

# 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

© 2011 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.

## Report Builder

### Mascot 2.3

Proteins (3024)

Quantitation (3248)

Unassigned (39559)

Protein families 1-10 (out of 3024)

### Mascot 2.4

Proteins (3024)

Report Builder

Unassigned (39559)

Protein families 1-10 (out of 3024)

**MASCOT** : Mascot Server 2.4

© 2011 Matrix Science



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

© 2011 Matrix Science

**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.

Proteins (3024) Report Builder Unassigned (39559) [\\$ permalink](#)

Protein hits (3248 proteins)

Columns (12 out of 27) **Edit columns**

Filters: (none)

Export as CSV **Export to Excel** **Top-level proteins**

Family	M	DB	Accession	Score	Mass	Matches	Pep(sig)	Sequence	q(sig)	emPAI	Description
1		IPi_human	IPi00784154	7185	68936	225	161		40	128.20	Tax_id=9606 H
2	1	IPi_human	IPi00396378	6596	40335	194	149	32	28	892.67	Tax_id=9606 H
2	2	IPi_human	IPi00215965	4059	41553	165	113	31	23	290.01	Tax_id=9606 H
2	3	IPi_human	IPi001176692	1885	36263	85	58	19	12	22.70	Tax_id=9606 H
2	4	IPi_human	IPi004	1112	42925	65	32	26	13	7.74	Tax_id=9606 H
2	5	IPi_human	IPi000			23	11	12	8	2.66	Tax_id=9606 H
3	1	IPi_human	IPi000			258	171	28	21	397.70	Tax_id=9606 A
3	2	IPi_human	IPi000			196	128	25	16	77.83	Tax_id=9606 A
3	3	IPi_human	IPi00003269	2042	44990	99	54	15	7	4.51	Tax_id=9606 H
3	4	IPi_human	IPi00248359	1059	84119	96	45	19	6	0.92	Tax_id=9606 P
4	1	IPi_human	IPi00003865	5231	78964	215	151	52	39	139.45	Tax_id=9606 H
4	2	IPi_human	IPi00304925	5056	77588	220	132	46	35	85.72	Tax_id=9606 H
4	3	IPi_human	IPi00514377	5033	77574	219	131	46	35	85.72	Tax_id=9606 H
4	4	IPi_human	IPi00003362	4199	81404	222	119	50	39	52.85	Tax_id=9606 H
4	5	IPi_human	IPi00007765	3068	81502	151	70	56	31	11.14	Tax_id=9606 S
4	6	IPi_human	IPi00301277	2308	78147	129	67	37	16	5.68	Tax_id=9606 H
4	7	IPi_human	IPi00339269	1828	77116	97	44	33	10	2.13	Tax_id=9606 H
4	8	IPi_human	IPi00397340	602	78659	45	19	22	8	1.31	Tax_id=9606 P
5	1	IPi_human	IPi00479186	5182	63836	275	161	53	40	159.18	Tax_id=9606 P
5	2	IPi_human	IPi00784179	4598	63984	245	142	51	38	102.35	Tax_id=9606 S
5	3	IPi_human	IPi00743867	458	10908	52	18	10	4	10.78	Tax_id=9606 S
5	1	IPi_human	IPi00219018	5147	39928	194	115	36	25	314.86	Tax_id=9606 G
7	1	IPi_human	IPi00011654	4879	52313	241	144	29	23	169.95	Tax_id=9606 T

Sort by any column

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

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.

Proteins (3024) Report Builder Unassigned (39559) [5 permalink](#)

**Protein hits (3248 proteins)**

▼Columns (12 out of 27)

**Enabled**

- Family
- Member
- Accession
- Score
- Mass
- Num. of matches
- Num. of significant matches
- Num. of sequences
- Num. of significant sequences
- 115/114
- 116/114
- 117/114
- Description

**Available**

**Protein hits**

- Database
- emPAI
- 115/114**
- Number of peptides (115/114)
- Significant (115/114)
- Not-normal (115/114)
- SD(geo) (115/114)
- 116/114**
- Number of peptides (116/114)
- Significant (116/114)
- Not-normal (116/114)
- SD(geo) (116/114)
- 117/114**
- Number of peptides (117/114)
- Significant (117/114)
- Not-normal (117/114)
- SD(geo) (117/114)

↑ ↓ Apply

►Filters: (none)

Export as CSV

Family	M	DB	Accession	Score	Mass	Matches	Pep(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	IPI_human	<a href="#">IPI00784154</a>	7185	68936	225	161	53	40	128.20	Tax_id=9606 P
2	1	IPI_human	<a href="#">IPI00396378</a>	6596	40335	194	149	32	28	892.67	Tax_id=9606 P
2	2	IPI_human	<a href="#">IPI00215965</a>	4059	41553	165	113	31	23	290.01	Tax_id=9606 P
2	3	IPI_human	<a href="#">IPI00176692</a>	1885	36263	85	58	19	12	22.70	Tax_id=9606 P

Done

---

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

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.

Proteins (3024) Report Builder Unassigned (32559) [\\$ permalink](#)

**Protein hits (3248 proteins)**

Columns (13 out of 27) **Filter out unwanted proteins**

Filters: (none)

Export as CSV

Family	M	Accession	Score	Mass	Matches	Pep(sig)	Sequences	Seq(sig)	115/114	116/114	117/114	Desc
1	1	hPI00784154	7185	68936	225	161	53	40	0.945	0.818	0.806	Tax_
2	1	hPI00396378	6596	40335	194	149	32	28	0.986	0.922	0.812	Tax_
2	2	hPI00215965	4059	41553	165	113	31	23	0.988	0.897	0.939	Tax_
2	3	hPI00176692	1885	36263	85	58	19	12	1.020	0.967	1.093	Tax_
2	4	hPI00419373	1112	42925	65	32	26	13	1.028	1.062	0.960	Tax_
2	5	hPI00011913	329	33698	23	11	12	8	1.047	1.037	1.221	Tax_
3	1	hPI00021439	6470	44868	258	171	28	21	1.115	0.837	0.900	Tax_
3	2	hPI00021428	4041	45182	196	128	25	16	1.110	0.832	0.896	Tax_
3	3	hPI00003269	2042	44990	99	54	15	7	1.160	0.866	0.936	Tax_
3	4	hPI00248359	1059	84119	96	45	19	6	1.131	0.859	0.874	Tax_
4	1	hPI00003865	5231	78964	215	151	52	39	0.920	0.788	0.864	Tax_
4	2	hPI00304925	5056	77588	220	132	46	35	0.961	0.709	0.897	Tax_
4	3	hPI00514377	5033	77574	219	131	46	35	0.961	0.710	0.897	Tax_
4	4	hPI00003362	4199	81404	222	119	50	39	0.989	0.932	0.838	Tax_
4	5	hPI00007765	3068	81502	151	70	56	31	0.924	0.854	0.775	Tax_
4	6	hPI00301277	2308	78147	129	67	37	16	0.964	0.694	0.819	Tax_
4	7	hPI00339269	1828	77116	97	44	33	10	0.977	0.746	0.870	Tax_
4	8	hPI00397340	602	78659	45	19	22	8	0.966	0.695	0.960	Tax_
5	1	hPI00479186	5182	63836	275	161	53	40	1.048	0.961	0.909	Tax_
5	2	hPI00784179	4598	63984	245	142	51	38	1.058	0.940	0.886	Tax_
5	3	hPI00743867	458	10908	52	18	10	4	1.010	1.020	0.911	Tax_
6	1	hPI00219018	5147	39928	194	115	36	25	0.927	0.973	0.959	Tax_
7	1	hPI00011654	4879	52313	241	144	29	23	1.049	1.119	1.007	Tax_

Done

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

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.

[Proteins \(3024\)](#) | [Report Builder](#) | [Unassigned \(39559\)](#)

**Protein hits (3248 proteins)**

▶ **Columns (13 out of 27)**

▼ **Filters: (none)**

Num. of significant sequences  ≥

<a href="#">↑ Family</a>	<a href="#">M</a>	<a href="#">Accession</a>	<a href="#">Score</a>	<a href="#">Mass</a>	<a href="#">Matches</a>	<a href="#">Pep(sig)</a>	<a href="#">Sequences</a>
<a href="#">1</a>	1	<a href="#">PI00784154</a>	7185	68936	225	161	53
<a href="#">2</a>	1	<a href="#">PI00396378</a>	6596	40335	194	149	32

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

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

[Proteins \(3024\)](#)
[Report Builder](#)
[Unassigned \(39559\)](#)

**Protein hits (2032 proteins)**
Number of hits changes

▶ **Columns (13 out of 27)**

▶ **Filters: "Num. of significant sequences" >= 2**
Active filters

<u>Family</u>	<u>M</u>	<u>Accession</u>	<u>Score</u>	<u>Mass</u>	<u>Matches</u>	<u>Pep(sig)</u>	<u>Seque</u>
<u>1</u>	1	<a href="#">PI00784154</a>	7185	68936	225	161	
<u>2</u>	1	<a href="#">PI00396378</a>	6596	40335	194	149	
<u>2</u>	2	<a href="#">PI00215965</a>	4059	41553	165	113	
<u>2</u>	3	<a href="#">PI00176692</a>	1885	36263	85	58	

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

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.

Proteins (3024) Report Builder Unassigned (39559)

**Protein hits (2032 proteins)**

► Columns (13 out of 27)

▼ Filters: "Num. of significant sequences"  $\geq 2$

Num. of significant sequences  $\geq$  2 Remove

AND 115/114 > 1

Update

Export as CSV

Family	M	Accession	Score	Mass	Matches	Pep(sig)	Sequences	Seq(sig)
<u>1</u>	1	<a href="#">iPI00784154</a>	7185	68936	225	161	53	40
<u>2</u>	1	<a href="#">iPI00396378</a>	6596	40335	194	149	32	28

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

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.

[Proteins \(3024\)](#) | [Report Builder](#) | [Unassigned \(39559\)](#)

**Protein hits (917 proteins)**

► **Columns (13 out of 27)**

▼ **Filters: (115/114 > 1 AND "Num. of significant sequences" >= 2)**

115/114 > 1  Remove

AND Num. of significant sequences >= 2  Remove

AND 116/114 < 1

<a href="#">↑Family</a>	<a href="#">M</a>	<a href="#">Accession</a>	<a href="#">Score</a>	<a href="#">Mass</a>	<a href="#">Matches</a>	<a href="#">Pep(sig)</a>	<a href="#">Sequences</a>	<a href="#">Seq(sig)</a>
<a href="#">2</a>	3	<a href="#">PI00176692</a>	1885	36263	85	58	19	12
<a href="#">2</a>	4	<a href="#">PI00419373</a>	1112	42925	65	32	26	13

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

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

Proteins (3024) Report Builder Unassigned (39559) [\\$ permalink](#)

Protein hits (388 proteins) **Family view**

Columns (13 out of 27)

Filters: (115/114 > 1 AND 116/114 < 1 AND "Num. of significant sequences" >= 2)

Export as CSV

Family	Accession	Score	Mass	Matches	Pep(sig)	Sequences	Seq(sig)	115/114	116/114	117/114	Desc
2	PI00176692	1885	36263	85	58	19	12	1.020	0.967	1.093	Tax_H
3	PI0021439	6470	44868	258	171	28	21	1.115	0.837	0.900	Tax_H
3	PI0021428	4041	45182	196	128	25	16	1.110	0.832	0.896	Tax_H
3	PI0003269	2042	44990	99	54	15	7	1.160	0.866	0.936	Tax_H
3	PI0024759	1059	84119	96	45	19	6	1.131	0.859	0.874	Tax_H
3	PI0047935	5182	63836	275	161	53	40	1.048	0.961	0.909	Tax_H
5	PI007841	4598	63984	245	142	51	38	1.058	0.940	0.886	Tax_H
9	PI0045531	4380	43388	201	123	40	26	1.035	0.999	0.929	Tax_H
10	PI00334775	1173	96496	214	126	60	44	1.031	0.953	0.937	Tax_H
10	PI00784295	145	96470	184	108	58	42	1.007	0.938	0.979	Tax_H
10	PI00555655	66227	72	38	22	11	1.019	0.985	1.019	Tax_H	
11	PI00216171	51275	76	33	16	5	1.051	0.901	0.730	Tax_H	
12	PI00010951	73563	159	15	92	10	1.071	0.943	0.912	Tax_H	
13	PI00736008	201	1203	156	76	34	21	1.045	0.903	0.907	Tax_H
20	PI00186290	103	60	41	1.064	0.961	0.963	Tax_H			
20	PI00003519	13	27	11	1.296	0.946	0.954	Tax_H			
26	PI00303476	2378	39983	142	64	30	17	1.026	0.917	0.873	Tax_H
27	PI00479359	1179	77531	126	55	46	31	1.031	0.931	0.935	Tax_H
27	PI00017367	794	78412	114	49	50	28	1.008	0.950	0.942	Tax_H
27	PI00384282	406	18445	56	23	30	11	1.056	0.957	0.957	Tax_H
34	PI00021263	2150	30892	77	50	20	14	1.019	0.889	0.838	Tax_H
34	PI00216318	1293	31183	80	44	21	13	1.030	0.889	0.812	Tax_H
34	PI00000816	1207	32031	73	41	22	15	1.016	0.972	0.993	Tax_H

Done

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

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

© 2011 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

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.

Filter    Significance threshold  $p <$      Max. number of families     [\[help\]](#)  
 ions score or expect cut-off     Dendrograms cut at   
 Preferred taxonomy

	Score	Mass
<input checked="" type="checkbox"/> CAPR1_MOUSE	77	78121
▼ 3 same-sets of CAPR1_MOUSE		
<input checked="" type="checkbox"/> CAPR1_RAT	77	78073
<input checked="" type="checkbox"/> CAPR1_BOVIN	77	78280
<input checked="" type="checkbox"/> CAPR1_HUMAN	77	78318

	Score	Mass
<input checked="" type="checkbox"/> CAPR1_RAT	77	78073
▼ 3 same-sets of CAPR1_RAT		
<input checked="" type="checkbox"/> CAPR1_MOUSE	77	78121
<input checked="" type="checkbox"/> CAPR1_BOVIN	77	78280
<input checked="" type="checkbox"/> CAPR1_HUMAN	77	78318



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

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

© 2011 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.

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.

## Auto-adjust to a required FDR

**Protein Family Summary**

**Filter**    Significance threshold p<     Max. number of families     [\[help\]](#)  
ions score or expect cut-off     Dendrograms cut at   
Show Percolator scores   
Preferred taxonomy

▼ **Decoy search summary**

Peptide matches	in target	in Decoy	FDR	
- above identity threshold	2343	96	4.10%	<input type="button" value="Adjust to"/> <input type="text" value="1%"/>
- above identity or homology threshold	2851	220	7.72%	<input type="button" value="Adjust to"/> <input type="text" value="1%"/>

Decoy results are available in [the decoy report](#).

Tweak here

Or here in Mascot 2.4

**MASCOT** : Mascot Server 2.4

© 2011 Matrix Science

**MATRIX**  
**SCIENCE**

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.

**Protein Family Summary**

Filter Significance threshold p< 0.00666 Max. number of families AUTO [help]  
 Ions score or expect cut-off 0 Dendrograms cut at 0  
 Show Percolator scores   
 Preferred taxonomy All entries

**▼Decoy search summary**

Peptide matches	in target	in Decoy	FDR		
- above identity threshold	1514	15	0.99%	Adjust to	1% ↕
- above identity or homology threshold	1879	21	1.12%	Adjust to	1% ↕

Decoy results are available in the decoy report.

---

**Protein Family Summary**

Filter Significance threshold p< 0.0059 Max. number of families AUTO [help]  
 Ions score or expect cut-off 0 Dendrograms cut at 0  
 Show Percolator scores   
 Preferred taxonomy All entries

**▼Decoy search summary**

Peptide matches	in target	in Decoy	FDR		
- above identity threshold	1477	14	0.95%	Adjust to	1% ↕
- above identity or homology threshold	1832	18	0.98%	Adjust to	1% ↕

Decoy results are available in the decoy report.

**MASCOT** 

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

© 2011 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)

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<sup>®</sup>

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

*Molecular & Cellular Proteomics* 10.2

10.1074/mcp.M110.003830-1

**MASCOT** : Mascot Server 2.4

© 2011 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

NCBI BLAST search of **GLKEGMNPSYDEYADSEDEQHDAYLER** (AM30)  
 (Parameters: blastp, nr protein database)  
 Other BLAST [web gateways](#)

All matches to this query

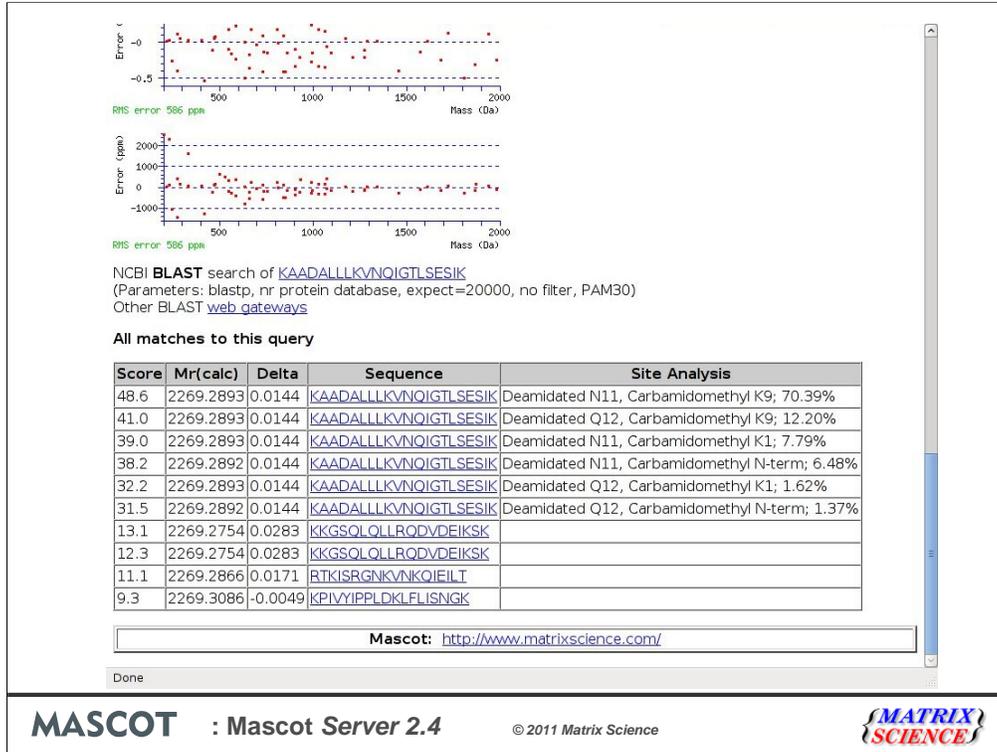
Score	Mr(calc)	Delta	Sequence	Site Analysis
58.3	3226.2710	0.9821	GLKEGMNPSYDEYADSEDEQHDAYLER	Phospho S16 99.02%
36.9	3226.2710	0.9821	GLKEGMNPSYDEYADSEDEQHDAYLER	Phospho Y13 0.71%
32.1	3226.2710	0.9821	GLKEGMNPSYDEYADSEDEQHDAYLER	Phospho Y24 0.23%
21.6	3226.2710	0.9821	GLKEGMNPSYDEYADSEDEQHDAYLER	Phospho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

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

MASCOT : Mascot Server 2.4 © 2011 Matrix Science

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.



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

© 2011 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

© 2011 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 utility, 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 NCBIInr, 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.

**Database Manager**  
Databases (6)  
Parse rules (22)  
Tasks (1)  
Databases  
Create new

### Databases

**Databases using public definitions**

Databases using public definitions will be automatically updated whenever database formats change. This requires the Mascot Server machine to be connected to the internet and allowed to access the Matrix Science website.

Public definitions were last updated on Wed May 25 12:44:19 2011. [Update now](#)

<input type="checkbox"/>	Name	Active?	Status	Next scheduled download
<input type="checkbox"/>	EST_human	yes	In use	Sat Jun 4 10:50:12 2011
<input type="checkbox"/>	NCBI nr	yes	In use	Sun Jun 5 10:50:12 2011
<input type="checkbox"/>	SwissProt	yes	In use	Mon Jun 6 10:50:12 2011
<input type="checkbox"/>	Trembl	yes	In use	Tue Jun 7 10:50:12 2011
<input type="checkbox"/>	cRAP	yes	In use	Wed Jun 8 10:50:12 2011

<action> selected items. [Do it](#)

**Databases using local definitions**

Locally defined databases cannot receive automatic configuration updates from the Matrix Science website if database formats change. The FASTA and other files can still be scheduled for automatic downloading.

<input type="checkbox"/>	Name	Active?	Status	Next scheduled download
<input type="checkbox"/>	h_pylori	yes	In use	never

<action> selected items. [Do it](#)

Done

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

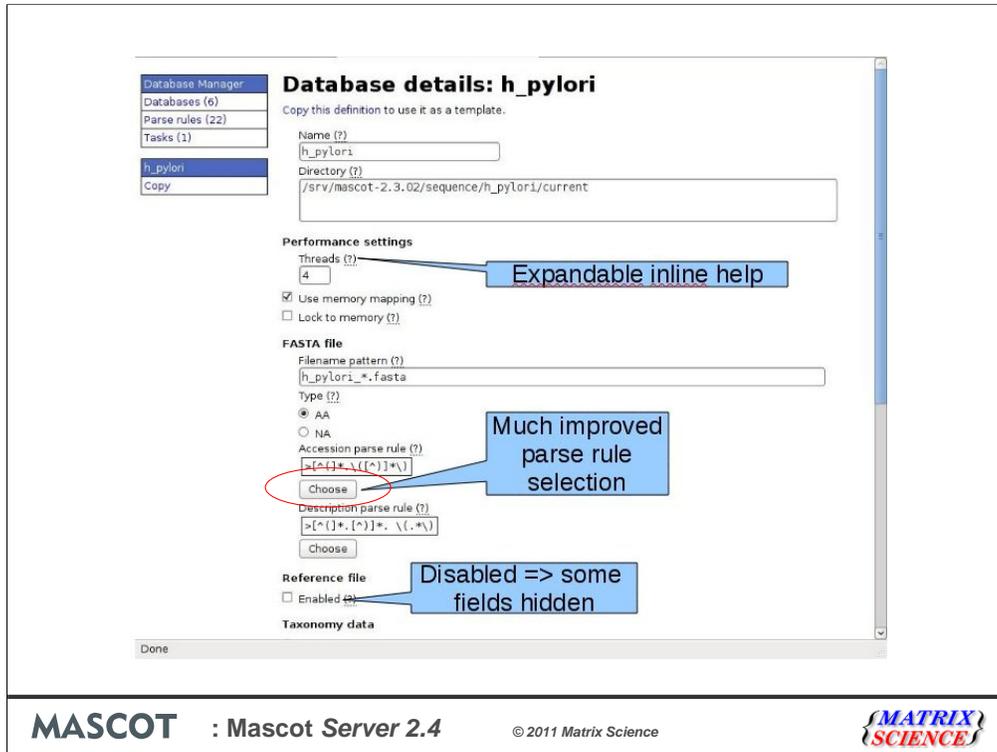
There are many improvements and I won't have time to go through all of them, but we can look at a few key ones.

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...



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,

<b>Database Manager</b>	<b>Database details: h_pylori</b>
Databases (6)	Copy this definition to use it as a template.
Parse rules (22)	
Tasks (1)	
<b>h_pylori</b>	
Copy	

**Name (?)**

A short, case-sensitive name. This is used to identify this database in forms and reports. The names of active databases must be unique. Allowed characters are alphanumerics and `_` `-` `%&()` [hide](#)

**Directory (?)**

**Performance settings**

**Threads (?)**

Use memory mapping (?)

Lock to memory (?)

**FASTA file**

**Filename pattern (?)**

**Type (?)**

AA

NA

**Accession parse rule (?)**

**Description parse rule (?)**

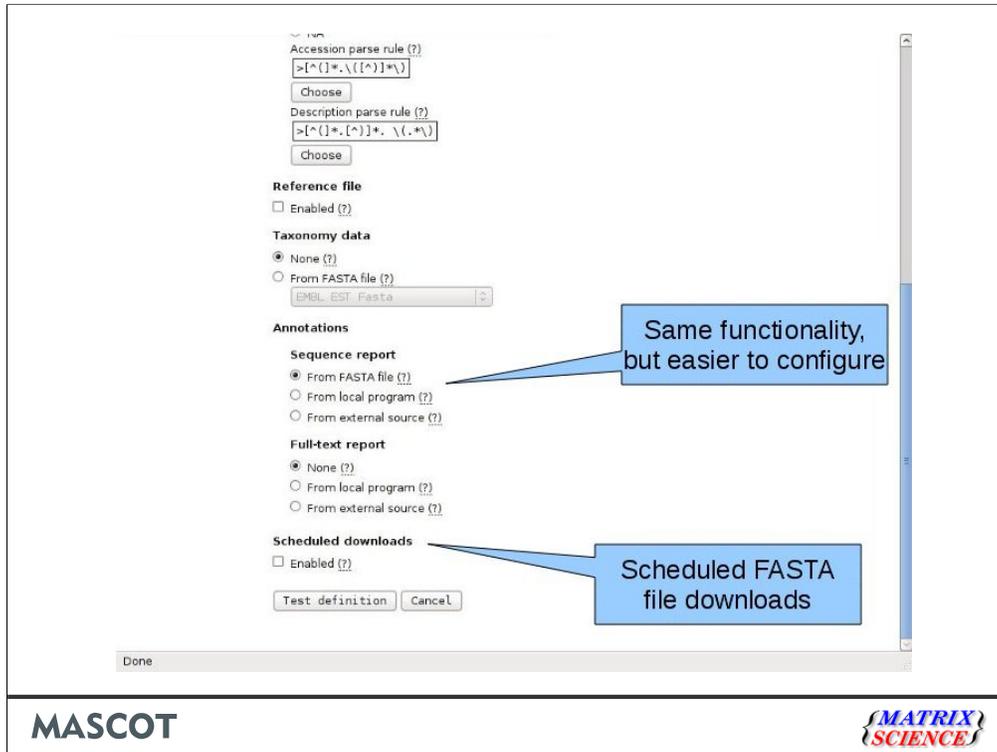
**Reference file**

Enabled (?)

Done

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

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.

Database Manager
Databases (6)
Parse rules (22)
Tasks (1)
Databases
Create new

## Databases

### Databases using public definitions

Databases using public definitions will be automatically updated whenever database formats change. This requires the Mascot Server machine to be connected to the Internet and allowed to access the Matrix Science website.

Public definitions were last updated on Wed May 25 12:44:19 2011. [Update now](#)

<input type="checkbox"/>	Name	Active?	Status	Next scheduled download
<input type="checkbox"/>	EST_human	yes	In use	Sat Jun 4 10:50:12 2011
<input type="checkbox"/>	NCBINr	yes	In use	Sun Jun 5 10:50:12 2011
<input type="checkbox"/>	SwissProt	yes	In use	Mon Jun 6 10:50:12 2011
<input type="checkbox"/>	Trembl	yes	In use	Tue Jun 7 10:50:12 2011
<input type="checkbox"/>	cRAP	yes	In use	Wed Jun 8 10:50:12 2011

selected items.

### Databases using local definitions

Locally defined databases cannot receive automatic configuration updates from the Matrix Science website if database formats change. The FASTA and other files can still be scheduled for automatic downloading.

<input type="checkbox"/>	Name	Active?	Status	Next scheduled download
<input type="checkbox"/>	h_pylori	yes	In use	never

selected items.




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.

**Database Manager**

- Databases (6)
- Parse rules (22)
- Tasks (1)

**EST\_human**

- Copy

### Database details: EST\_human

This definition is linked to the public definition EST\_human. Read-only fields contain data from the public definition, which cannot be edited and are collapsed by default. Copy this definition to create a non-linked copy.

The public definitions file was last updated Wed May 25 12:44:19 2011.

Name (?)  
EST\_human

Directory (?)  
/srv/mascot-2.3.02/sequence/EST\_human/current

**Performance settings**

Threads (?)  
4

Use memory mapping (?)  
 Lock to memory (?)

►FASTA file

►Reference file

►Taxonomy data

**Annotations**

►Sequence report

►Full-text report

**Scheduled downloads**

Enabled (?)

Download directory (for temporary files) (?)  
/srv/mascot-2.3.02/sequence/EST\_human/incoming

Done

**MASCOT** **MATRIX SCIENCE**

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.

<b>Database Manager</b>	<b>Database details: EST_human</b>
Databases (6)	<p>This definition is linked to the public definition EST_human. Read-only fields contain data from the public definition, which cannot be edited and are collapsed by default. Copy this definition to create a non-linked copy.</p> <p>The public definitions file was last updated Wed May 25 12:44:19 2011.</p>
Parse rules (22)	
Tasks (1)	
<b>EST_human</b>	<p>Name (?)</p> <input type="text" value="EST_human"/>
Copy	<p>Directory (?)</p> <input type="text" value="/srv/mascot-2.3.02/sequence/EST_human/current"/>
	<p><b>Performance settings</b></p> <p>Threads (?)</p> <input type="text" value="4"/>
	<p><input checked="" type="checkbox"/> Use memory mapping (?)</p> <p><input type="checkbox"/> Lock to memory (?)</p>
	<p><b>FASTA file</b></p> <p>Filename pattern (?)</p> <input type="text" value="EST_human_*.fasta"/>
	<p>Type (?)</p> <p><input type="radio"/> AA</p> <p><input checked="" type="radio"/> NA</p>
	<p>Accession parse rule (?)</p> <input type="text" value="&gt;\(gi  [0-9]*\)"/>
	<p>Description parse rule (?)</p> <input type="text" value="&gt;[^\ ]* \(.*\)"/>
	<p>▶ Reference file</p> <p>▶ Taxonomy data</p> <p>Annotations</p>
Done	

**MASCOT**


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 NCBIInr 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

© 2011 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

## Quantitation of modified peptides

**UNIMOD**

protein modifications for mass spectrometry

[Help](#)

Unimod, View record [ Accession #: 986 ]

[Back to list](#)

Accession #	986	MS Name		Interim Name	Label:13C(6)+Dimethyl
Description	Dimethyl 13C(6) silac label				
Composition	H(4) C(-4) 13C(6)	Monoisotopic	34.051429	Average	34.0091
<b>Specificity Definition 1</b>					
Site	K	Position	Anywhere	Classification	Isotopic label
Comment	SILAC+PTM				
<b>Notes and References</b>					
Source	FindMod	Reference	DIMETH		
Source	Misc. URL	Reference	SILAC introduction		
Source	PubMed PMID	Reference	Stable-isotope dimethyl labeling for quantitative proteomics		
Curator	glick2	Last Modified	2010-05-14 16:19:43		No

[Back to list](#)

Dimethyl: H(4), C(2)

SILAC: C(-6), 13C(6)

**MASCOT** : Mascot Server 2.4

© 2011 Matrix Science

**MATRIX**  
**SCIENCE**

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 C<sub>2</sub>H<sub>4</sub>, while the SILAC label substitutes six <sup>13</sup>C atoms for <sup>12</sup>C. So, the net change is H(4) C(-4) <sup>13</sup>C(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

© 2011 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!