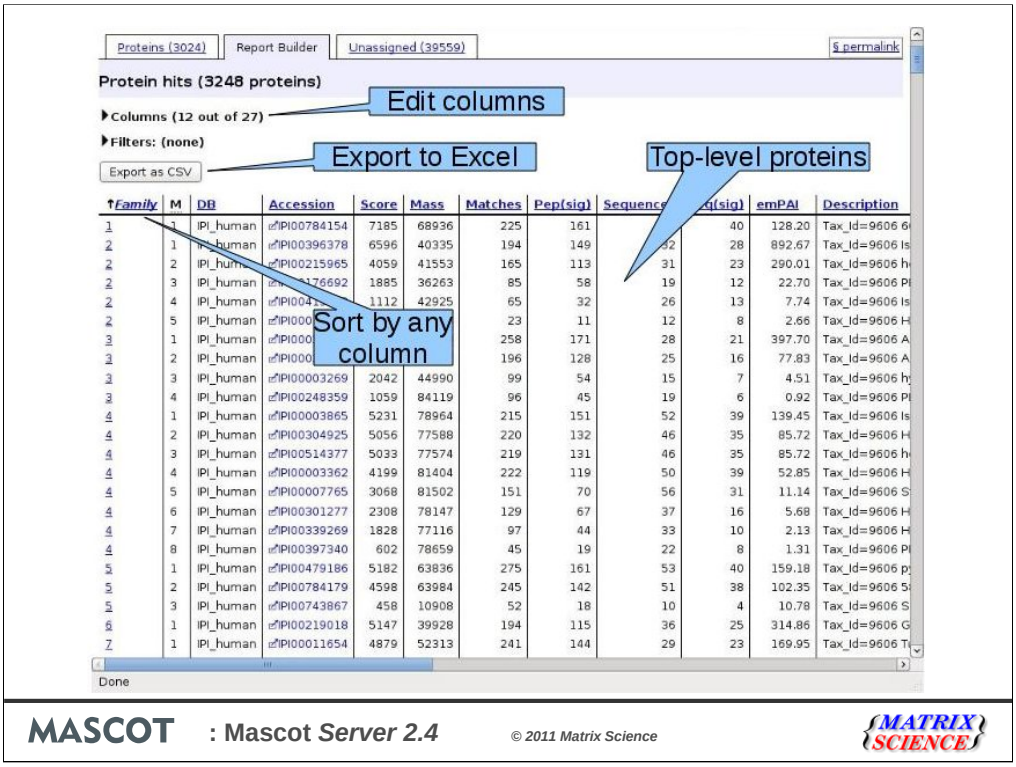
In Mascot 2.3, reports for large searches default to the Protein Family Summary. This has a Quantitation tab, which contains a simple table of all the top-level protein hits, and provides an overview of the search result, particularly when the main interest is protein level quantitation. In Mascot 2.4, this tab has been renamed to Report Builder to reflect the fact that it now includes a great deal of additional functionality.

Report Builder

- Show all top-level proteins in a table
- Add, remove and re-order columns
- Sort by any column
- Rule based filters
- Export as CSV (for Excel).

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

Report Builder does what it says: allows you to build a customised table of protein hits. You can choose which columns to include, and their order, filter out proteins that are of no interest, such as one-hit wonders, and export the table in CSV format directly to Excel. I'll now show you some of the basic functionality using a result report for an iTRAQ experiment as an example.

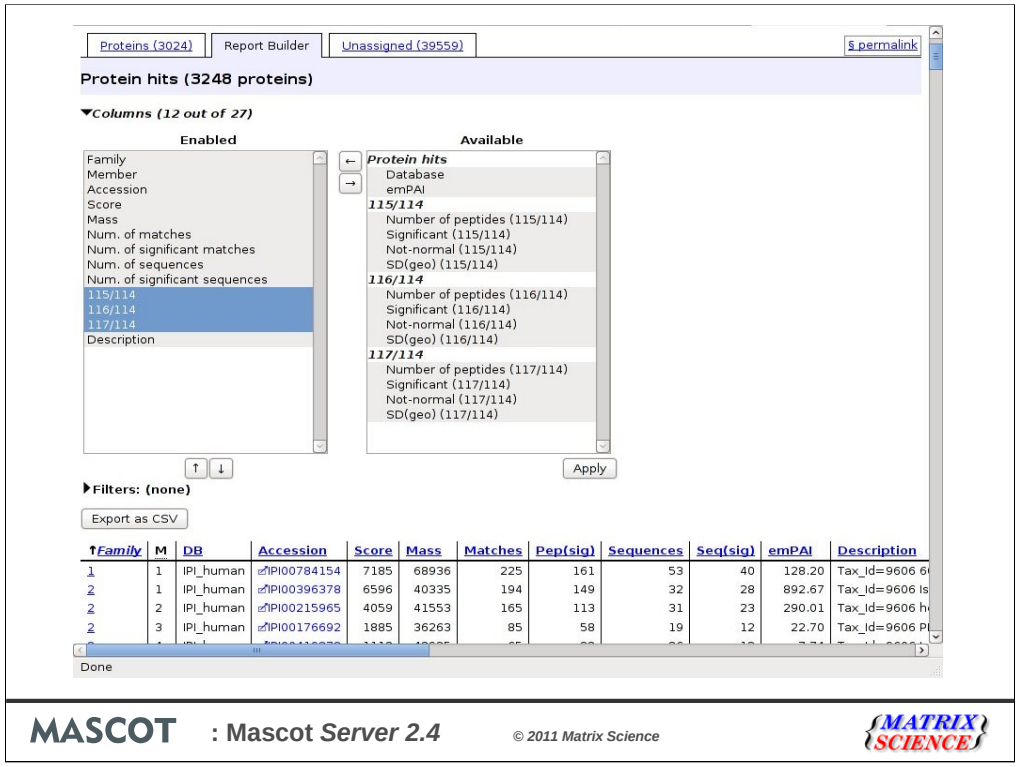


The table has one row for each of the top-level protein hits, sometimes called anchor proteins. Same-set or subset proteins do not appear. If you need to see a subset hits, you just go to the Proteins tab. Since a family can contain several protein hits, there will be as many rows per family as there are top-level proteins in the family.

If your search gives a lot of same-set proteins, you can select which same-set protein to show in the list by choosing a preferred taxonomy. We'll come back to that later in the talk.

The table can be exported as CSV with one click.

You can sort the table by clicking on a column header. The currently active sort order is shown by an arrow in the relevant column; up means ascending, down means descending. The sort order is also preserved when you export as CSV.

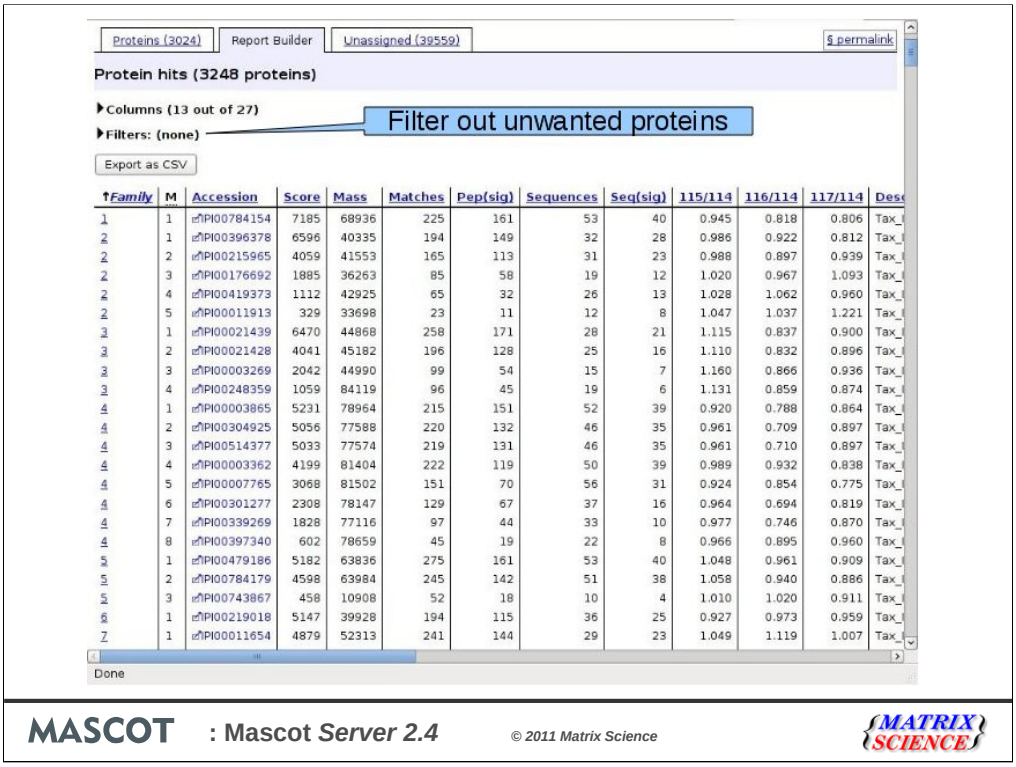


You can choose which columns to include and exclude by moving them between these two lists. The columns are categorised into groups. The basic set of columns that are always available are under "Protein hits". Each quantitation method gets its own set of columns, so these groups depend on which method you are using, if any.

Here we've added the quantitation ratio columns and removed database and emPAI to make some room on the slides. This is a single-database search, so the database column has the same value in each row -- not very interesting.

You can re-order the columns by using the up and down buttons here.

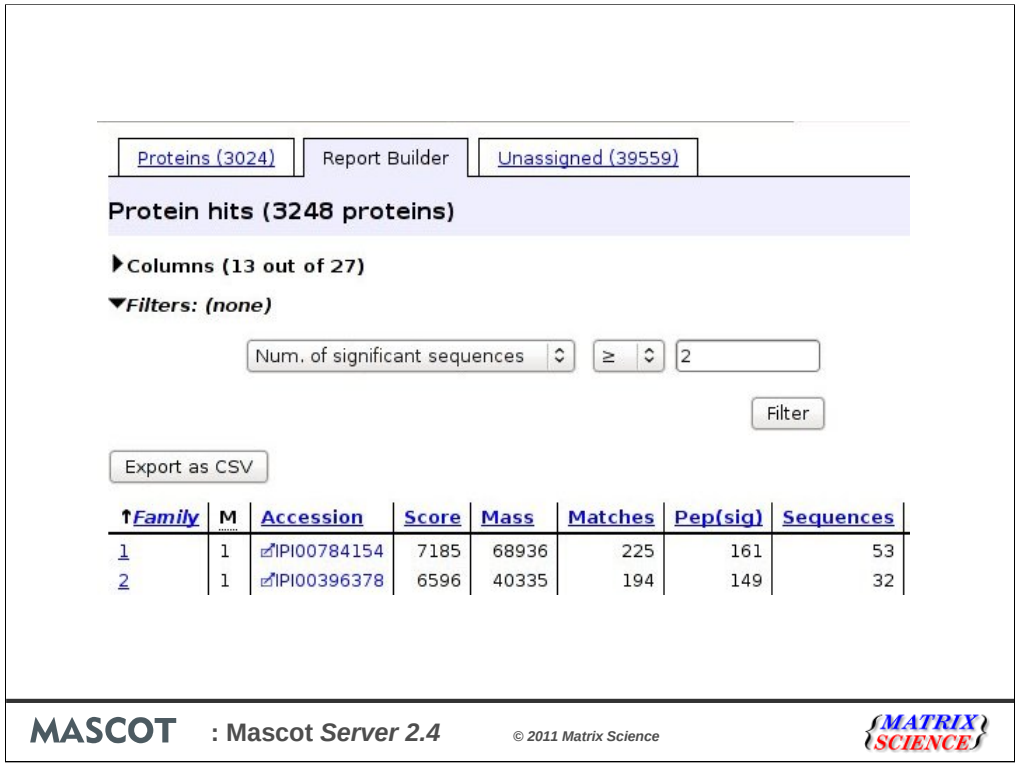
The changes take effect when you click Apply. The table will be reloaded with the new columns. The columns you've chosen to view and their order will be the same if you export as CSV.



↑Family	M	Accession	Score	Mass	Matches	Pep(sig)	Sequences	Seq(sig)	115/114	116/114	117/114	Desc
1	1	IPI00784154	7185	68936	225	161	53	40	0.945	0.818	0.806	Tax_
2	1	IPI00396378	6596	40335	194	149	32	28	0.986	0.922	0.812	Tax_
2	2	IPI00215965	4059	41553	165	113	31	23	0.988	0.897	0.939	Tax_
2	3	IPI00176692	1885	36263	85	58	19	12	1.020	0.967	1.093	Tax_
2	4	IPI00419373	1112	42925	65	32	26	13	1.028	1.062	0.960	Tax_
2	5	IPI00011913	329	33698	23	11	12	8	1.047	1.037	1.221	Tax_
3	1	IPI00021439	6470	44868	258	171	28	21	1.115	0.837	0.900	Tax_
3	2	IPI00021428	4041	45182	196	128	25	16	1.110	0.832	0.896	Tax_
3	3	IPI00003269	2042	44990	99	54	15	7	1.160	0.866	0.936	Tax_
3	4	IPI00248359	1059	84119	96	45	19	6	1.131	0.859	0.874	Tax_
4	1	IPI00003865	5231	78964	215	151	52	39	0.920	0.788	0.864	Tax_
4	2	IPI00304925	5056	77588	220	132	46	35	0.961	0.709	0.897	Tax_
4	3	IPI00514377	5033	77574	219	131	46	35	0.961	0.710	0.897	Tax_
4	4	IPI00003362	4199	81404	222	119	50	39	0.989	0.932	0.838	Tax_
4	5	IPI00007765	3068	81502	151	70	56	31	0.924	0.854	0.775	Tax_
4	6	IPI00301277	2308	78147	129	67	37	16	0.964	0.694	0.819	Tax_
4	7	IPI00339269	1828	77116	97	44	33	10	0.977	0.746	0.870	Tax_
4	8	IPI00397340	602	78659	45	19	22	8	0.966	0.895	0.960	Tax_
5	1	IPI00479186	5182	63836	275	161	53	40	1.048	0.961	0.909	Tax_
5	2	IPI00784179	4598	63984	245	142	51	38	1.058	0.940	0.886	Tax_
5	3	IPI00743867	458	10908	52	18	10	4	1.010	1.020	0.911	Tax_
6	1	IPI00219018	5147	39928	194	115	36	25	0.927	0.973	0.959	Tax_
7	1	IPI00011654	4879	52313	241	144	29	23	1.049	1.119	1.007	Tax_

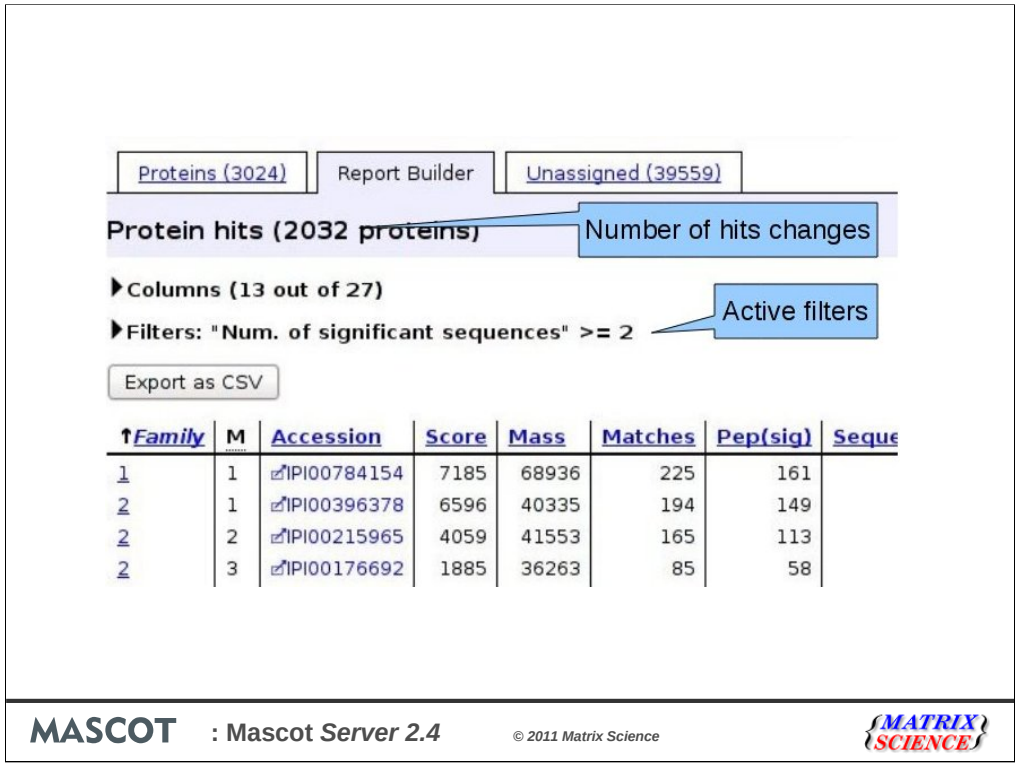
After choosing apply, the table is displayed with the new settings. We have some 3200 hits here, but maybe they are not all of interest. This is where filtering comes in.

First, let's filter out one-hit wonders, to comply with MCP guidelines. The relevant column is Seq(sig), which is the number of significant distinct sequence matches in the protein. There are no one-hit wonders in the first 24 hits, shown here, but there are dozens further down in the list. By the way, the column names are abbreviated here to save space, but you can always see the full title in a tooltip.

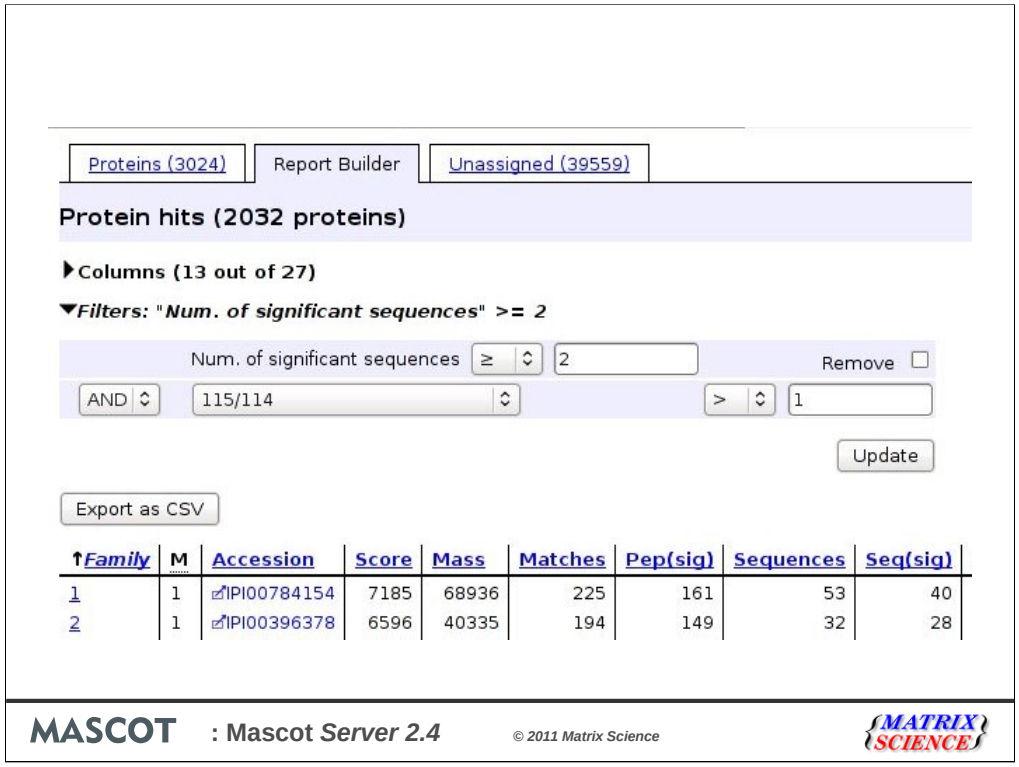


Family	M	Accession	Score	Mass	Matches	Pep(sig)	Sequences
1	1	IPI00784154	7185	68936	225	161	53
2	1	IPI00396378	6596	40335	194	149	32

To filter out the one-hit wonder proteins, you simply select the column by which to filter, enter a value, and click Filter.

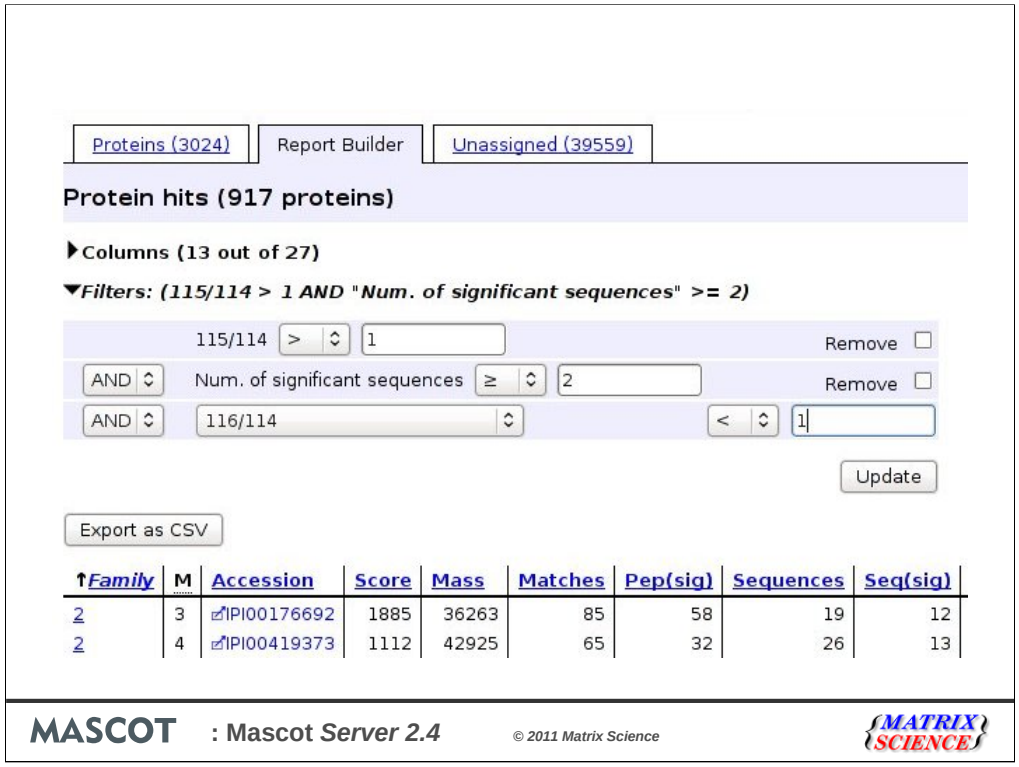


The table reloads and the number of proteins has reduced from 3248 to 2032, so there were quite a few one-hit wonders. Any currently active filters are shown above the table, and the number of proteins is updated every time a filter is applied. If the number doesn't change, it means no protein hits were removed.

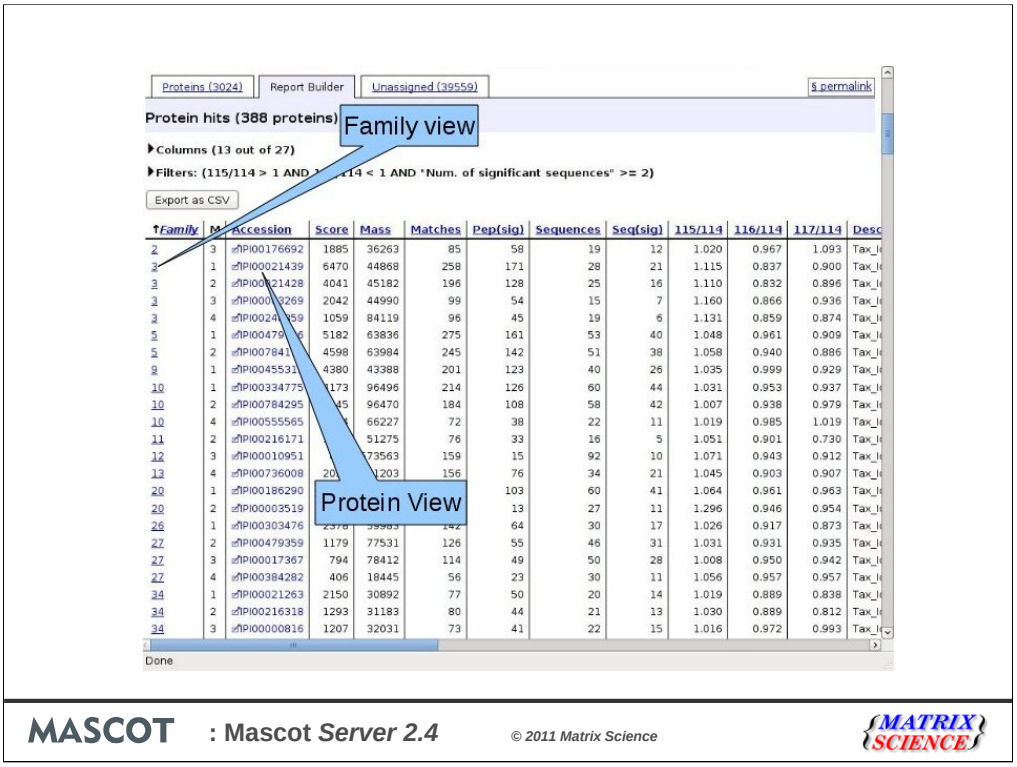


↑Family	M	Accession	Score	Mass	Matches	Pep(sig)	Sequences	Seq(sig)
1	1	IPI00784154	7185	68936	225	161	53	40
2	1	IPI00396378	6596	40335	194	149	32	28

Let's assume you want to know which proteins were up regulated in the sample labelled with iTRAQ 115 relative to the 114 sample and simultaneously down regulated in 116. You just need to add two more filters, the first of which is shown here.



The filters need to be added one at a time, and the table is updated between each addition.



Now we're down to 388 proteins, of which all are up regulated in 115, as you can see here, and down regulated in 116, relative to 114. One-hit wonders excluded.

If you want, you can drill down into the data: click on the accession string to get the regular Protein View. Click on the family number and the report will switch to the Proteins tab, scrolling and changing page if necessary to display the selected protein. This allows you to examine the peptide level data.

If you export as CSV now, the exported data contains exactly what you see in this table, apart from the formatting.

Mascot 2.4

Improvements in reports

- Report Builder
- Preferred taxonomy
- FDR button
- Site analysis (delta score)

Database Manager

Other improvements

- Quantitation of modified peptides

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

That was an overview of Report Builder. Now, let's take a brief look at a new format control called Preferred Taxonomy

Preferred taxonomy

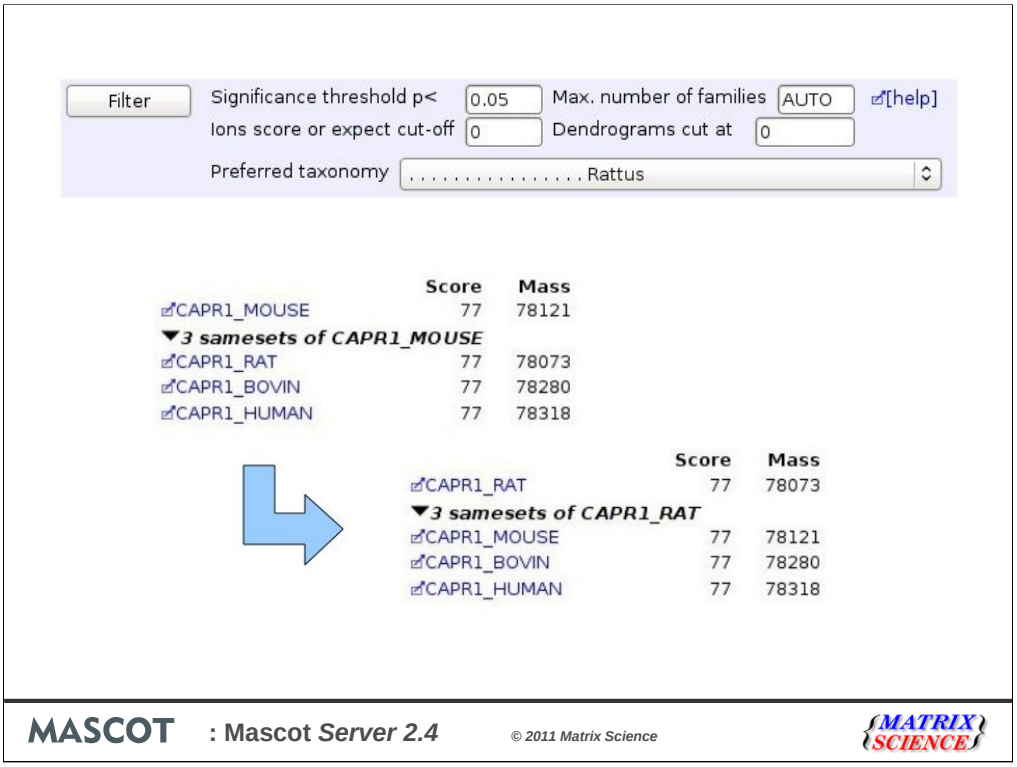
"I selected Bos taurus as a taxonomy filter, but the report shows a protein from chimpanzee?"

NCBI nr entries have multiple taxa

Same-set proteins may have different taxa

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

In some databases, like NCBI nr, each entry represents multiple proteins. In Mascot 2.3 and earlier, it is always the first protein in the title line that is displayed in reports. If you applied a taxonomy filter for, say, Bos taurus, and the protein entry that matched had something else as the first protein, you might see something you didn't expect. Similarly, if you get same-set hits for a protein, there's by definition no evidence to differentiate between them. You might get a hit for, say, the same protein in Mus musculus and Rattus norvegicus, which would be grouped as same-sets. Which one is shown in Mascot 2.3 and earlier depends on simple sort order rules applied to the accession string.



The new format control allows you to specify a preferred taxonomy to be selected as the anchor protein in such cases.

In the NCBI nr case, you can select Bos taurus here to make the Bos taurus entry the default one. This means the correct accession string and description line are swapped in, if the protein entry has a Bos taurus protein.

In the same-set protein case, as you can see here, you can select Rattus as the preferred taxonomy to make CAPR1_RAT the preferred same-set protein.

In Mascot 2.4, the additional taxonomy information required for this control is saved in the results file. The preferred taxonomy control will always be available for new searches. If a file is from Mascot 2.3 or earlier, the databases that were used in the search need to be configured and online. Otherwise the control will be hidden, because there would be no way of getting the required taxonomy information.

Mascot 2.4

Improvements in reports

- Report Builder
- Preferred taxonomy
- FDR button
- Site analysis (delta score)

Database Manager

Other improvements

- Quantitation of modified peptides

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

False Discovery Rate (FDR) is the number of significant peptide matches in the decoy results divided by the number of significant matches in the target results.

Auto-adjust to a required FDR

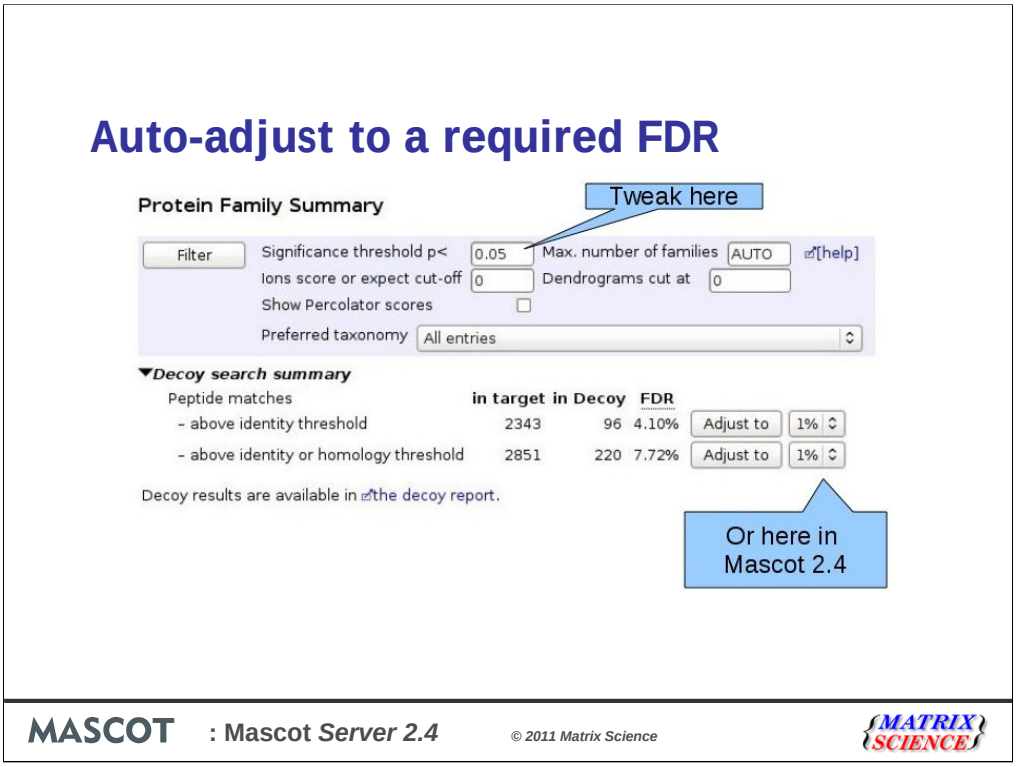
How do I get to 1% FDR?

- In Mascot 2.3: adjust the significance threshold by trial and error
- In Mascot 2.4: click a button.

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

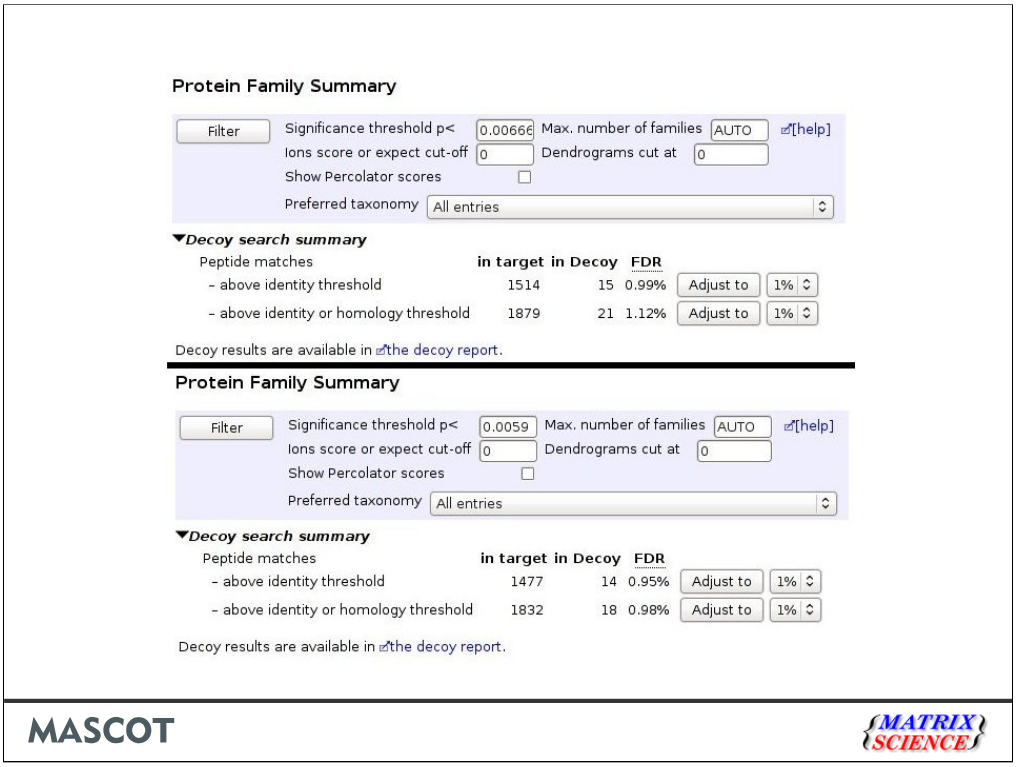
The way to change the FDR in the report is to change the significance threshold in the format controls.

In Mascot 2.3 and earlier, you had to use trial and error to find the significance threshold that gave the required FDR. Slightly tedious. In Mascot 2.4, you can click a button and have make the adjustment automatically.



For example, this search of the ABRF iPRG2008 data set has a 4.1% FDR when the significance threshold is set to 5%. By trial and error, you can get the count above the homology threshold to 0.99% FDR by setting significance threshold to 0.0065. But you need to wait for caching to finish after each change in format controls.

In Mascot 2.4, you can select a target FDR from a dropdown list and click "Adjust" to have the significance threshold adjusted so that you get 1%FDR, or whichever FDR you chose. The list of FDR percentages is configurable by editing mascot.dat.



Note that the setting to get a given FDR for the identity threshold will be a little different from the setting for the homology threshold. Usually, you will see better sensitivity by using the homology threshold. Here, we see 1832 matches by using the homology threshold compared with 1514 matches for the identity threshold

Mascot 2.4

Improvements in reports

- Report Builder
- Preferred taxonomy
- FDR button
- Site analysis (delta score)

Database Manager

Other improvements

- Quantitation of modified peptides

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

The final new feature in the reports that I'd like to tell you about is modification site analysis.

Site specificity delta score

Which sites were modified?

```
Score > 30 indicates identity
Score > 25 indicates homology
304 1   58   2.4e-05 ▼1   U  R.GLKEGMNPSYDEYADSDEDQHDAYLER.M + Phospho (ST)
304 1   37    0.0033 2      R.GLKEGMNPSYDEYADSDEDQHDAYLER.M + Phospho (Y)
304 1   32      0.01 3      R.GLKEGMNPSYDEYADSDEDQHDAYLER.M + Phospho (Y)
304 1   22      0.11 4      R.GLKEGMNPSYDEYADSDEDQHDAYLER.M + Phospho (Y)
304 1   18      0.23 5      R.GLKEGMNPSYDEYADSDEDQHDAYLER.M + Phospho (ST)
```

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

You have no doubt seen cases like this. These are the top scoring matches to a particular spectrum. The peptide is clearly phosphorylated, but can we be sure of the phosphorylation site? Is only the top match correct, or do we have a mixture of peptides modified at different sites?

Site specificity delta score

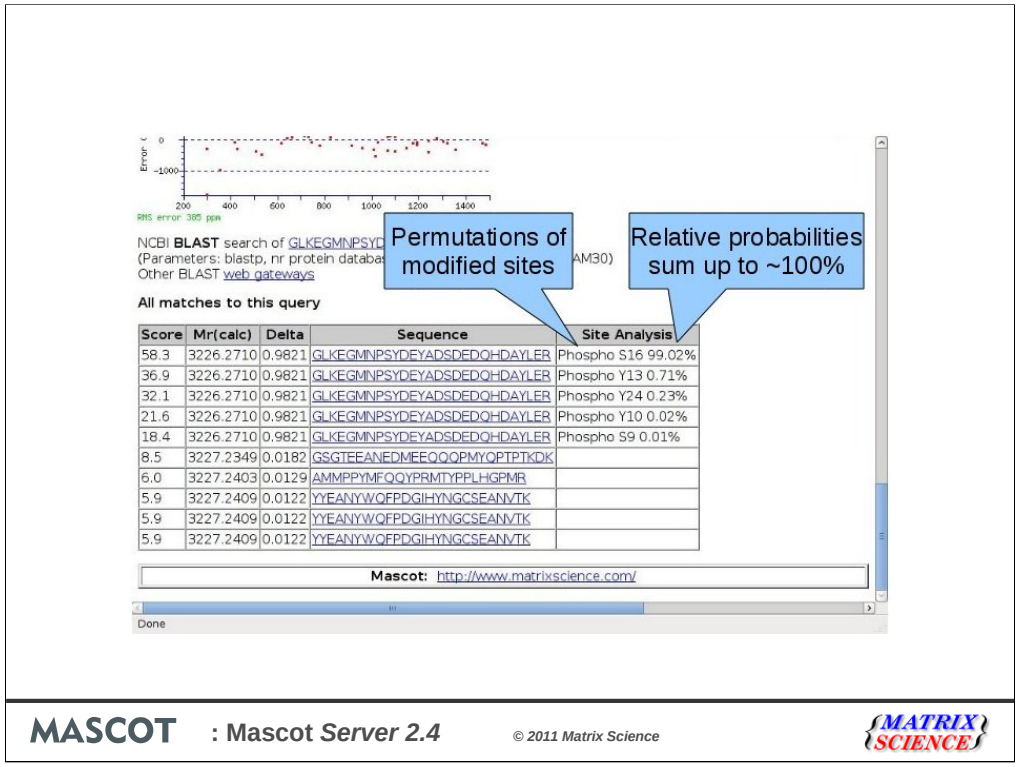
Confident Phosphorylation Site Localization Using the Mascot Delta Score

Mikhail M. Savitski‡, Simone Lemeer§, Markus Boesche‡, Manja Lang‡, Toby Mathieson‡, Marcus Bantscheff‡||, and Bernhard Kuster§¶||

Molecular & Cellular Proteomics 10.2 10.1074/mcp.M110.003830–1

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

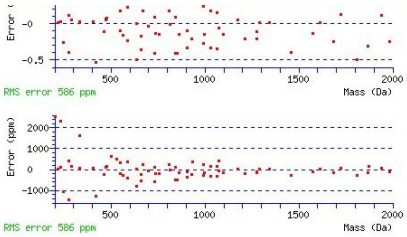
At last year's user meeting, Bernhard Kuster described how the Mascot score difference could be used for reliable site assignment when there are matches to the same spectrum with different arrangements of modifications. This work has now appeared in MCP. We have implemented this approach in the Mascot 2.4 Peptide View report. The only real difference is that we have generalised it to work for any variable modification, not just phosphorylation



Score	Mr(calc)	Delta	Sequence	Site Analysis
58.3	3226.2710	0.9821	GLKEGMNPSYDEYADSDEDQHDAYLER	Phospho S16 99.02%
36.9	3226.2710	0.9821	GLKEGMNPSYDEYADSDEDQHDAYLER	Phospho Y13 0.71%
32.1	3226.2710	0.9821	GLKEGMNPSYDEYADSDEDQHDAYLER	Phospho Y24 0.23%
21.6	3226.2710	0.9821	GLKEGMNPSYDEYADSDEDQHDAYLER	Phospho Y10 0.02%
18.4	3226.2710	0.9821	GLKEGMNPSYDEYADSDEDQHDAYLER	Phospho S9 0.01%
8.5	3227.2349	0.0182	GSGTEEANEDMEEQQQPMYQPTPTKDK	
6.0	3227.2403	0.0129	AMMPPYMFQQYPRMTYPPLHGPMR	
5.9	3227.2409	0.0122	YYEANYWQFPDGIHYNGCSEANVTK	
5.9	3227.2409	0.0122	YYEANYWQFPDGIHYNGCSEANVTK	
5.9	3227.2409	0.0122	YYEANYWQFPDGIHYNGCSEANVTK	

This is the same query as in the earlier slide. You can see that probabilities have been assigned to the five alternative site assignments. Because all possible arrangements are listed, the probabilities sum up to 100% (well, 99.99% due to rounding error). By definition, a score difference of 10 means a factor of 10 in probability. If there are only two possible assignments and their scores are 48 and 38, the relative confidence that the first one is correct is 91%. The relative confidence that the second one is correct must then be 9%.

In this query, there are more than two matches, but you can still see that the matches with scores 32 and 22 differ by about a factor of 10 in relative probability. This calculation assumes that at least the first match in the query is above significance threshold, so it can be assumed to be a correct match. If there are no alternative sites in the sequence where modification could have occurred, then there is no ambiguity and no site assignment probabilities are reported.



NCBI **BLAST** search of KAADALLLKVNQIGTLSESIK
(Parameters: blastp, nr protein database, expect=20000, no filter, PAM30)
Other BLAST web gateways

All matches to this query

Score	Mr(calc)	Delta	Sequence	Site Analysis
48.6	2269.2893	0.0144	KAADALLLKVNQIGTLSESIK	Deamidated N11, Carbamidomethyl K9; 70.39%
41.0	2269.2893	0.0144	KAADALLLKVNQIGTLSESIK	Deamidated Q12, Carbamidomethyl K9; 12.20%
39.0	2269.2893	0.0144	KAADALLLKVNQIGTLSESIK	Deamidated N11, Carbamidomethyl K1; 7.79%
38.2	2269.2892	0.0144	KAADALLLKVNQIGTLSESIK	Deamidated N11, Carbamidomethyl N-term; 6.48%
32.2	2269.2893	0.0144	KAADALLLKVNQIGTLSESIK	Deamidated Q12, Carbamidomethyl K1; 1.62%
31.5	2269.2892	0.0144	KAADALLLKVNQIGTLSESIK	Deamidated Q12, Carbamidomethyl N-term; 1.37%
13.1	2269.2754	0.0283	KKGSQLQLLRQDVDEIKSK	
12.3	2269.2754	0.0283	KKGSQLQLLRQDVDEIKSK	
11.1	2269.2866	0.0171	RTKISRGNKVNKQIEILT	
9.3	2269.3086	-0.0049	KPIVYIPPLDKLFLISNGK	

Mascot: http://www.matrixscience.com/

Done

MASCOT : Mascot *Server 2.4* © 2011 Matrix Science MATRIX SCIENCE

Here's another example, this time with deamidation and carbamidomethylation. In order to calculate site specificities, the results file has to be from Mascot 2.2 or newer. If you want to take advantage of the feature, you will need to re-run a search if it is from an older version.

Mascot 2.4

Improvements in reports

- Report Builder
- Preferred taxonomy
- FDR button
- Site analysis (delta score)

Database Manager

Other improvements

- Quantitation of modified peptides

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

The next new feature I'd like to mention is Database Manager, which makes managing and updating databases very much easier.

Database Manager

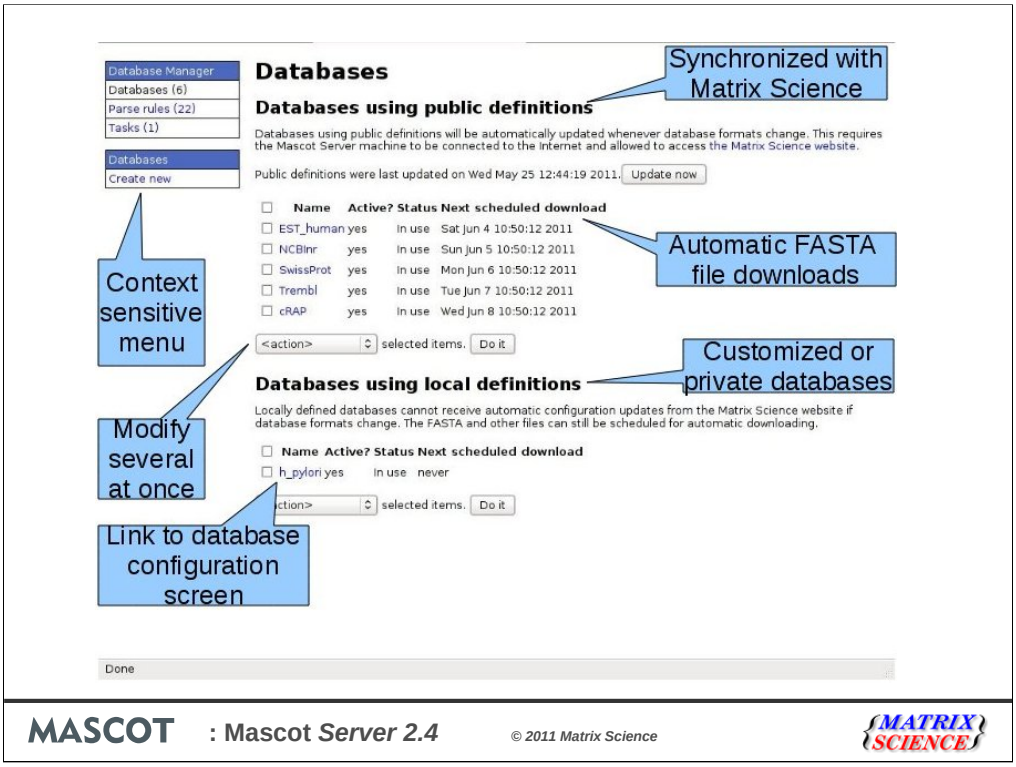
- Web browser interface
- Automated configuration for ‘popular’ databases
- Automatic updates

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

You can do everything in Database Manager that you used to do with Database Maintenance and Database Update, but it is a single, browser-based utiltiy, which makes it a lot easier to use.

Database formats change from time to time. If you've used the old database update script or just downloaded new FASTA files manually, you may have encountered a problem: the accession or description string formats can change, and suddenly the database parse rules don't work anymore. As a result, what should have been a minor update ends up being a debugging session.

The solution we offer is automatically updating the configurations for databases such as SwissProt and NCBInr, based on settings downloaded from the Matrix Science website. This means that, for popular databases, the only thing you need to do is choose which one you want to have enabled on the server, allow updates from the public website, and you're done. If a database format changes, Database Manager will update the configuration before downloading the next new FASTA file.



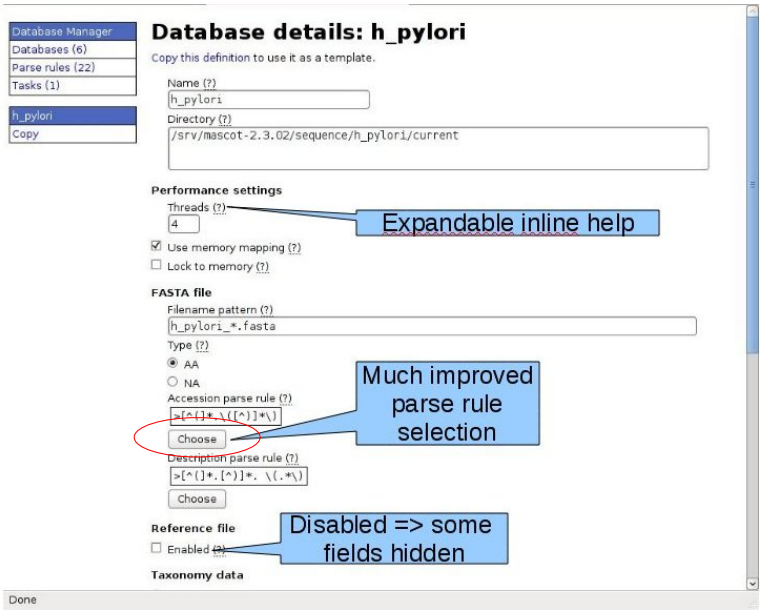
There are many improvements and I won't have time to go through all of them, but we can look at a few key ones.

The overview page shows all the databases you have configured, whether they are active or not. The databases are divided into two categories: public definitions and local definitions.

Public definition means the configuration comes from the Matrix Science website. These are databases like SwissProt and NCBI nr. Public definitions are automatically kept up to date by periodically synchronising with the Matrix Science website. You can choose which databases are to be updated automatically and how often.

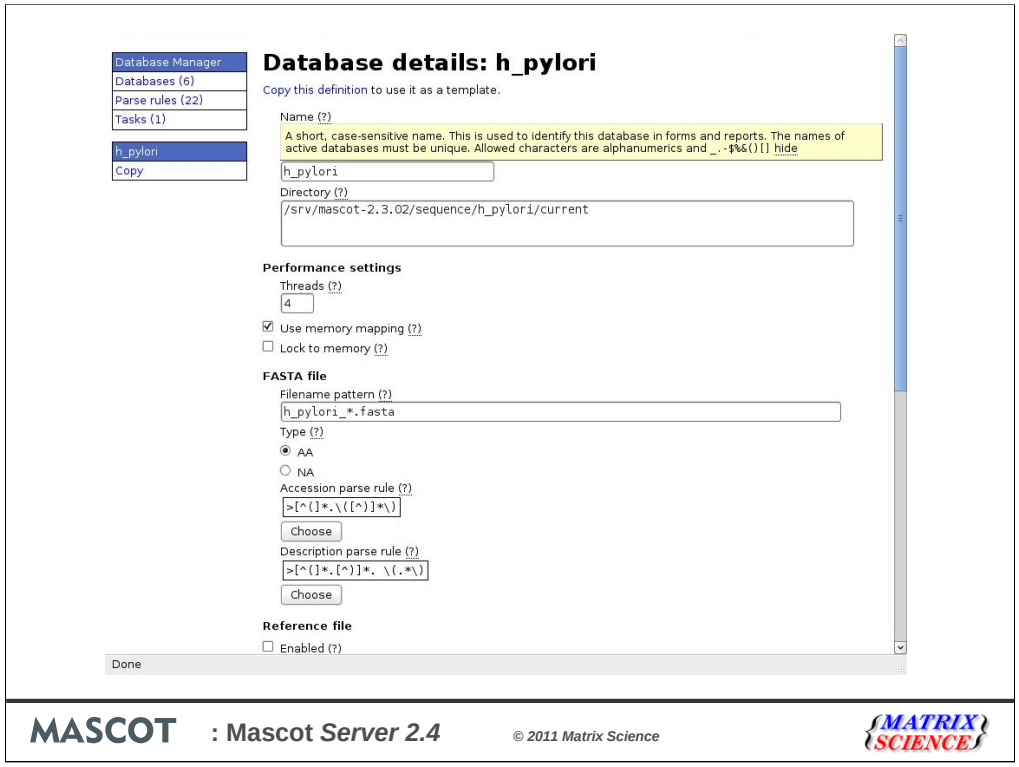
Local definitions are databases that we don't have configuration information for. For example, if you have a database compiled in-house or a single organism genome. You can still use one of the public database configurations as a template, but the parse rules and URLs won't be kept synchronised with the public definitions.

There's a context-sensitive menu on the left that allows you to switch between sections (databases, parse rules) additional actions are shown, such as "Create a new database" here. You can also activate or deactivate several databases at once by ticking the checkboxes and choosing an action from the dropdown menu. If you now click on the h_pylori link...

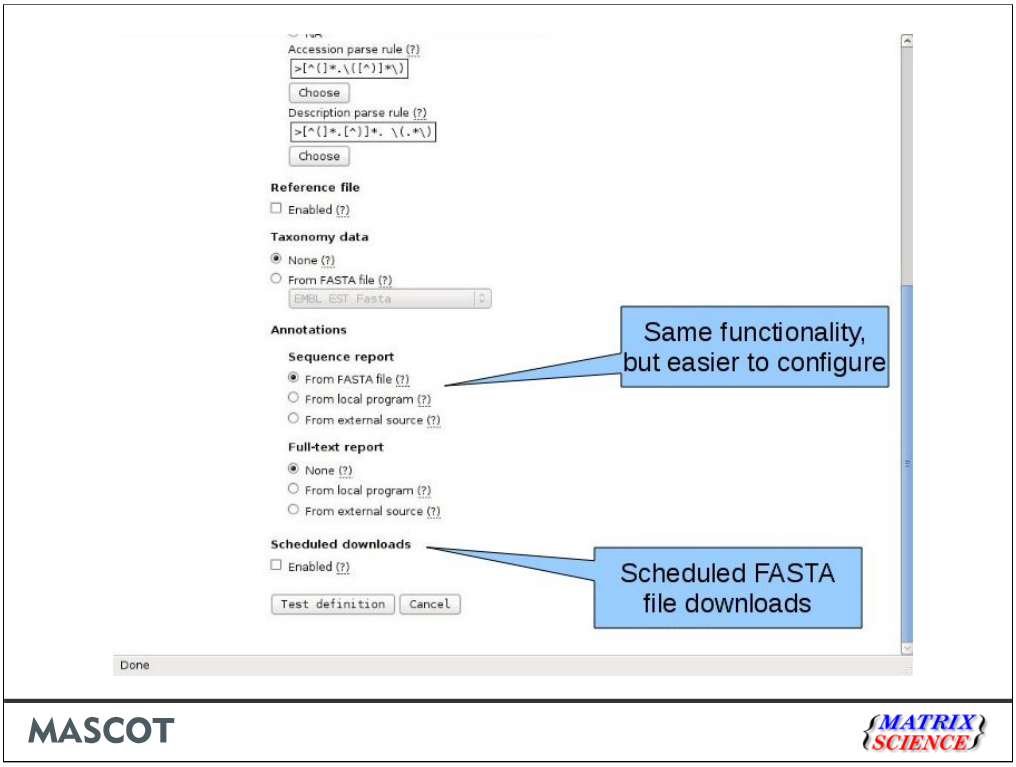


MASCOT : Mascot *Server 2.4* © 2011 Matrix Science MATRIX SCIENCE

This is a local definition for a custom database: protein sequences translated from the heliobacter pylori genome. The configuration information is very similar to that displayed in database maintenance, but slightly restructured. For example, the database directory is separated from the FASTA filename pattern. Selecting parse rules is much improved. Previously you had a dropdown list and there was some guesswork and trial and error involved which parse rule would work. Here, when you click on Choose, you will get a table of all possible parse rules and what they return from the FASTA file. So you will be able to see in a glance which parse rules work or which ones almost work. Controls that are not relevant are hidden. For example, if your database doesn't have a reference file, parse rules for the reference file and related taxonomy entries won't be shown. This reduces clutter quite a bit, and there's less uncertainty which fields you need to fill in. Each little question mark expands to inline help text that is related to the data field. For example,

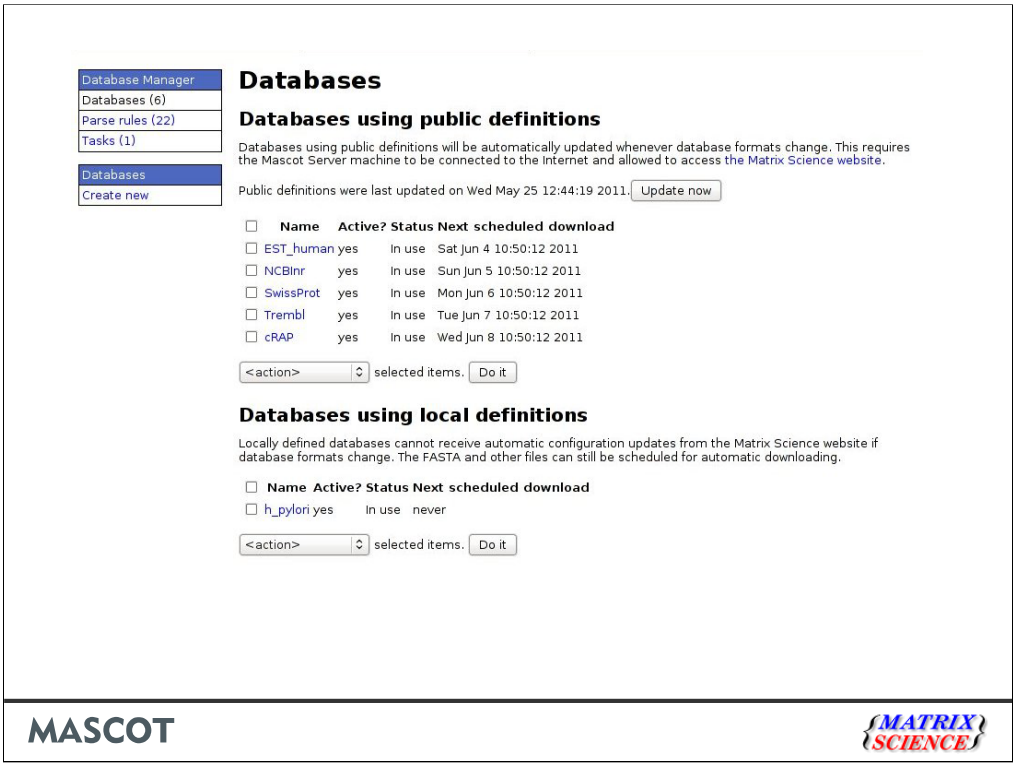


Expanding the help for database name shows which characters are allowed.

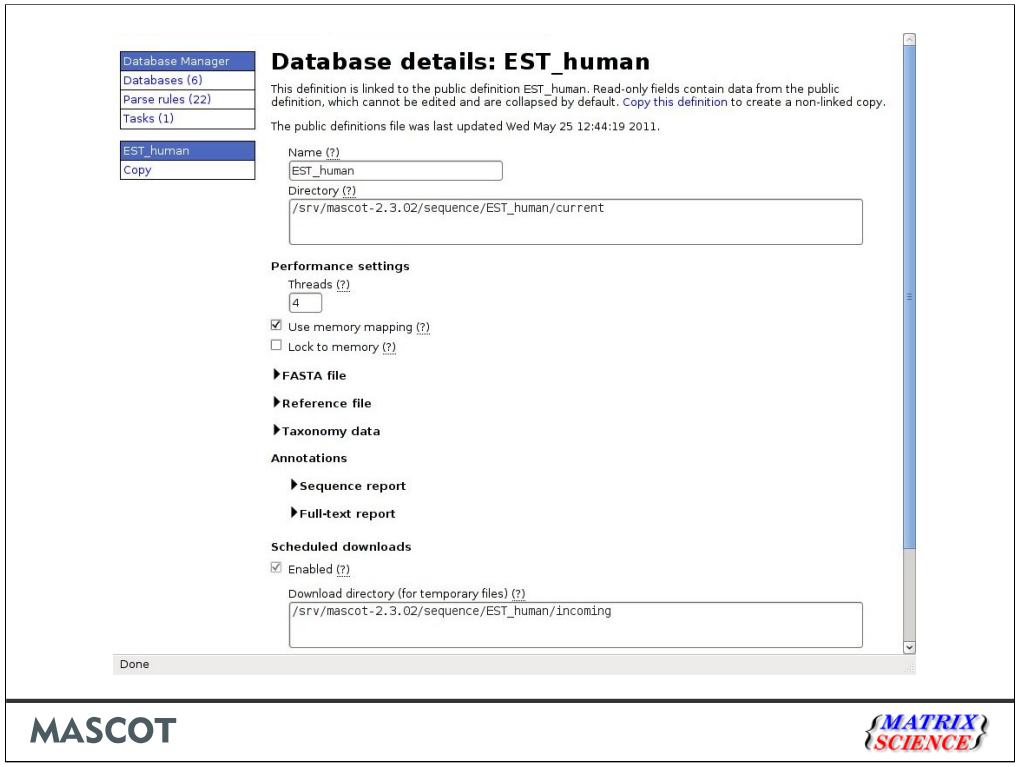


Further down the page, the sequence and full-text report functionality is the same as before. You can take the sequence report from the FASTA file and full-text annotations from the reference file (if the reference file is enabled), or you can configure an external source. However, the configuration has been made easier by dividing the three cases into three selections.

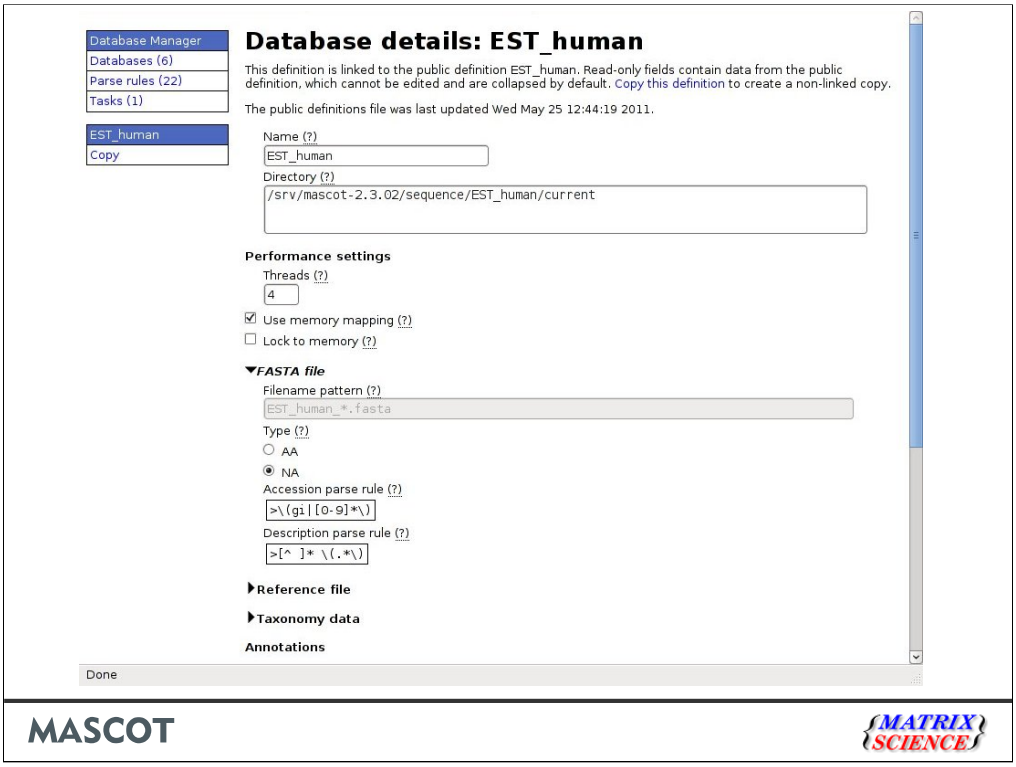
Scheduled downloads replaces the db_update.pl script. If you tick the checkbox here, you'll get fields that allow you to choose when to download (say, every first Tuesday of the month) and where to download from. Database Manager will queue up the downloads and do them in the background.



You can see an item called Tasks in the menu here. This is where you can view which downloads are executing at the moment and how complete they are. You can also pause and cancel downloads. If you want to update the files before the scheduled time, just tick the checkbox and select the download action from the dropdown menu. Let's look briefly at EST_human, which is a public definition.



You can see that the configuration looks mostly the same, but most of the sections are collapsed. Almost all of the fields and controls in a public definition are read only. They are set by downloading configuration details from the Matrix Science website. Let's expand the FASTA file configuration.



There's no way to edit the file name pattern and no way to edit the parse rules. This is deliberate, because one of the main reasons database updates go wrong is because parse rules change. All of the read only fields will be kept synchronised with the Matrix Science website. So, if database formats change, you don't need to do anything as long as you're using one of the public definitions. If you want to have NCBInr on your Mascot server, all you need to do is select it from the list, click a button, and the files will be downloaded and the database kept up to date. That's all!

Mascot 2.4

Improvements in reports

- Report Builder
- Preferred taxonomy
- FDR button
- Site analysis (delta score)

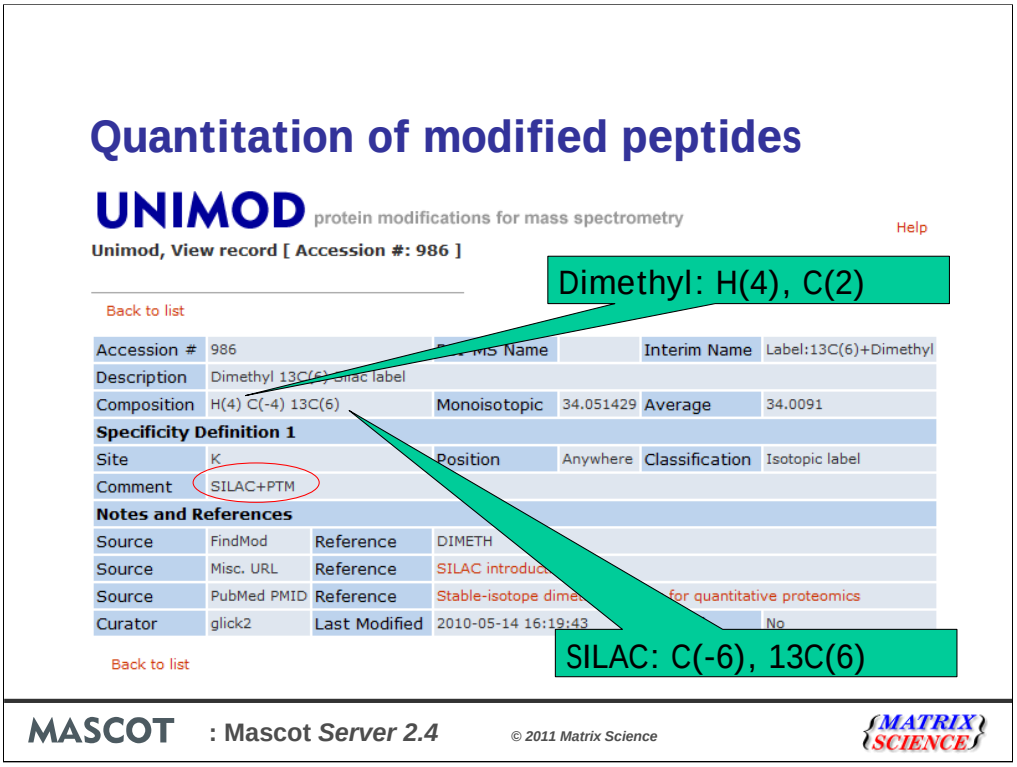
Database Manager

Other improvements

- Quantitation of modified peptides

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

There are many other new features and bug fixes in Mascot 2.4. One I'd like to mention in closing relates to quantitation of peptides that have multiple modifications on the same site



SILAC, for example, is implemented as modifications, often to K or R. Mascot has only ever allowed one modification per site so, if you have a peptide with an artefactual or post-translational modification on K or R, you have to define combination modifications, such as the one shown here.

Dimethylation adds C2H4, while the SILAC label substitutes six 13C atoms for 12C. So, the net change is H(4) C(-4) 13C(6). This is clumsy. You have to specify and select permutations of modifications and labels, which can quickly bring you up against the limits on the maximum number of variable modifications allowed in a search

In Mascot 2.4, if you use a quantitation method with exclusive modifications and you select a variable modification for the same residue, then Mascot will also test for a match to the double modification. There is no longer a requirement to create these combined modifications and it keeps the search space as small as possible.

Mascot Server 2.4

Improvements in reports

- Report Builder
- Preferred taxonomy
- FDR button
- Site analysis (delta score)

Database Manager

Other improvements

- Quantitation of modified peptides

MASCOT : Mascot *Server 2.4* MATRIX SCIENCE

I hope that has been a useful preview of some of the new features in Mascot Server 2.4

When will it be released? We're very close. Hopefully, later this summer!