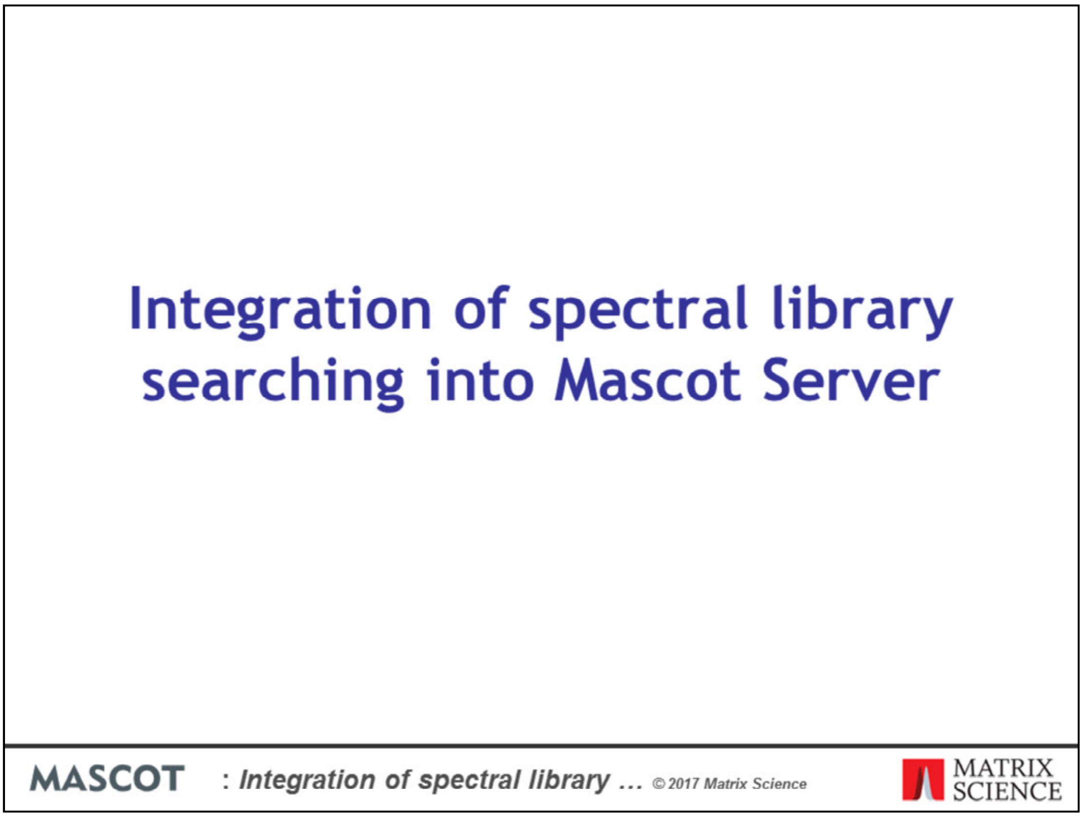
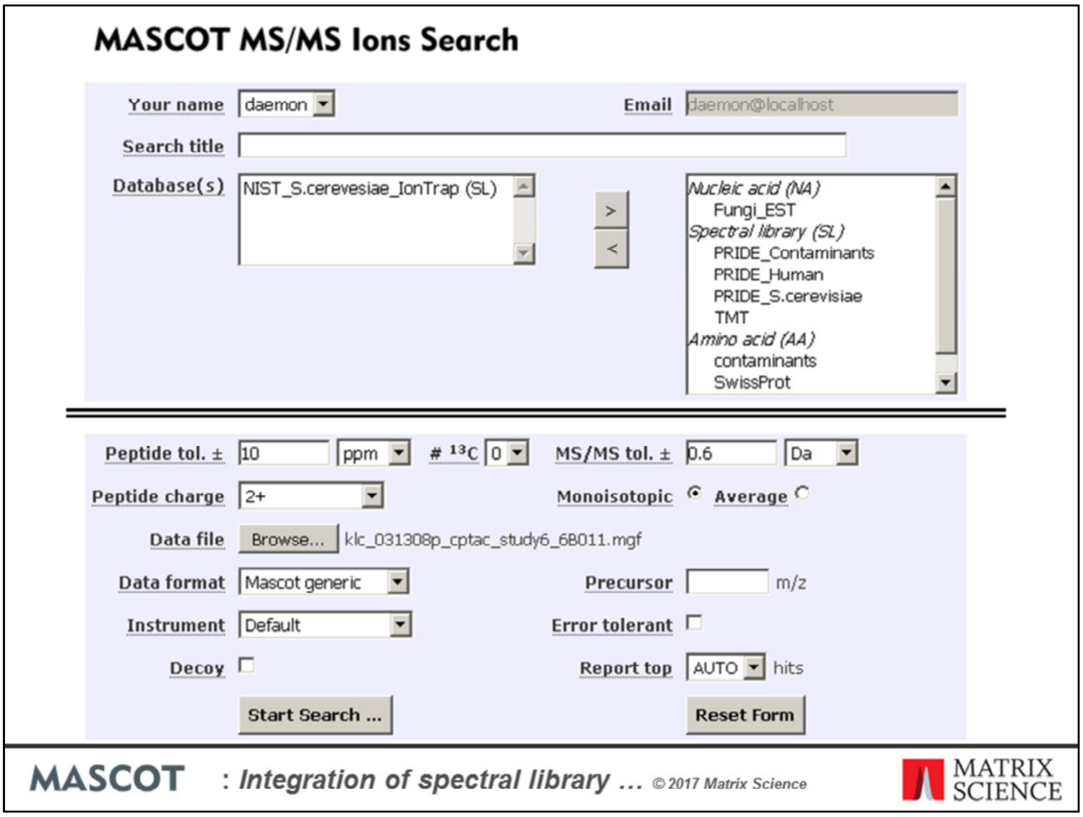


Integration of spectral library searching into Mascot Server



MASCOT MS/MS Ions Search

Your name: daemon ▾ Email: daemon@localhost

Search title:

Database(s): NIST_S.cerevesiae_IonTrap (SL)

Nucleic acid (NA)
- Fungi_EST

Spectral library (SL)
- PRIDE_Contaminants
- PRIDE_Human
- PRIDE_S.cerevisiae
- TMT

Amino acid (AA)
- contaminants
- SwissProt

Peptide tol. ± 10 ppm ▾ # ^{13}C 0 ▾ MS/MS tol. ± 0.6 Da ▾

Peptide charge 2+ ▾ Monoisotopic ● Average ○

Data file Browse... klc_031308p_cptac_study6_6B011.mgf

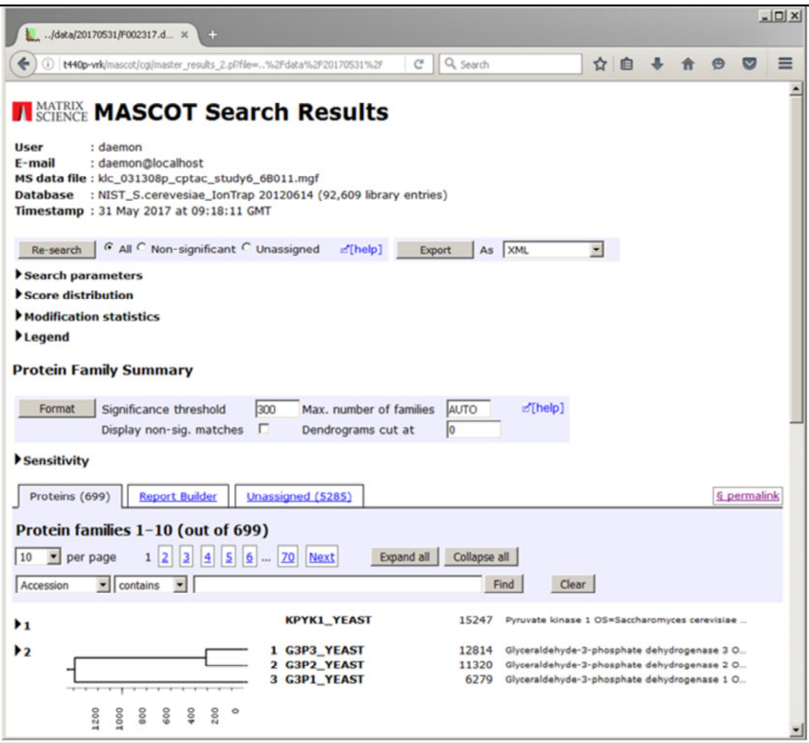
Data format Mascot generic ▾ Precursor m/z

Instrument Default ▾ Error tolerant □

Decoy □ Report top AUTO ▾ hits

In Mascot Server 2.6, we added the capability to search spectral libraries using MSPepSearch from Steven Stein's group at NIST. When submitting a search, any combination of amino acid Fasta files, nucleic acid Fasta files, and spectral library files can be selected. Here, we perform a simple search of some data from CPTAC study 6 against a NIST yeast library

Most search parameters – modifications, enzyme, missed cleavages, taxonomy, and instrument – simply don't apply to a library search. All that matters is how well the experimental spectrum matches the one in the library. The main exceptions are the precursor and fragment mass tolerances.


.../data/20170531/F002317.d...
t440p-vrk/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20170531%2F
Search
Re-search
Export
Format
Expand all
Collapse all
Find
Clear
1200
1000
800
600
400
200
0

MASCOT Search Results

User : daemon
E-mail : daemon@localhost
MS data file : klc_031308p_cptac_study6_6B011.mgf
Database : NIST_S.cerevesiae_IonTrap 20120614 (92,609 library entries)
Timestamp : 31 May 2017 at 09:18:11 GMT

● All ○ Non-significant ○ Unassigned [help] As XML

- **Search parameters**
- **Score distribution**
- **Modification statistics**
- **Legend**

Protein Family Summary

Significance threshold 300 Max. number of families AUTO [help]
Display non-sig. matches □ Dendrograms cut at 0

- **Sensitivity**

Proteins (699) | Report Builder | Unassigned (5285) | § permalink

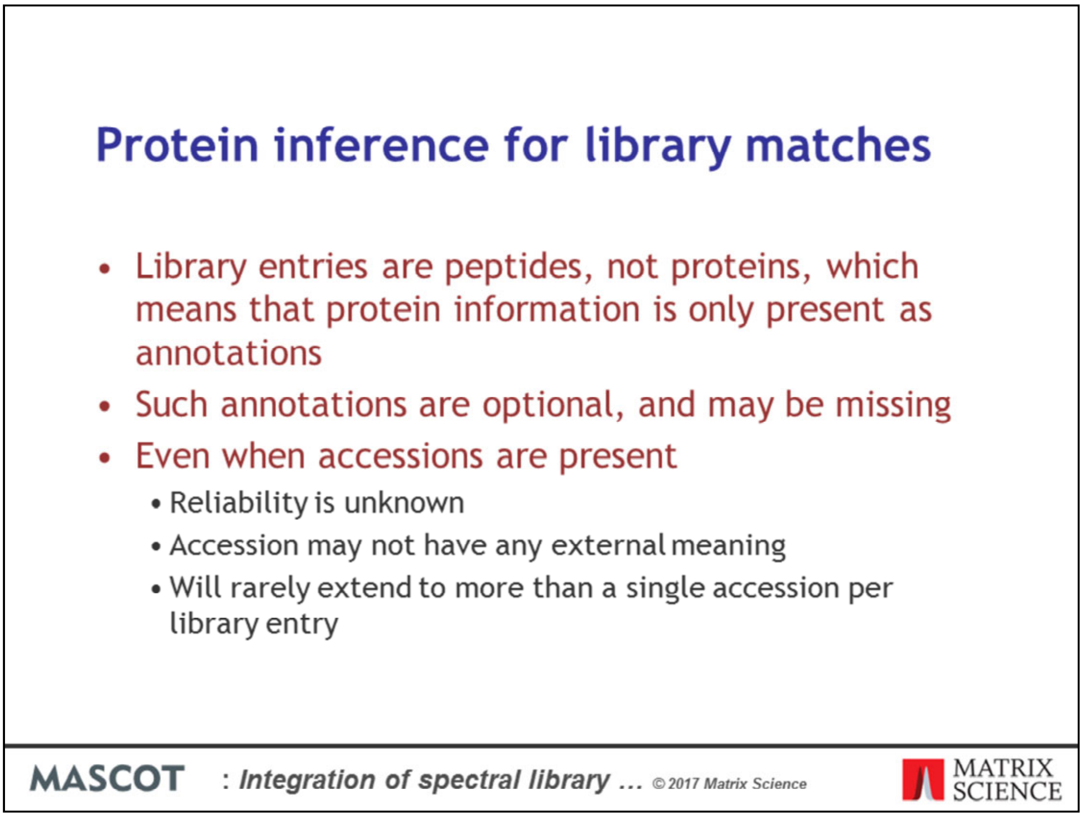
Protein families 1–10 (out of 699)

10 per page 1 2 3 4 5 6 ... 70 Next

Accession contains

1		KPYK1_YEAST	15247	Pyruvate kinase 1 OS=Saccharomyces cerevisiae ...
2	1	G3P3_YEAST	12814	Glyceraldehyde-3-phosphate dehydrogenase 3 O...
	2	G3P2_YEAST	11320	Glyceraldehyde-3-phosphate dehydrogenase 2 O...
	3	G3P1_YEAST	6279	Glyceraldehyde-3-phosphate dehydrogenase 1 O...

On completion of the search, the matches are reported in a protein family summary. In order to generate such a report, we need reliable and accurate protein inference



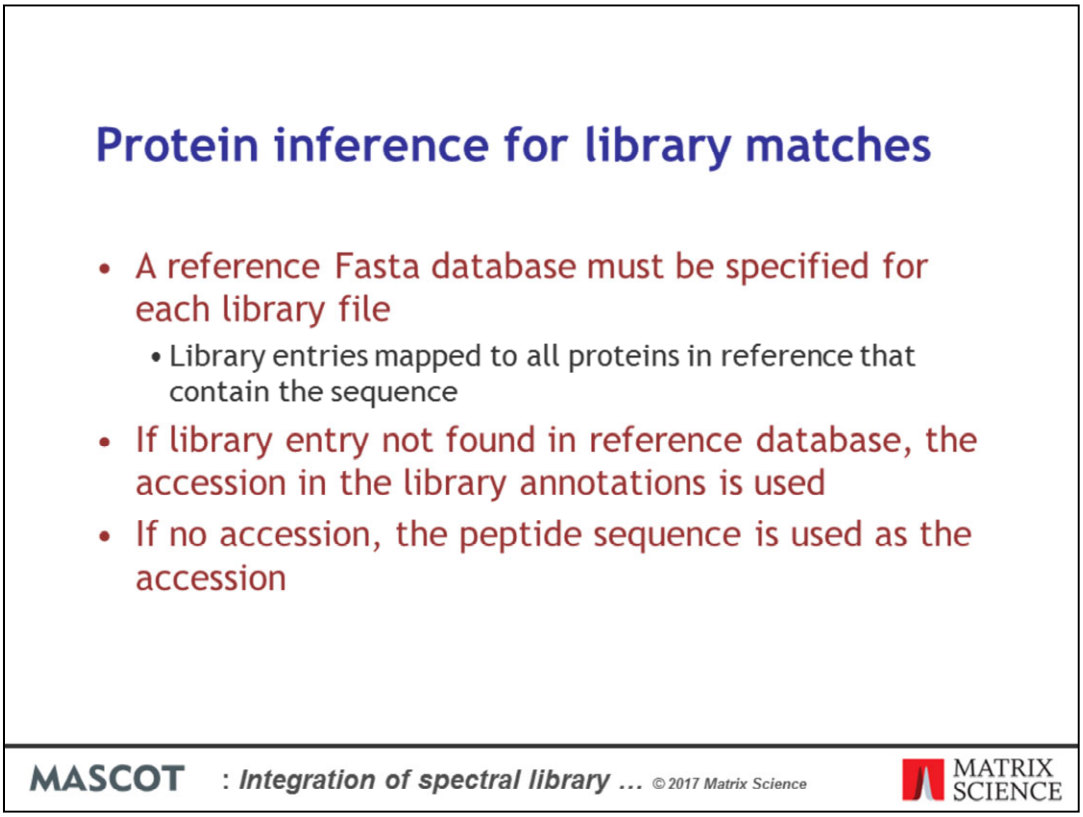
Protein inference for library matches

- Library entries are peptides, not proteins, which means that protein information is only present as annotations
- Such annotations are optional, and may be missing
- Even when accessions are present
  - Reliability is unknown
  - Accession may not have any external meaning
  - Will rarely extend to more than a single accession per library entry

MASCOT : *Integration of spectral library …* © 2017 Matrix Science MATRIX SCIENCE

The are some difficulties associated with Protein inference for library matches. First of all, library entries are peptides, not proteins, which means that protein information is only ever present as annotations. Such annotations are optional, and may be missing, as in the case of most PRIDE libraries.

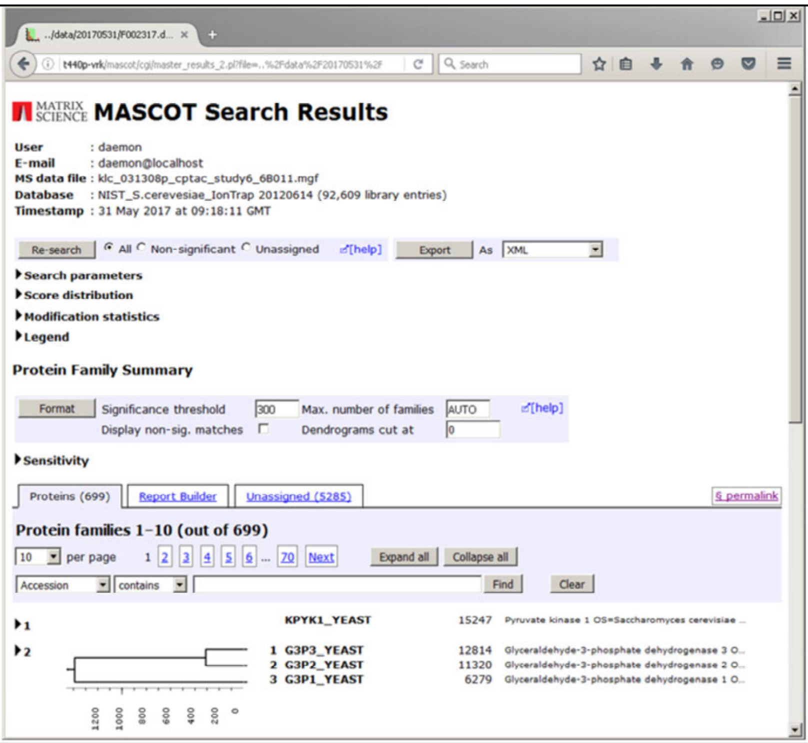
Even when present, the reliability is unknown. The accession could be a meaningless number or string. And, I've never seen a library with more than a single accession per library entry, so protein inference will be inaccurate for shared peptides.



Protein inference for library matches

- A reference Fasta database must be specified for each library file
  - Library entries mapped to all proteins in reference that contain the sequence
- If library entry not found in reference database, the accession in the library annotations is used
- If no accession, the peptide sequence is used as the accession

Our solution is to require a reference Fasta database to be assigned to each library file when it is added to the system. The default is SwissProt, with an appropriate taxonomy filter, but any online Fasta can be chosen. This allows Mascot to map most of the library peptides to accessions in the reference database. This mapping is done at the sequence level, with no constraints from enzyme specificity. If a library entry has a novel sequence, not found in the reference database, the accession in the library annotations is used. If there is no accession, the peptide sequence is treated as the accession, so that duplicate matches to the same peptide can be grouped, if nothing else.



MASCOT Search Results

User : daemon
E-mail : daemon@localhost
MS data file : klc_031308p_cptac_study6_6B011.mgf
Database : NIST_S.cerevesiae_IonTrap 20120614 (92,609 library entries)
Timestamp : 31 May 2017 at 09:18:11 GMT

● All ○ Non-significant ○ Unassigned   As XML

Protein Family Summary

Significance threshold 300   Max. number of families AUTO
Display non-sig. matches □   Dendrograms cut at 0

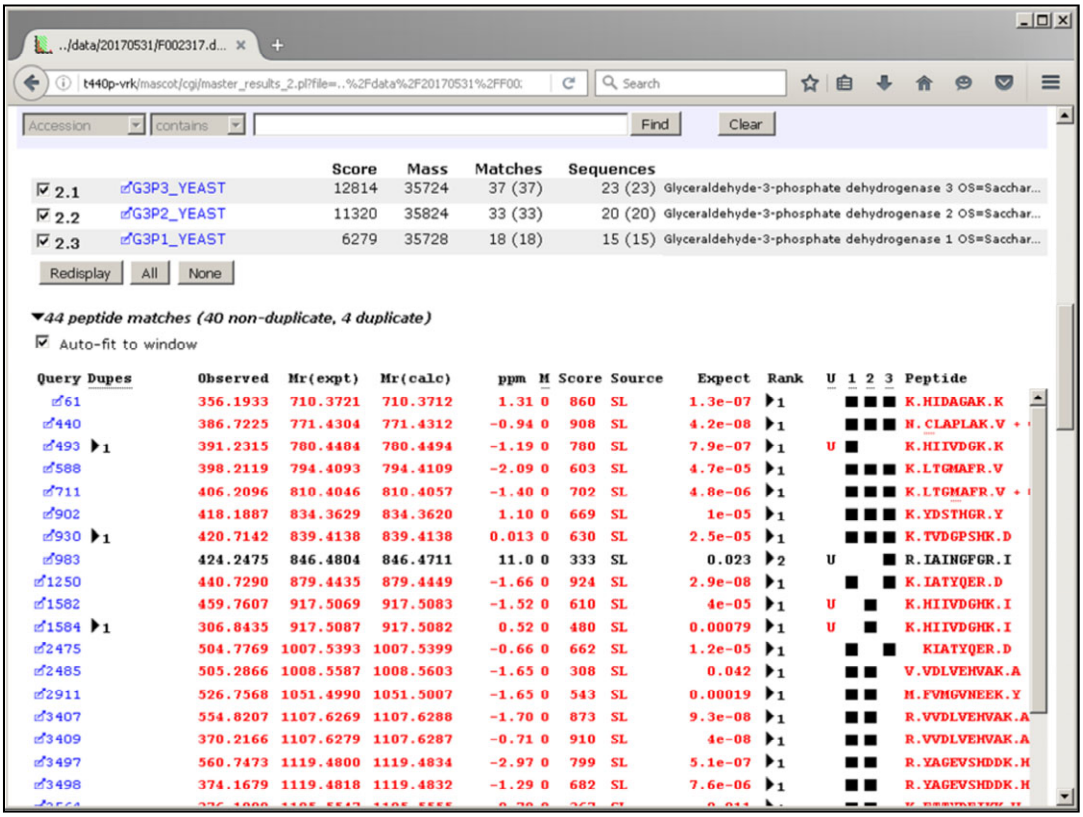
Protein families 1–10 (out of 699)

10 per page   1 2 3 4 5 6 ... 70 Next

Accession   contains

1		KPYK1_YEAST	15247	Pyruvate kinase 1 OS=Saccharomyces cerevisiae ...
2	1	G3P3_YEAST	12814	Glyceraldehyde-3-phosphate dehydrogenase 3 O...
	2	G3P2_YEAST	11320	Glyceraldehyde-3-phosphate dehydrogenase 2 O...
	3	G3P1_YEAST	6279	Glyceraldehyde-3-phosphate dehydrogenase 1 O...

Here’s the library search report again. Protein inference allows us to create a report for library matches that is near identical to a report for Fasta matches.



../data/20170531/F002317.d...
t440p-vrk/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20170531%2FF00:
Search
Accession
contains
Find
Clear

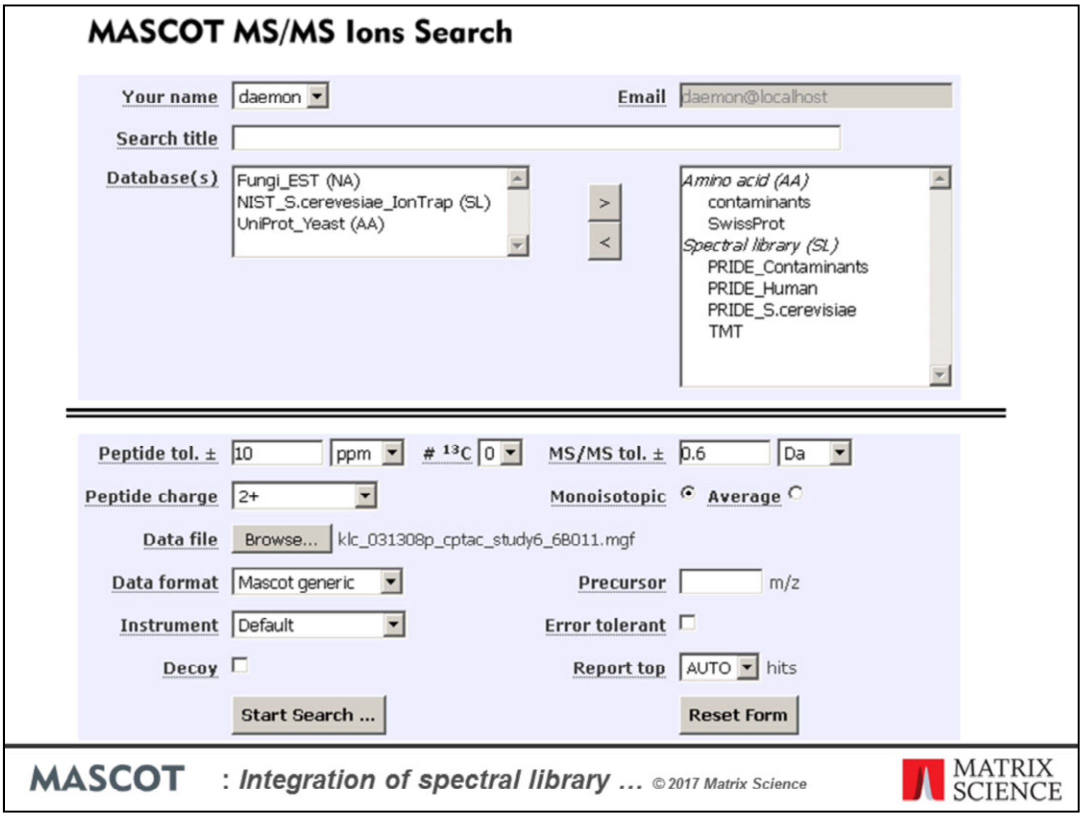
		Score	Mass	Matches	Sequences	
☑ 2.1	G3P3_YEAST	12814	35724	37 (37)	23 (23)	Glyceraldehyde-3-phosphate dehydrogenase 3 OS=Sacchar...
☑ 2.2	G3P2_YEAST	11320	35824	33 (33)	20 (20)	Glyceraldehyde-3-phosphate dehydrogenase 2 OS=Sacchar...
☑ 2.3	G3P1_YEAST	6279	35728	18 (18)	15 (15)	Glyceraldehyde-3-phosphate dehydrogenase 1 OS=Sacchar...

▼44 peptide matches (40 non-duplicate, 4 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Source	Expect	Rank	U	1	2	3	Peptide
61		356.1933	710.3721	710.3712	1.31	0	860	SL	1.3e-07	▶1		■	■	■	K.HIDAGAK.K
440		386.7225	771.4304	771.4312	-0.94	0	908	SL	4.2e-08	▶1		■	■	■	N.CLAPLAK.V +
493	▶1	391.2315	780.4484	780.4494	-1.19	0	780	SL	7.9e-07	▶1	U	■			K.HIIVDGK.K
588		398.2119	794.4093	794.4109	-2.09	0	603	SL	4.7e-05	▶1		■	■	■	K.LTGMAFR.V
711		406.2096	810.4046	810.4057	-1.40	0	702	SL	4.8e-06	▶1		■	■	■	K.LTGMAFR.V +
902		418.1887	834.3629	834.3620	1.10	0	669	SL	1e-05	▶1		■	■	■	K.YDSTHGR.Y
930	▶1	420.7142	839.4138	839.4138	0.013	0	630	SL	2.5e-05	▶1		■	■	■	K.TVDGPSHK.D
983		424.2475	846.4804	846.4711	11.0	0	333	SL	0.023	▶2	U			■	R.IAINGFGR.I
1250		440.7290	879.4435	879.4449	-1.66	0	924	SL	2.9e-08	▶1		■		■	K.IATYQER.D
1582		459.7607	917.5069	917.5083	-1.52	0	610	SL	4e-05	▶1	U		■		K.HIIVDGHK.I
1584	▶1	306.8435	917.5087	917.5082	0.52	0	480	SL	0.00079	▶1	U		■		K.HIIVDGHK.I
2475		504.7769	1007.5393	1007.5399	-0.66	0	662	SL	1.2e-05	▶1		■		■	KIATYQER.D
2485		505.2866	1008.5587	1008.5603	-1.65	0	308	SL	0.042	▶1		■	■		V.VDLVEHVAK.A
2911		526.7568	1051.4990	1051.5007	-1.65	0	543	SL	0.00019	▶1		■	■		M.FVMGVNEEK.Y
3407		554.8207	1107.6269	1107.6288	-1.70	0	873	SL	9.3e-08	▶1		■	■		R.VVDLVEHVAK.A
3409		370.2166	1107.6279	1107.6287	-0.71	0	910	SL	4e-08	▶1		■	■		R.VVDLVEHVAK.A
3497		560.7473	1119.4800	1119.4834	-2.97	0	799	SL	5.1e-07	▶1		■	■		R.YAGEVSHDDK.H
3498		374.1679	1119.4818	1119.4832	-1.29	0	682	SL	7.6e-06	▶1		■	■		R.YAGEVSHDDK.H

If only libraries are searched, MSPepSearch scores are converted to arbitrary expect values. A score of 300 becomes an expect value of 0.05 and the maximum score of 1000 becomes an expect of 5E-9



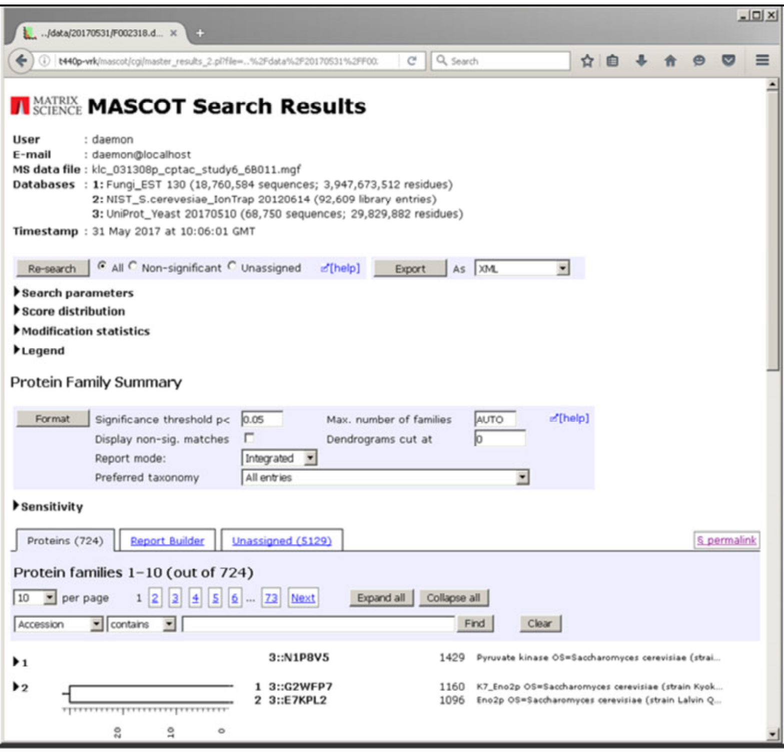
MASCOT MS/MS Ions Search

Your name	daemon	Email	daemon@localhost
Search title			
Database(s)	Fungi_EST (NA) NIST_S.cerevesiae_IonTrap (SL) UniProt_Yeast (AA)	> <	*Amino acid (AA)* contaminants SwissProt *Spectral library (SL)* PRIDE_Contaminants PRIDE_Human PRIDE_S.cerevisiae TMT
Peptide tol. ±	10 ppm # ^{13}C 0	MS/MS tol. ±	0.6 Da
Peptide charge	2+	Monoisotopic	● Average ○
Data file	Browse... klc_031308p_cptac_study6_6B011.mgf		
Data format	Mascot generic	Precursor	m/z
Instrument	Default	Error tolerant	□
Decoy	□	Report top	AUTO hits
	Start Search ...		Reset Form

MASCOT : *Integration of spectral library …* © 2017 Matrix Science

MATRIX SCIENCE

Coming back to the search form, you can search any mixture of amino acid and nucleic acid databases and spectral libraries.



../data/20170531/F002318.d...
t440p-vrk/mascot/cgi/master_results_2.pl?file=...%2Fdata%2F20170531%2FF00
Search
MATRIX SCIENCE

MASCOT Search Results

```
User          : daemon
E-mail        : daemon@localhost
MS data file  : klc_031308p_cptac_study6_6B011.mgf
Databases     : 1: Fungi_EST 130 (18,760,584 sequences; 3,947,673,512 residues)
                2: NIST_S.cerevesiae_IonTrap 20120614 (92,609 library entries)
                3: UniProt_Yeast 20170510 (68,750 sequences; 29,829,882 residues)
Timestamp     : 31 May 2017 at 10:06:01 GMT
```

Re-search
● All ○ Non-significant ○ Unassigned
[help]
Export
As XML
Search parameters
Score distribution
Modification statistics
Legend

Protein Family Summary

Significance threshold p< 0.05
Max. number of families AUTO
Display non-sig. matches □
Dendrograms cut at 0
Report mode: Integrated
Preferred taxonomy All entries

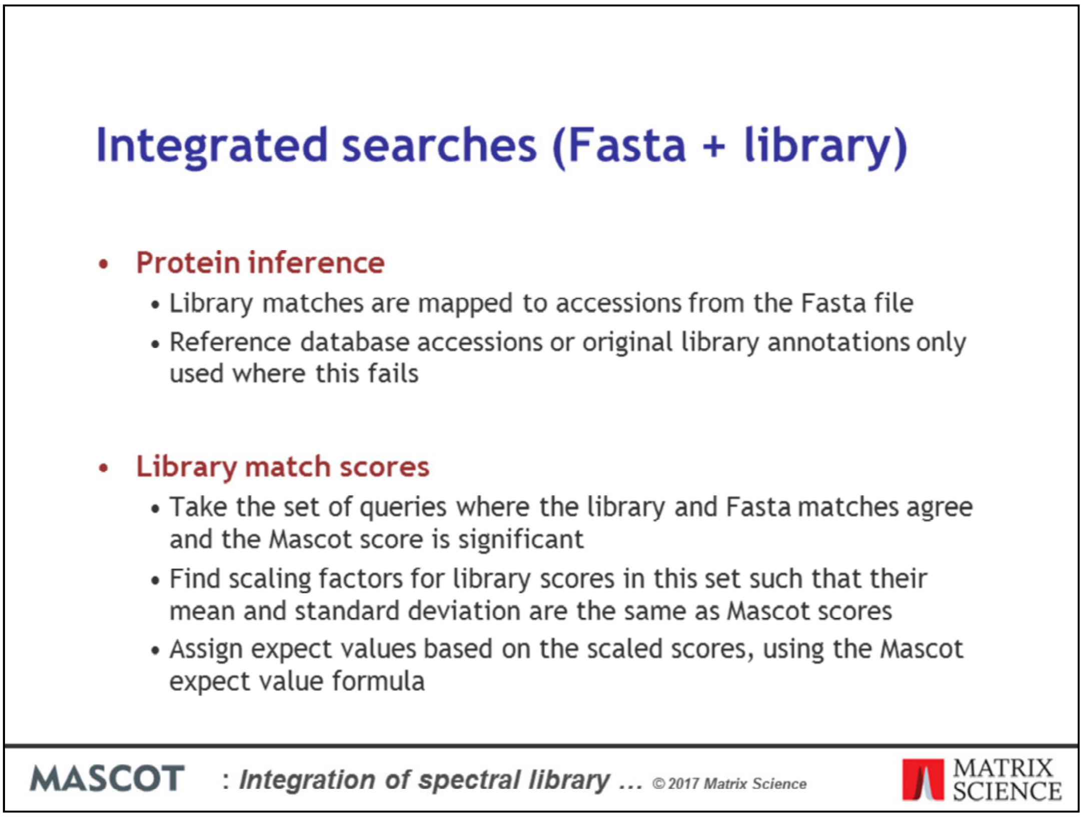
Protein families 1–10 (out of 724)

```
1          3::N1P8V5   1429  Pyruvate kinase OS=Saccharomyces cerevisiae (strai...
2        1 3::G2WFP7   1160  K7_Eno2p OS=Saccharomyces cerevisiae (strain Kyok...
         2 3::E7KPL2   1096  Eno2p OS=Saccharomyces cerevisiae (strain Lalvin Q...
```

MASCOT : *Integration of spectral library …* © 2017 Matrix Science

MATRIX SCIENCE

This is a report for a search of both library and Fasta files. We can see all three databases listed at the top of the result report, and each is assigned an index so that we know where each accession comes from. The top hit has an index 3 which corresponds to the UniProt proteome. There are two important differences between this ‘integrated’ report and a library-only report.



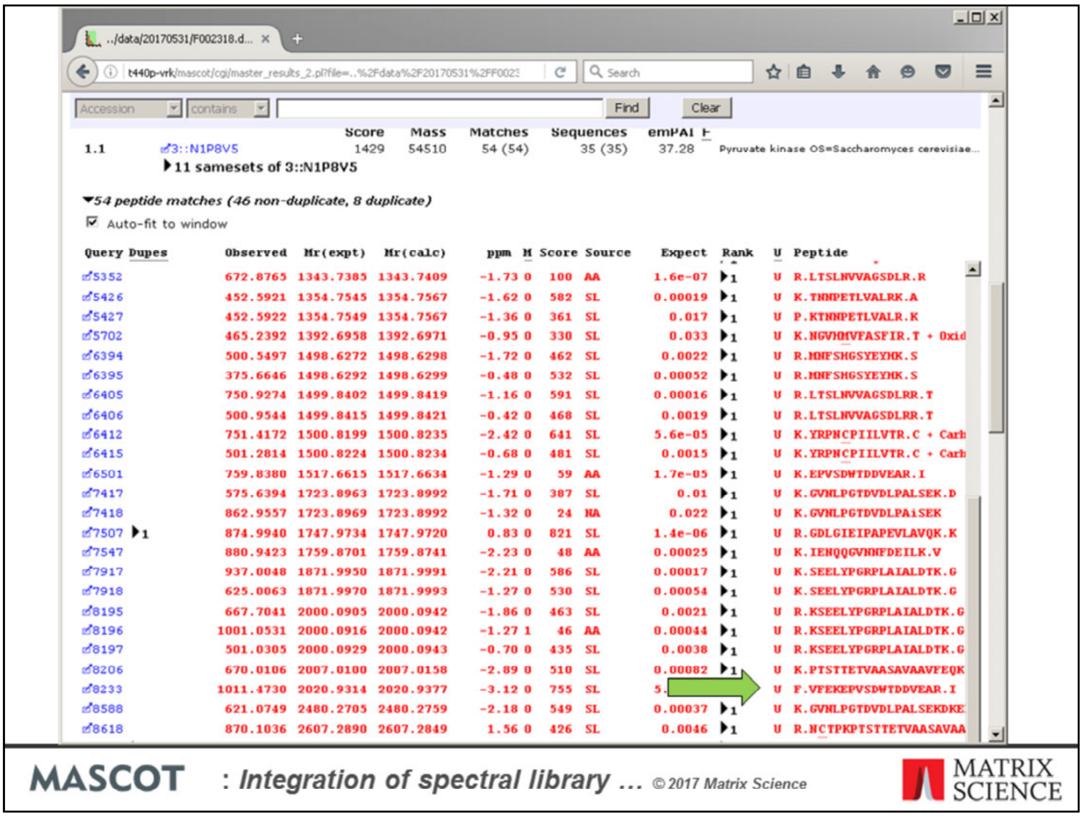
Integrated searches (Fasta + library)

- **Protein inference**
  - Library matches are mapped to accessions from the Fasta file
  - Reference database accessions or original library annotations only used where this fails
- **Library match scores**
  - Take the set of queries where the library and Fasta matches agree and the Mascot score is significant
  - Find scaling factors for library scores in this set such that their mean and standard deviation are the same as Mascot scores
  - Assign expect values based on the scaled scores, using the Mascot expect value formula

MASCOT : *Integration of spectral library ...* © 2017 Matrix Science MATRIX SCIENCE

For protein inference, if the peptide sequence can be mapped to one of the Fasta files being searched, this becomes the preferred accession. The accession from the reference database is only used when this fails.

In an integrated search, we can use the Fasta matches to create a simple empirical estimate of library score significance. This is achieved by calibrating library scores based on the set of queries where the library and Fasta searches return the same match and the Mascot score is significant. The shapes of the library and Mascot score distributions in this set are similar and they often have a fairly high correlation. Next, scale these library scores so that they have the same mean and standard deviation as Mascot scores. This produces values on the same scale as Mascot scores. We can now assign expect values to library matches using the same expression as for Mascot matches


../data/20170531/F002318.d...
t440p-vrk/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20170531%2FF0023
Search
Accession
contains
Find
Clear

		Score	Mass	Matches	Sequences	emPAI	
1.1	3::N1P8V5	1429	54510	54 (54)	35 (35)	37.28	Pyruvate kinase OS=Saccharomyces cerevisiae...

11 samesets of 3::N1P8V5

54 peptide matches (46 non-duplicate, 8 duplicate)

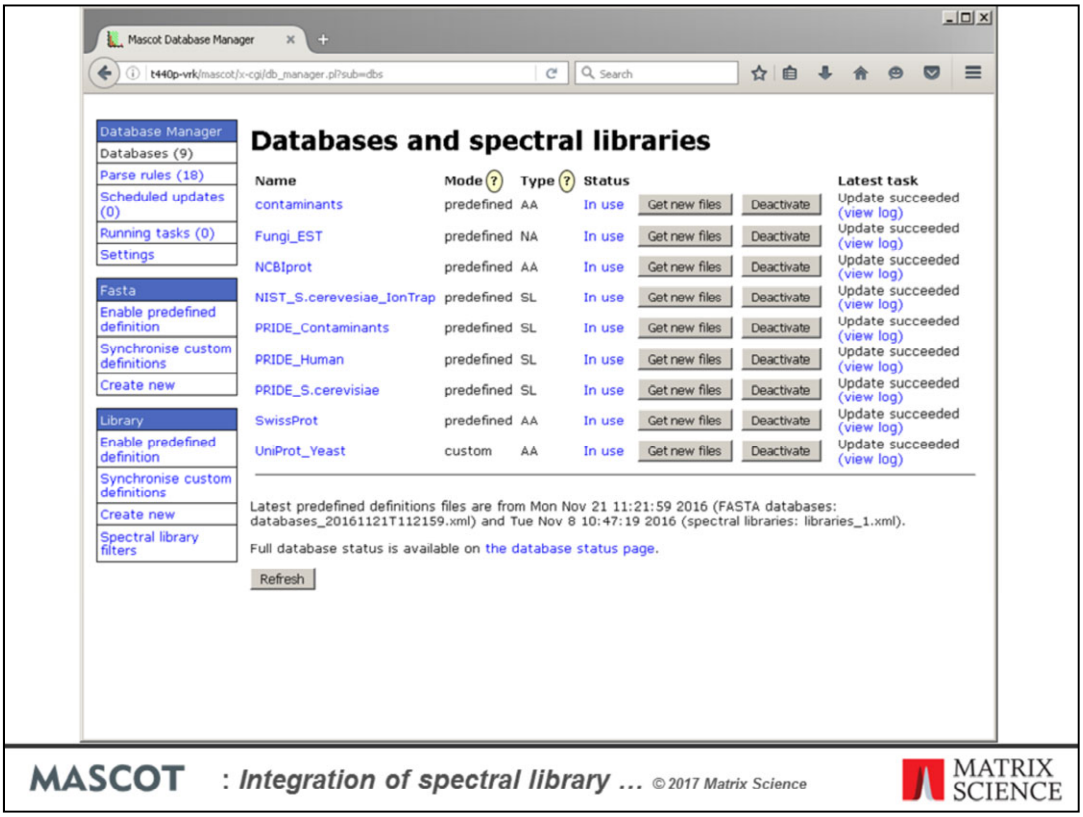
☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Source	Expect	Rank	U	Peptide
5352		672.8765	1343.7385	1343.7409	-1.73	0	100	AA	1.6e-07	1	U	R.LTSLNVVAGSDLR.R
5426		452.5921	1354.7545	1354.7567	-1.62	0	582	SL	0.00019	1	U	K.TNNPETLVALRK.A
5427		452.5922	1354.7549	1354.7567	-1.36	0	361	SL	0.017	1	U	P.KTNNPETLVALR.K
5702		465.2392	1392.6958	1392.6971	-0.95	0	330	SL	0.033	1	U	K.NGVHMVFASFIR.T + Oxid
6394		500.5497	1498.6272	1498.6298	-1.72	0	462	SL	0.0022	1	U	R.MNFSHGSYEYHK.S
6395		375.6646	1498.6292	1498.6299	-0.48	0	532	SL	0.00052	1	U	R.MNFSHGSYEYHK.S
6405		750.9274	1499.8402	1499.8419	-1.16	0	591	SL	0.00016	1	U	R.LTSLNVVAGSDLRR.T
6406		500.9544	1499.8415	1499.8421	-0.42	0	468	SL	0.0019	1	U	R.LTSLNVVAGSDLRR.T
6412		751.4172	1500.8199	1500.8235	-2.42	0	641	SL	5.6e-05	1	U	K.YRPNCPIILVTR.C + Carb
6415		501.2814	1500.8224	1500.8234	-0.68	0	481	SL	0.0015	1	U	K.YRPNCPIILVTR.C + Carb
6501		759.8380	1517.6615	1517.6634	-1.29	0	59	AA	1.7e-05	1	U	K.EPVSDWTDDVEAR.I
7417		575.6394	1723.8963	1723.8992	-1.71	0	387	SL	0.01	1	U	K.GVNLPGTDVDLPALSEK.D
7418		862.9557	1723.8969	1723.8992	-1.32	0	24	NA	0.022	1	U	K.GVNLPGTDVDLPAiSEK
7507	1	874.9940	1747.9734	1747.9720	0.83	0	821	SL	1.4e-06	1	U	R.GDLGIEIPAPEVLAVQK.K
7547		880.9423	1759.8701	1759.8741	-2.23	0	48	AA	0.00025	1	U	K.IENQQGVNNFDEILK.V
7917		937.0048	1871.9950	1871.9991	-2.21	0	586	SL	0.00017	1	U	K.SEELYPGRPLAIALDTK.G
7918		625.0063	1871.9970	1871.9993	-1.27	0	530	SL	0.00054	1	U	K.SEELYPGRPLAIALDTK.G
8195		667.7041	2000.0905	2000.0942	-1.86	0	463	SL	0.0021	1	U	R.KSEELYPGRPLAIALDTK.G
8196		1001.0531	2000.0916	2000.0942	-1.27	1	46	AA	0.00044	1	U	R.KSEELYPGRPLAIALDTK.G
8197		501.0305	2000.0929	2000.0943	-0.70	0	435	SL	0.0038	1	U	R.KSEELYPGRPLAIALDTK.G
8206		670.0106	2007.0100	2007.0158	-2.89	0	510	SL	0.00082	1	U	K.PTSTTETVAASAVAAVFEQK
8233		1011.4730	2020.9314	2020.9377	-3.12	0	755	SL	5.		U	F.VFEKEPVSDWTDDVEAR.I
8588		621.0749	2480.2705	2480.2759	-2.18	0	549	SL	0.00037	1	U	K.GVNLPGTDVDLPALSEKDKE
8618		870.1036	2607.2890	2607.2849	1.56	0	426	SL	0.0046	1	U	R.NCTPKPTSTTETVAASAVAA

MASCOT : *Integration of spectral library …* © 2017 Matrix Science

MATRIX SCIENCE

Here, the top hit has been expanded. You can see that the top ranking PSMs come from both library and Fasta. In most cases, the same match is found in two or all three databases, and the listed match is the one with the lowest expect value. An exception can be seen here for query 8233. This peptide is non-specific at the amino terminus, and is only found in the library. It will not be matched in the Fasta files because the enzyme for the search was strict trypsin.



Databases and spectral libraries

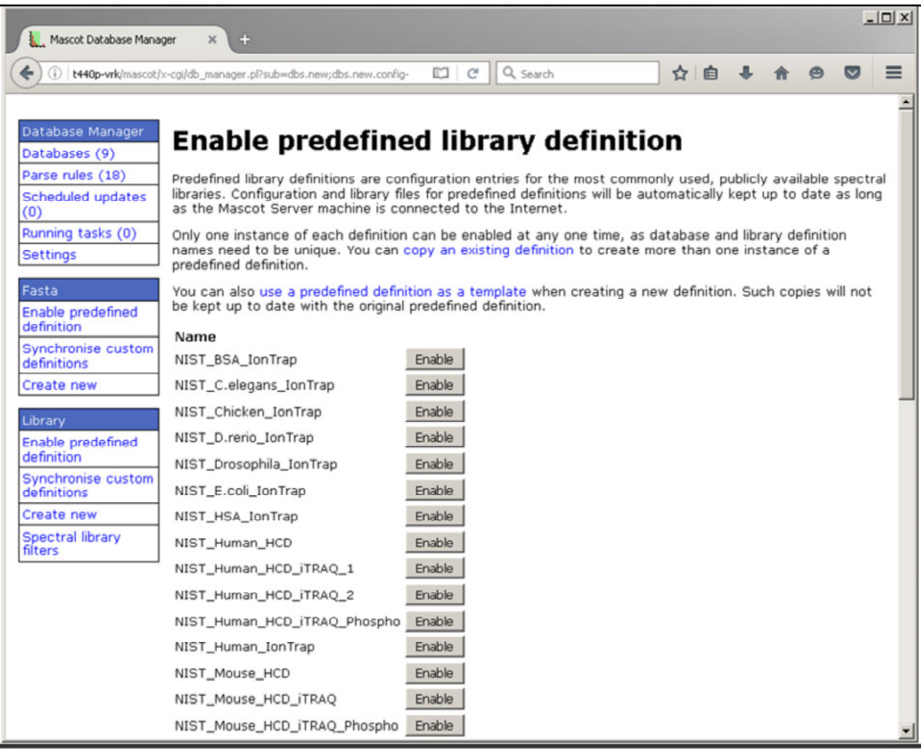
Name	Mode ?	Type ?	Status			Latest task
contaminants	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
Fungi_EST	predefined	NA	In use	Get new files	Deactivate	Update succeeded (view log)
NCBIprot	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
NIST_S.cerevesiae_IonTrap	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
PRIDE_Contaminants	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
PRIDE_Human	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
PRIDE_S.cerevisiae	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
SwissProt	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
UniProt_Yeast	custom	AA	In use	Get new files	Deactivate	Update succeeded (view log)

Latest predefined definitions files are from Mon Nov 21 11:21:59 2016 (FASTA databases: databases_20161121T112159.xml) and Tue Nov 8 10:47:19 2016 (spectral libraries: libraries_1.xml).

Full database status is available on the database status page.

MASCOT : Integration of spectral library ... © 2017 Matrix Science
MATRIX SCIENCE

Let’s turn our attention to administration aspects. Library files in MSP format are handled in Database Manager much the same as Sequence databases in Fasta format. This slide shows the top level screen of Database Manager, with a mixture of Fasta and library databases configured for searching. The ‘Type’ column shows which are AA or NA Fasta and which are spectral library. Most have ‘predefined’ configuration settings – that is, Matrix Science maintains a master file of configuration settings that is downloaded by Database Manager.



Mascot Database Manager
t440p-vrk/mascot/x-cgi/db_manager.pl?sub=dbs.new;dbs.new.config-
Search

Database Manager
Databases (9)
Parse rules (18)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Enable predefined library definition

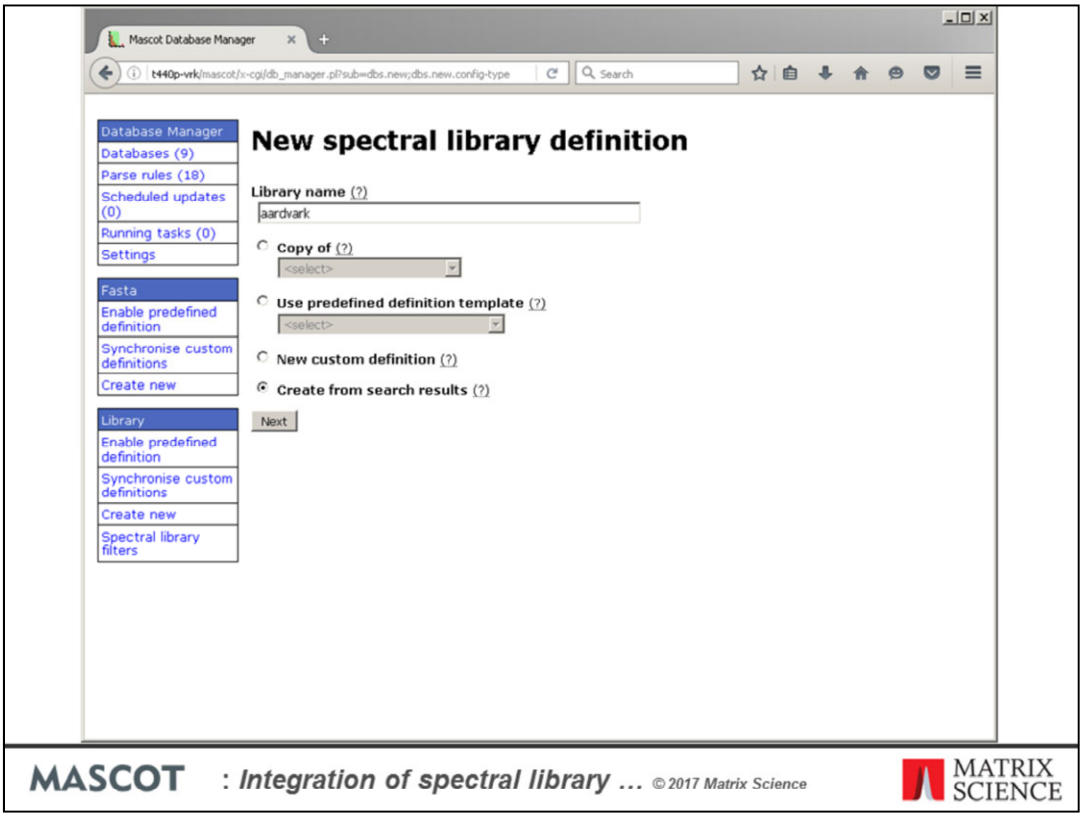
Predefined library definitions are configuration entries for the most commonly used, publicly available spectral libraries. Configuration and library files for predefined definitions will be automatically kept up to date as long as the Mascot Server machine is connected to the Internet.

Only one instance of each definition can be enabled at any one time, as database and library definition names need to be unique. You can copy an existing definition to create more than one instance of a predefined definition.

You can also use a predefined definition as a template when creating a new definition. Such copies will not be kept up to date with the original predefined definition.

Name	
NIST_BSA_IonTrap	Enable
NIST_C.elegans_IonTrap	Enable
NIST_Chicken_IonTrap	Enable
NIST_D.rerio_IonTrap	Enable
NIST_Drosophila_IonTrap	Enable
NIST_E.coli_IonTrap	Enable
NIST_HSA_IonTrap	Enable
NIST_Human_HCD	Enable
NIST_Human_HCD_iTRAQ_1	Enable
NIST_Human_HCD_iTRAQ_2	Enable
NIST_Human_HCD_iTRAQ_Phospho	Enable
NIST_Human_IonTrap	Enable
NIST_Mouse_HCD	Enable
NIST_Mouse_HCD_iTRAQ	Enable
NIST_Mouse_HCD_iTRAQ_Phospho	Enable

To enable a predefined library is a matter of a few mouse clicks



Mascot Database Manager
t440p-vrk/mascot/x-cgi/db_manager.pl?sub=dbs.new;dbs.new.config-type
Search
Database Manager
Databases (9)
Parse rules (18)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

New spectral library definition

Library name (?)
aardvark

○ Copy of (?)
<select>

○ Use predefined definition template (?)
<select>

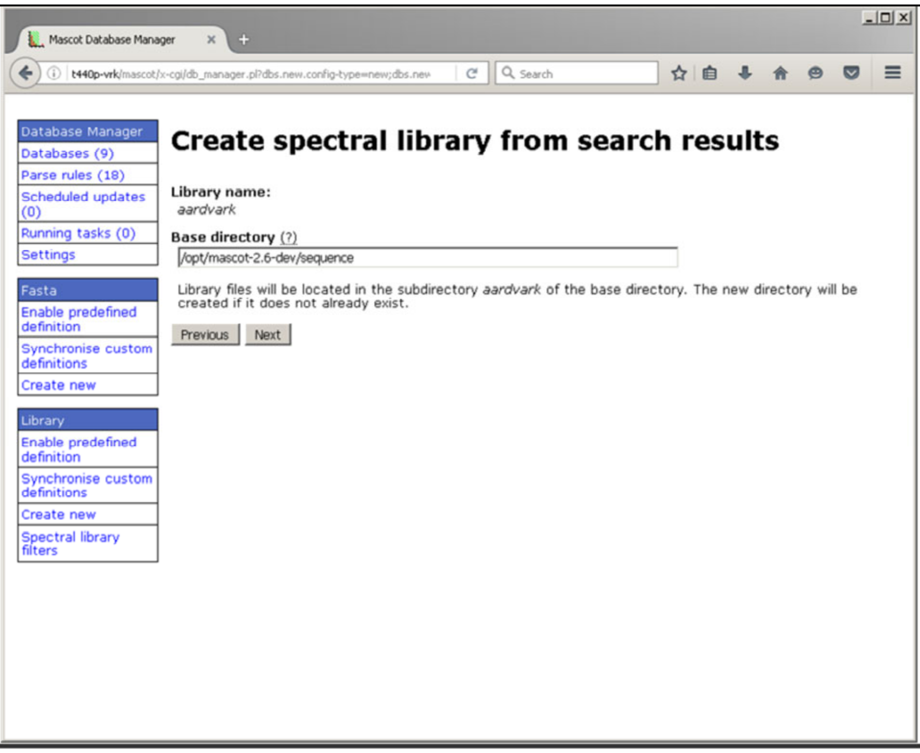
○ New custom definition (?)

● Create from search results (?)

MASCOT : *Integration of spectral library …* © 2017 Matrix Science

MATRIX SCIENCE

If the library you want to search is not on the predefined list, you use the 'Create New' Wizard to configure it as a custom database. A particularly interesting case is if you want to create your own library from Mascot search results. This is easily accomplished, as illustrated in the next few slides. Suppose that we are working on aardvark, and want to make a custom library for the aardvark proteome. We choose a name and select 'Create from search results'



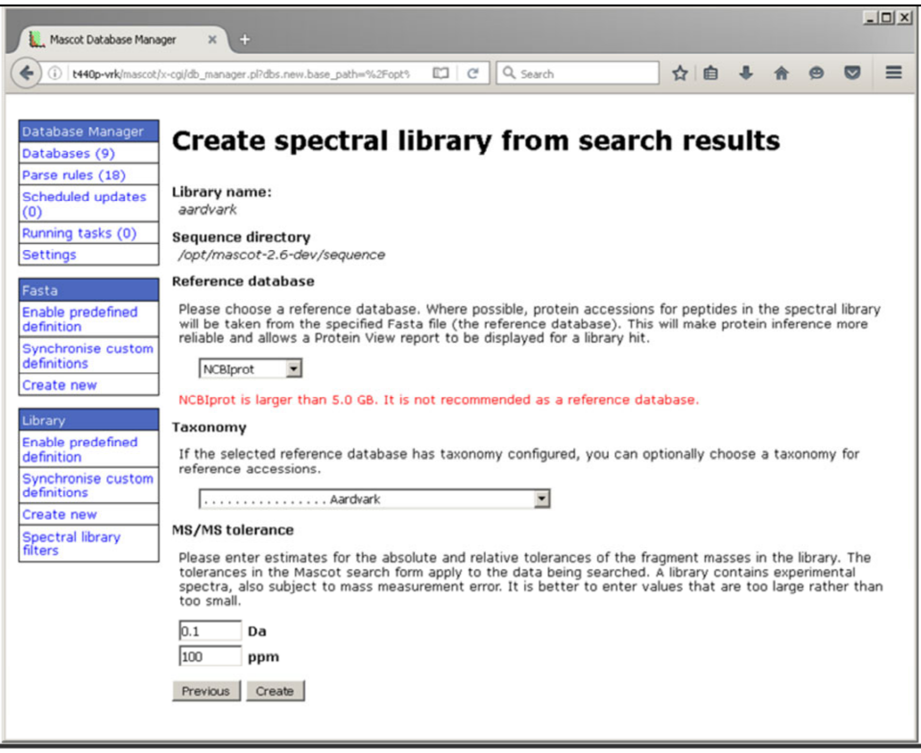
Create spectral library from search results

Library name:
aardvark

Base directory (?)
/opt/mascot-2.6-dev/sequence

Library files will be located in the subdirectory *aardvark* of the base directory. The new directory will be created if it does not already exist.

The next screen just gives an opportunity to change the default location for the files


Mascot Database Manager
t440p-vrk/mascot/x-cgi/db_manager.pl?dbs.new.base_path=%2Fopt%
Search
Database Manager
Databases (9)
Parse rules (18)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Create spectral library from search results

Library name:
aardvark

Sequence directory
/opt/mascot-2.6-dev/sequence

Reference database

Please choose a reference database. Where possible, protein accessions for peptides in the spectral library will be taken from the specified Fasta file (the reference database). This will make protein inference more reliable and allows a Protein View report to be displayed for a library hit.

NCBIprot

NCBIprot is larger than 5.0 GB. It is not recommended as a reference database.

Taxonomy

If the selected reference database has taxonomy configured, you can optionally choose a taxonomy for reference accessions.

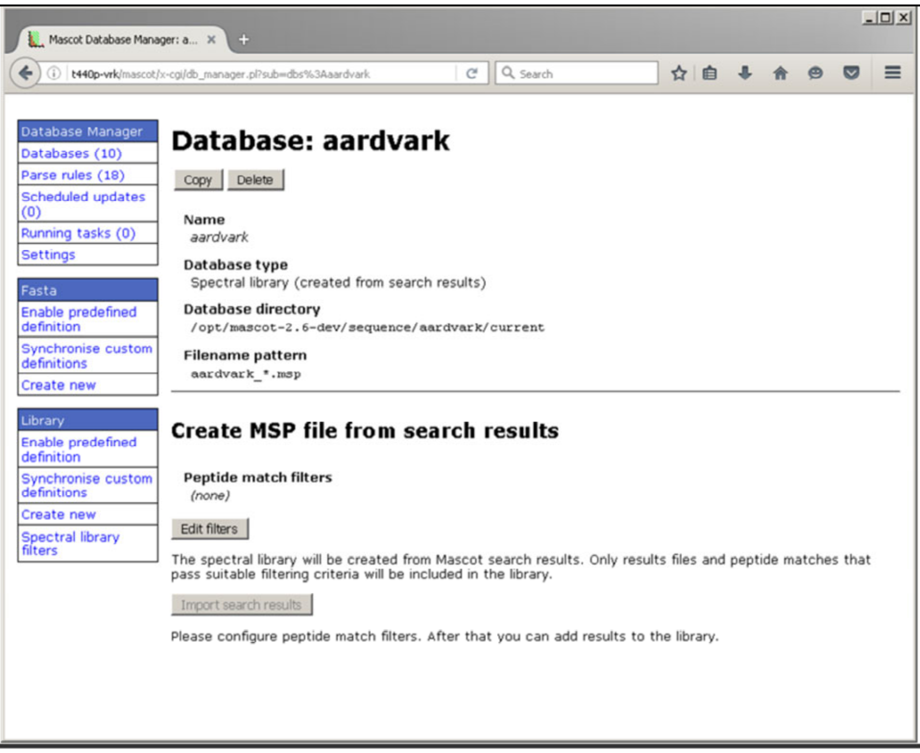
. . . . Aardvark

MS/MS tolerance

Please enter estimates for the absolute and relative tolerances of the fragment masses in the library. The tolerances in the Mascot search form apply to the data being searched. A library contains experimental spectra, also subject to mass measurement error. It is better to enter values that are too large rather than too small.

0.1 **Da**
100 **ppm**

The reference database is used to assign protein accessions to the library entries. Normally, you wouldn’t choose NCBIprot because it is such a large and redundant database. But, since SwissProt only contains 10 aardvark entries, we don’t have much choice. We must also provide an estimate of suitable MS/MS tolerances for the library contents. If the search results come from multiple instruments, you need to base this on the least accurate of them.


Mascot Database Manager: a...
t440p-vrk/mascot/x-cgi/db_manager.pl?sub=dbs%3Aardvark
Search
Database Manager
Databases (10)
Parse rules (18)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters
Copy
Delete
Edit filters
Import search results

Database: aardvark

Name
aardvark

Database type
Spectral library (created from search results)

Database directory
`/opt/mascot-2.6-dev/sequence/aardvark/current`

Filename pattern
`aardvark_*.msp`

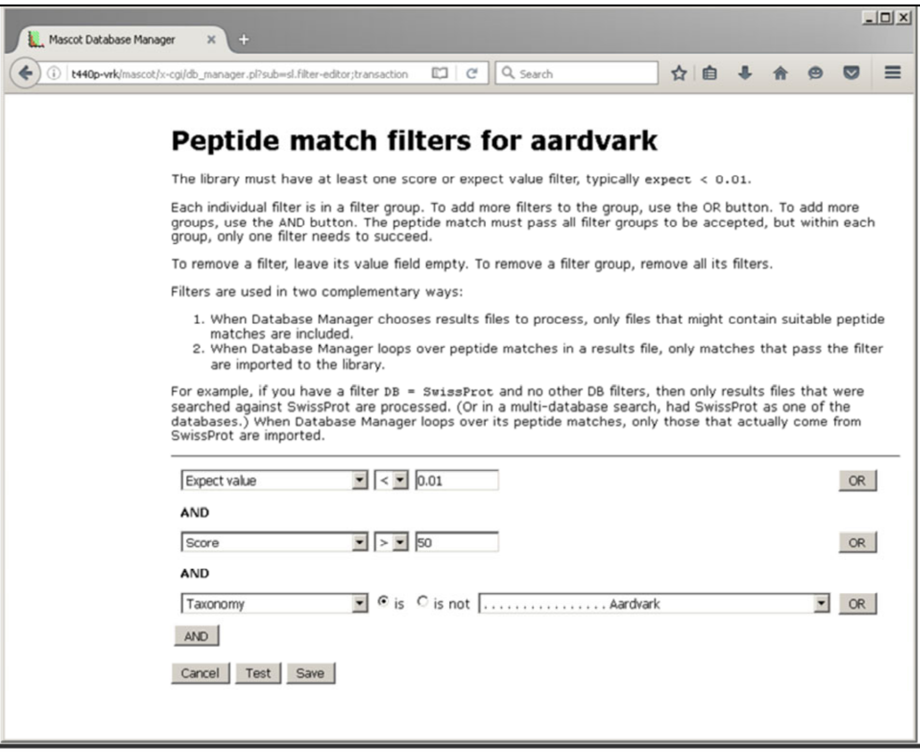
Create MSP file from search results

Peptide match filters
(none)

The spectral library will be created from Mascot search results. Only results files and peptide matches that pass suitable filtering criteria will be included in the library.

Please configure peptide match filters. After that you can add results to the library.

Peptide match filters are used to select matches for inclusion in the library. We choose ‘Edit filters’



Mascot Database Manager
t440p-vrk/mascot/x-cgi/db_manager.pl?sub=sl.filter-editor;transaction
Search

Peptide match filters for aardvark

The library must have at least one score or expect value filter, typically `expect < 0.01`.

Each individual filter is in a filter group. To add more filters to the group, use the OR button. To add more groups, use the AND button. The peptide match must pass all filter groups to be accepted, but within each group, only one filter needs to succeed.

To remove a filter, leave its value field empty. To remove a filter group, remove all its filters.

Filters are used in two complementary ways:

1. When Database Manager chooses results files to process, only files that might contain suitable peptide matches are included.
2. When Database Manager loops over peptide matches in a results file, only matches that pass the filter are imported to the library.

For example, if you have a filter `DB = SwissProt` and no other DB filters, then only results files that were searched against SwissProt are processed. (Or in a multi-database search, had SwissProt as one of the databases.) When Database Manager loops over its peptide matches, only those that actually come from SwissProt are imported.

Expect value < 0.01

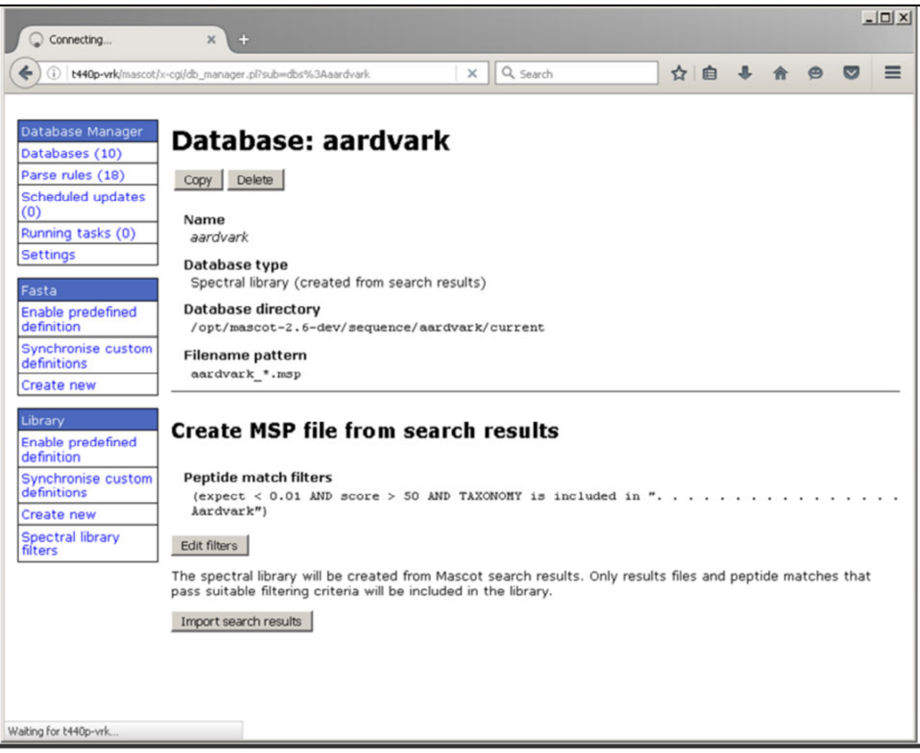
AND

Score > 50

AND

Taxonomy ● is ○ is not .... Aardvark

There is a lot of flexibility here, and we don’t have time to go into all the possibilities. This would be a simple filter for PSMs that can be assigned to a specific organism. We only want strong, confident matches in our library, so we require the match to have an expect value less than 0.01 and a score greater than 50. If the set of search results includes duplicate PSMs, only the one with the highest score goes into the library. We choose Save …



Database: aardvark

Name
aardvark

Database type
Spectral library (created from search results)

Database directory
`/opt/mascot-2.6-dev/sequence/aardvark/current`

Filename pattern
`aardvark_*.msp`

Create MSP file from search results

Peptide match filters

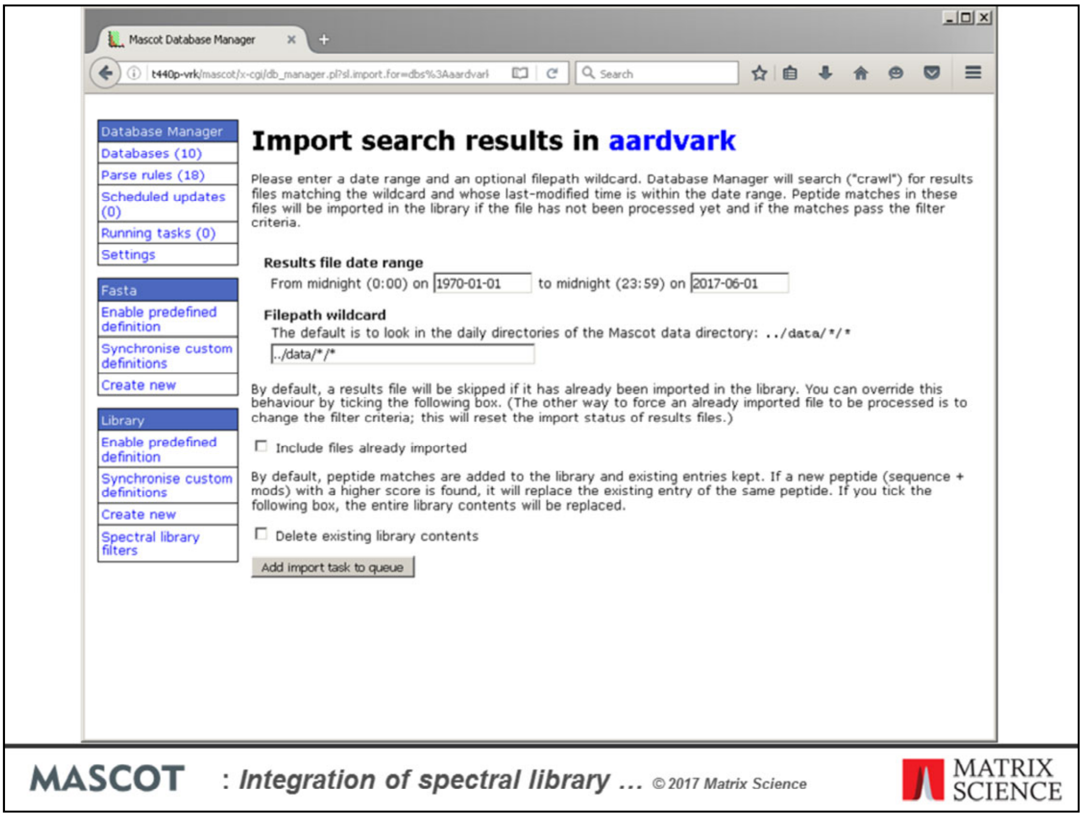
```
(expect < 0.01 AND score > 50 AND TAXONOMY is included in ". . . . . . . . . . . . . . . .
Aardvark")
```

The spectral library will be created from Mascot search results. Only results files and peptide matches that pass suitable filtering criteria will be included in the library.

MASCOT : *Integration of spectral library …* © 2017 Matrix Science

MATRIX SCIENCE

Which takes us back to the previous page, and we are ready to import search results


Mascot Database Manager
t440p-vrk/mascot/x-cgi/db_manager.pl?sl.import.for=dbs%3Aaardvark
Search
Database Manager
Databases (10)
Parse rules (18)
Scheduled updates (0)
Running tasks (0)
Settings
Fasta
Enable predefined definition
Synchronise custom definitions
Create new
Library
Enable predefined definition
Synchronise custom definitions
Create new
Spectral library filters

Import search results in aardvark

Please enter a date range and an optional filepath wildcard. Database Manager will search ("crawl") for results files matching the wildcard and whose last-modified time is within the date range. Peptide matches in these files will be imported in the library if the file has not been processed yet and if the matches pass the filter criteria.

Results file date range

From midnight (0:00) on 1970-01-01 to midnight (23:59) on 2017-06-01

Filepath wildcard

The default is to look in the daily directories of the Mascot data directory: `../data/*/*`

../data/*/*

By default, a results file will be skipped if it has already been imported in the library. You can override this behaviour by ticking the following box. (The other way to force an already imported file to be processed is to change the filter criteria; this will reset the import status of results files.)

□ Include files already imported

By default, peptide matches are added to the library and existing entries kept. If a new peptide (sequence + mods) with a higher score is found, it will replace the existing entry of the same peptide. If you tick the following box, the entire library contents will be replaced.

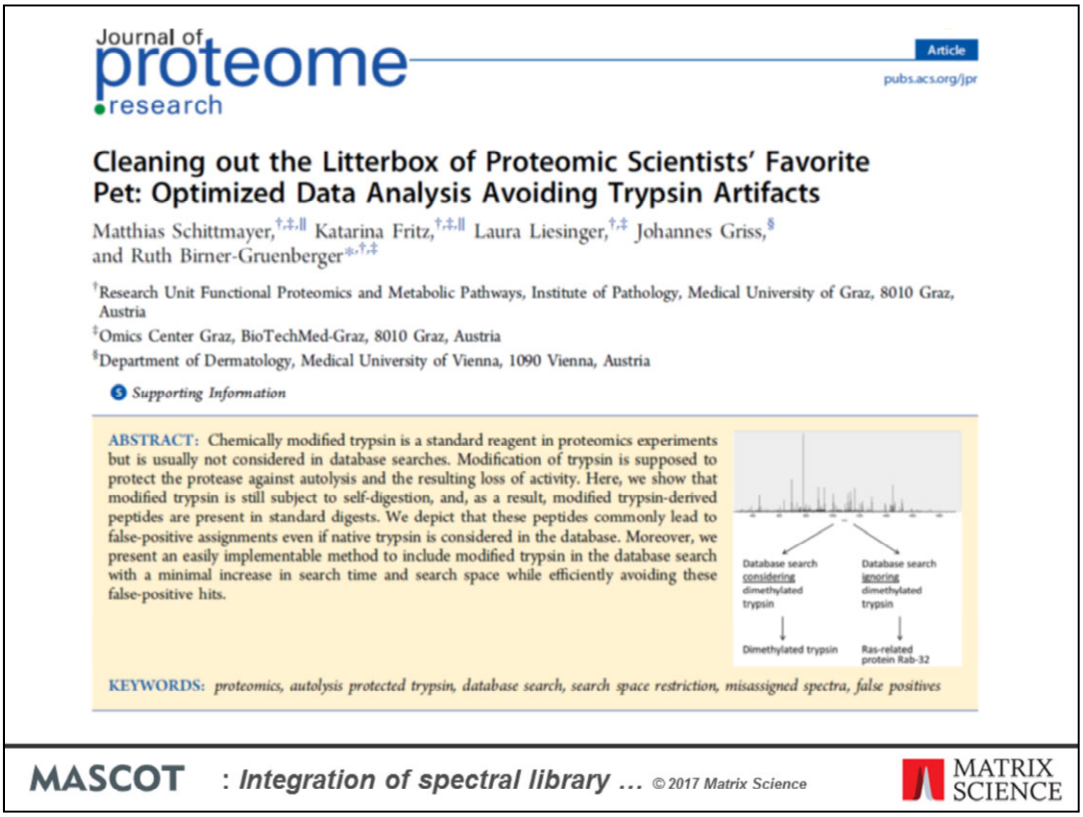
□ Delete existing library contents

MASCOT : *Integration of spectral library …* © 2017 Matrix Science

MATRIX SCIENCE

The only other thing we need to decide is which search result files to crawl. This can be specified as a date range or a wild card file path or some combination of the two. Finally, we add the import task to the queue and the selected files will be crawled as a background task.

You can even schedule automatic updates for such a database, which means that matches can be imported from new result files, created since the last import



Journal of proteome research

Article

pubs.acs.org/jpr

Cleaning out the Litterbox of Proteomic Scientists' Favorite Pet: Optimized Data Analysis Avoiding Trypsin Artifacts

Matthias Schittmayer,[†,‡,∥] Katarina Fritz,[†,‡,∥] Laura Liesinger,[†,‡] Johannes Griss,[§] and Ruth Birner-Gruenberger[*,†,‡]

[†] Research Unit Functional Proteomics and Metabolic Pathways, Institute of Pathology, Medical University of Graz, 8010 Graz, Austria

[‡] Omics Center Graz, BioTechMed-Graz, 8010 Graz, Austria

[§] Department of Dermatology, Medical University of Vienna, 1090 Vienna, Austria

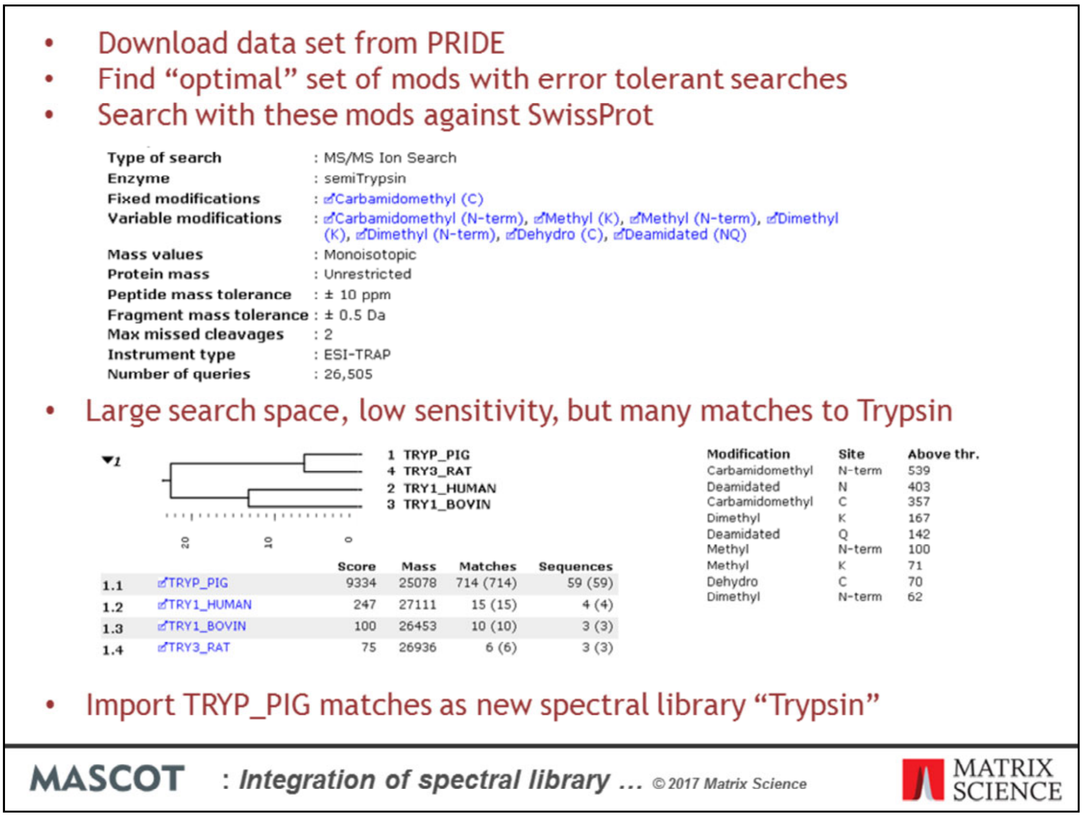
Supporting Information

ABSTRACT: Chemically modified trypsin is a standard reagent in proteomics experiments but is usually not considered in database searches. Modification of trypsin is supposed to protect the protease against autolysis and the resulting loss of activity. Here, we show that modified trypsin is still subject to self-digestion, and, as a result, modified trypsin-derived peptides are present in standard digests. We depict that these peptides commonly lead to false-positive assignments even if native trypsin is considered in the database. Moreover, we present an easily implementable method to include modified trypsin in the database search with a minimal increase in search time and search space while efficiently avoiding these false-positive hits.

Database search considering dimethylated trypsin
Database search ignoring dimethylated trypsin
Dimethylated trypsin
Ras-related protein Rab-32

KEYWORDS: proteomics, autolysis protected trypsin, database search, search space restriction, misassigned spectra, false positives

Let's look at a practical example of how these new features might be used. This recent paper JPR reminded us that sequencing grade trypsin is modified by methylation or acetylation of the lysines. Unless these variable modifications are selected in a search, simply including a contaminants database will not be sufficient to catch all trypsin autolysis peptides. The authors suggested a solution based on editing the sequence of trypsin in the Fasta, replacing K with J, and defining J as the mass of dimethylated lysine. This is fine, as far as it goes, but it misses many of the other modifications that are present, not to mention extensive non-specific cleavage.



- Download data set from PRIDE
- Find “optimal” set of mods with error tolerant searches
- Search with these mods against SwissProt

Type of search	: MS/MS Ion Search
Enzyme	: semiTrypsin
Fixed modifications	: Carbamidomethyl (C)
Variable modifications	: Carbamidomethyl (N-term), Methyl (K), Methyl (N-term), Dimethyl (K), Dimethyl (N-term), Dehydro (C), Deamidated (NQ)
Mass values	: Monoisotopic
Protein mass	: Unrestricted
Peptide mass tolerance	: ± 10 ppm
Fragment mass tolerance	: ± 0.5 Da
Max missed cleavages	: 2
Instrument type	: ESI-TRAP
Number of queries	: 26,505

- Large search space, low sensitivity, but many matches to Trypsin

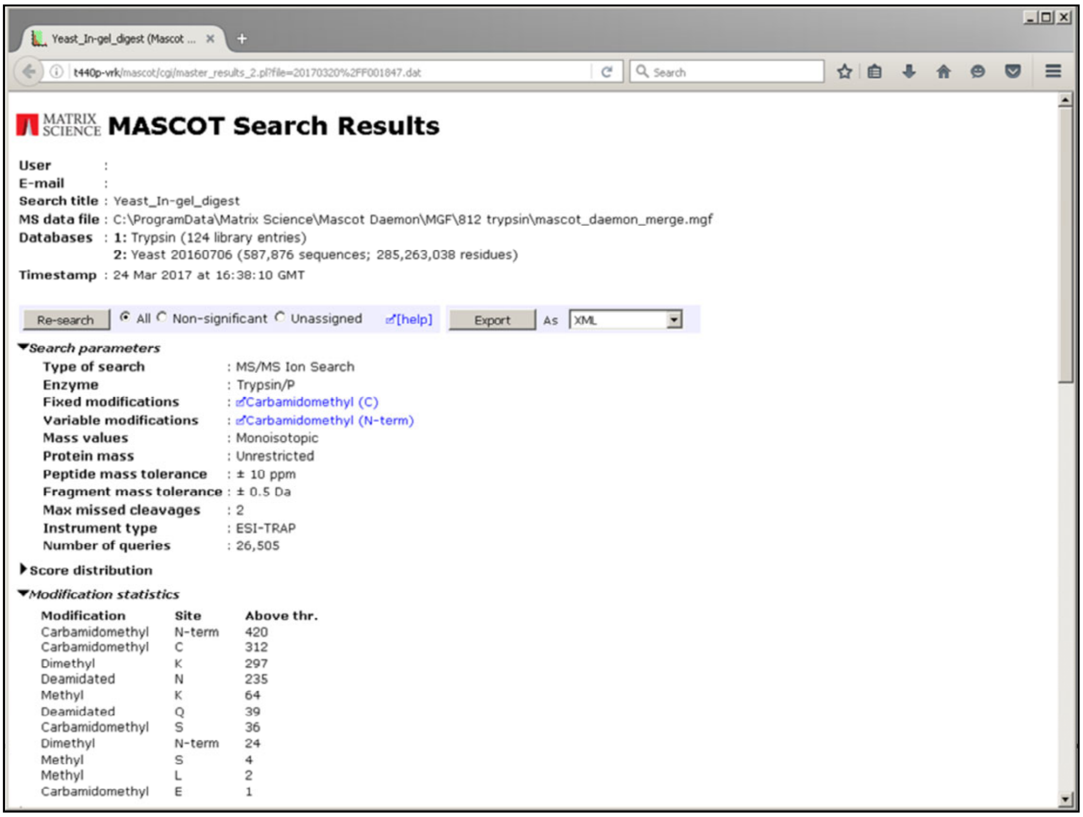
		Score	Mass	Matches	Sequences
1.1	TRYP_PIG	9334	25078	714 (714)	59 (59)
1.2	TRY1_HUMAN	247	27111	15 (15)	4 (4)
1.3	TRY1_BOVIN	100	26453	10 (10)	3 (3)
1.4	TRY3_RAT	75	26936	6 (6)	3 (3)

Modification	Site	Above thr.
Carbamidomethyl	N-term	539
Deamidated	N	403
Carbamidomethyl	C	357
Dimethyl	K	167
Deamidated	Q	142
Methyl	N-term	100
Methyl	K	71
Dehydro	C	70
Dimethyl	N-term	62

- Import TRYP_PIG matches as new spectral library “Trypsin”

We downloaded the raw files for one of the data sets in this study from PRIDE and tried a variety of error tolerant searches to discover exactly what was present. Based on these results, we chose these search settings. The enzyme specificity was semiTrypsin because peptides show very extensive C-terminal ‘ragged ends’

This makes the search space very large. The search takes a long time and overall sensitivity is not as good as it would be for a simple search with strict trypsin and only one or two variable modifications. The answer, of course, is to make a library of the trypsin matches and include this in the vanilla search. This is a very powerful option, since it allows any number of modified and non-specific peptides from any number of contaminants to be intercepted with no increase in the search space.



Yeast_In-gel_digest (Mascot ...
t440p-vrk/mascot/cgi/master_results_2.pl?file=20170320%2FF001847.dat
Search
MATRIX SCIENCE

MASCOT Search Results

User :
E-mail :
Search title : Yeast_In-gel_digest
MS data file : C:\ProgramData\Matrix Science\Mascot Daemon\MGF\812 trypsin\mascot_daemon_merge.mgf
Databases : **1:** Trypsin (124 library entries)
2: Yeast 20160706 (587,876 sequences; 285,263,038 residues)
Timestamp : 24 Mar 2017 at 16:38:10 GMT

Re-search ● All ○ Non-significant ○ Unassigned [help] Export As XML

Search parameters

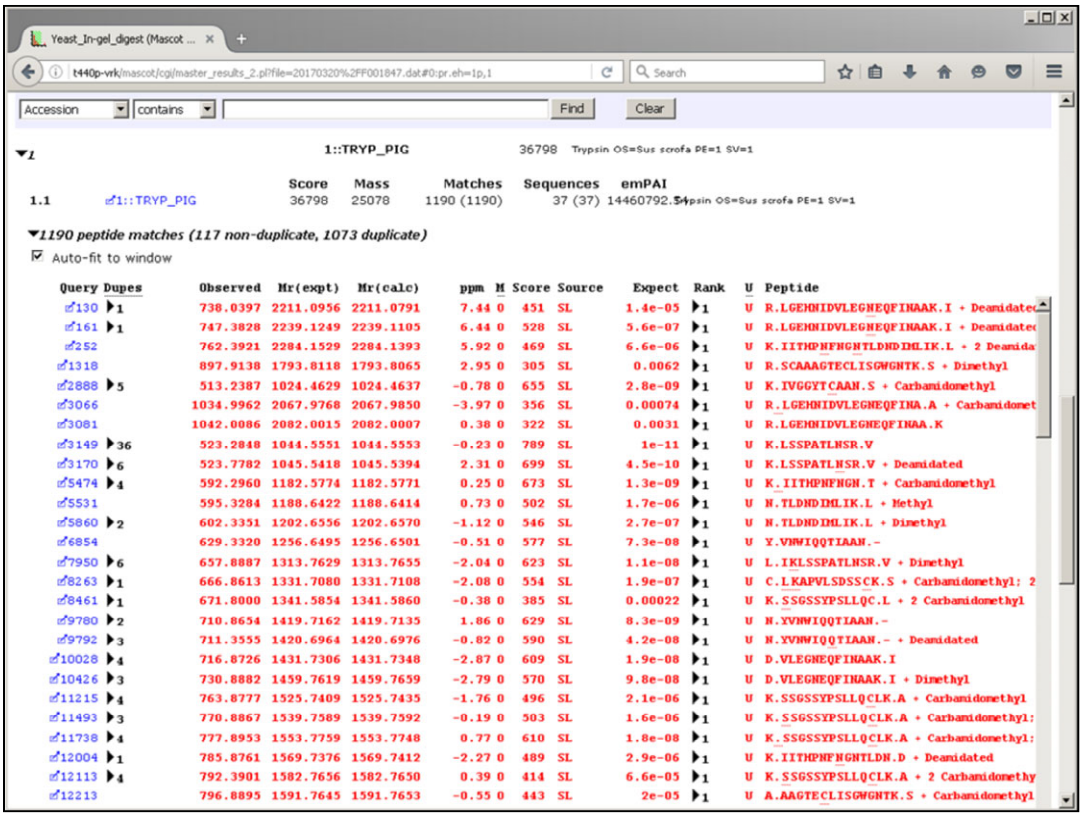
Type of search	: MS/MS Ion Search
Enzyme	: Trypsin/P
Fixed modifications	: Carbamidomethyl (C)
Variable modifications	: Carbamidomethyl (N-term)
Mass values	: Monoisotopic
Protein mass	: Unrestricted
Peptide mass tolerance	: ± 10 ppm
Fragment mass tolerance	: ± 0.5 Da
Max missed cleavages	: 2
Instrument type	: ESI-TRAP
Number of queries	: 26,505

Score distribution

Modification statistics

Modification	Site	Above thr.
Carbamidomethyl	N-term	420
Carbamidomethyl	C	312
Dimethyl	K	297
Deamidated	N	235
Methyl	K	64
Deamidated	Q	39
Carbamidomethyl	S	36
Dimethyl	N-term	24
Methyl	S	4
Methyl	L	2
Carbamidomethyl	E	1

Here, we search a yeast database plus the tryptic autolysis library with strict trypsin and a single variable mod - yet still obtain matches to all the modified and non-specific trypsin autolysis peptides


Yeast_In-gel_digest (Mascot ...
t440p-vrk/mascot/cgi/master_results_2.pl?file=20170320%2FF001847.dat#0:pr.eh=1p,1
Search
Accession
contains
Find
Clear

▼1 1::TRYP_PIG 36798 Trypsin OS=Sus scrofa PE=1 SV=1

		Score	Mass	Matches	Sequences	emPAI
1.1	1::TRYP_PIG	36798	25078	1190 (1190)	37 (37)	14460792.54psin OS=Sus scrofa PE=1 SV=1

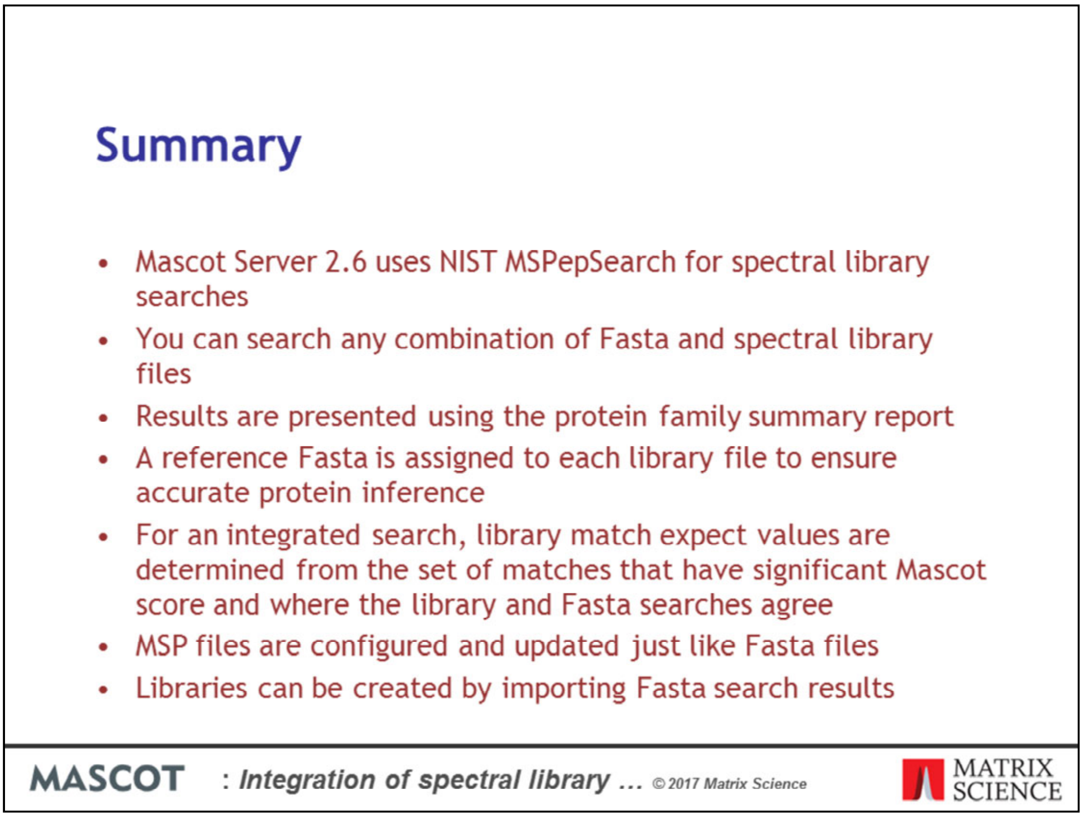
▼*1190 peptide matches (117 non-duplicate, 1073 duplicate)*

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Source	Expect	Rank	U	Peptide
130	▶1	738.0397	2211.0956	2211.0791	7.44	0	451	SL	1.4e-05	▶1	U	R.LGEHNIDVLEGNEQFINAAK.I + Deamidate
161	▶1	747.3828	2239.1249	2239.1105	6.44	0	528	SL	5.6e-07	▶1	U	R.LGEHNIDVLEGNEQFINAAK.I + Deamidate
252		762.3921	2284.1529	2284.1393	5.92	0	469	SL	6.6e-06	▶1	U	K.IITHPNFNGNTLDNDIMLIK.L + 2 Deamida
1318		897.9138	1793.8118	1793.8065	2.95	0	305	SL	0.0062	▶1	U	R.SCAAAGTECLISGWGNTK.S + Dimethyl
2888	▶5	513.2387	1024.4629	1024.4637	-0.78	0	655	SL	2.8e-09	▶1	U	K.IVGGYTCAAN.S + Carbamidomethyl
3066		1034.9962	2067.9768	2067.9850	-3.97	0	356	SL	0.00074	▶1	U	R.LGEHNIDVLEGNEQFINA.A + Carbamidomet
3081		1042.0086	2082.0015	2082.0007	0.38	0	322	SL	0.0031	▶1	U	R.LGEHNIDVLEGNEQFINAA.K
3149	▶36	523.2848	1044.5551	1044.5553	-0.23	0	789	SL	1e-11	▶1	U	K.LSSPATLNSR.V
3170	▶6	523.7782	1045.5418	1045.5394	2.31	0	699	SL	4.5e-10	▶1	U	K.LSSPATLNSR.V + Deamidated
5474	▶4	592.2960	1182.5774	1182.5771	0.25	0	673	SL	1.3e-09	▶1	U	K.IITHPNFNGN.T + Carbamidomethyl
5531		595.3284	1188.6422	1188.6414	0.73	0	502	SL	1.7e-06	▶1	U	N.TLDNDIMLIK.L + Methyl
5860	▶2	602.3351	1202.6556	1202.6570	-1.12	0	546	SL	2.7e-07	▶1	U	N.TLDNDIMLIK.L + Dimethyl
6854		629.3320	1256.6495	1256.6501	-0.51	0	577	SL	7.3e-08	▶1	U	Y.VNWIQQTIAAN.-
7950	▶6	657.8887	1313.7629	1313.7655	-2.04	0	623	SL	1.1e-08	▶1	U	L.IKLSSPATLNSR.V + Dimethyl
8263	▶1	666.8613	1331.7080	1331.7108	-2.08	0	554	SL	1.9e-07	▶1	U	C.LKAPVLSDSSCK.S + Carbamidomethyl; 2
8461	▶1	671.8000	1341.5854	1341.5860	-0.38	0	385	SL	0.00022	▶1	U	K.SSGSSYPSLLQC.L + 2 Carbamidomethyl
9780	▶2	710.8654	1419.7162	1419.7135	1.86	0	629	SL	8.3e-09	▶1	U	N.YVNWIQQTIAAN.-
9792	▶3	711.3555	1420.6964	1420.6976	-0.82	0	590	SL	4.2e-08	▶1	U	N.YVNWIQQTIAAN.- + Deamidated
10028	▶4	716.8726	1431.7306	1431.7348	-2.87	0	609	SL	1.9e-08	▶1	U	D.VLEGNEQFINAAK.I
10426	▶3	730.8882	1459.7619	1459.7659	-2.79	0	570	SL	9.8e-08	▶1	U	D.VLEGNEQFINAAK.I + Dimethyl
11215	▶4	763.8777	1525.7409	1525.7435	-1.76	0	496	SL	2.1e-06	▶1	U	K.SSGSSYPSLLQCLK.A + Carbamidomethyl
11493	▶3	770.8867	1539.7589	1539.7592	-0.19	0	503	SL	1.6e-06	▶1	U	K.SSGSSYPSLLQCLK.A + Carbamidomethyl;
11738	▶4	777.8953	1553.7759	1553.7748	0.77	0	610	SL	1.8e-08	▶1	U	K.SSGSSYPSLLQCLK.A + Carbamidomethyl;
12004	▶1	785.8761	1569.7376	1569.7412	-2.27	0	489	SL	2.9e-06	▶1	U	K.IITHPNFNGNTLDN.D + Deamidated
12113	▶4	792.3901	1582.7656	1582.7650	0.39	0	414	SL	6.6e-05	▶1	U	K.SSGSSYPSLLQCLK.A + 2 Carbamidomethy
12213		796.8895	1591.7645	1591.7653	-0.55	0	443	SL	2e-05	▶1	U	A.AAGTECLISGWGNTK.S + Carbamidomethyl

This removes 1190 spectra which otherwise might have given rise to false positives.

If you're wondering about the ridiculous emPAI value, it's because the assumption behind emPAI is strict tryptic cleavage. However, the library search is giving all kinds of semitryptic matches, so the model assumptions are not satisfied.



Summary

- Mascot Server 2.6 uses NIST MSPepSearch for spectral library searches
- You can search any combination of Fasta and spectral library files
- Results are presented using the protein family summary report
- A reference Fasta is assigned to each library file to ensure accurate protein inference
- For an integrated search, library match expect values are determined from the set of matches that have significant Mascot score and where the library and Fasta searches agree
- MSP files are configured and updated just like Fasta files
- Libraries can be created by importing Fasta search results

MASCOT : *Integration of spectral library …* © 2017 Matrix Science

MATRIX SCIENCE

To summarise