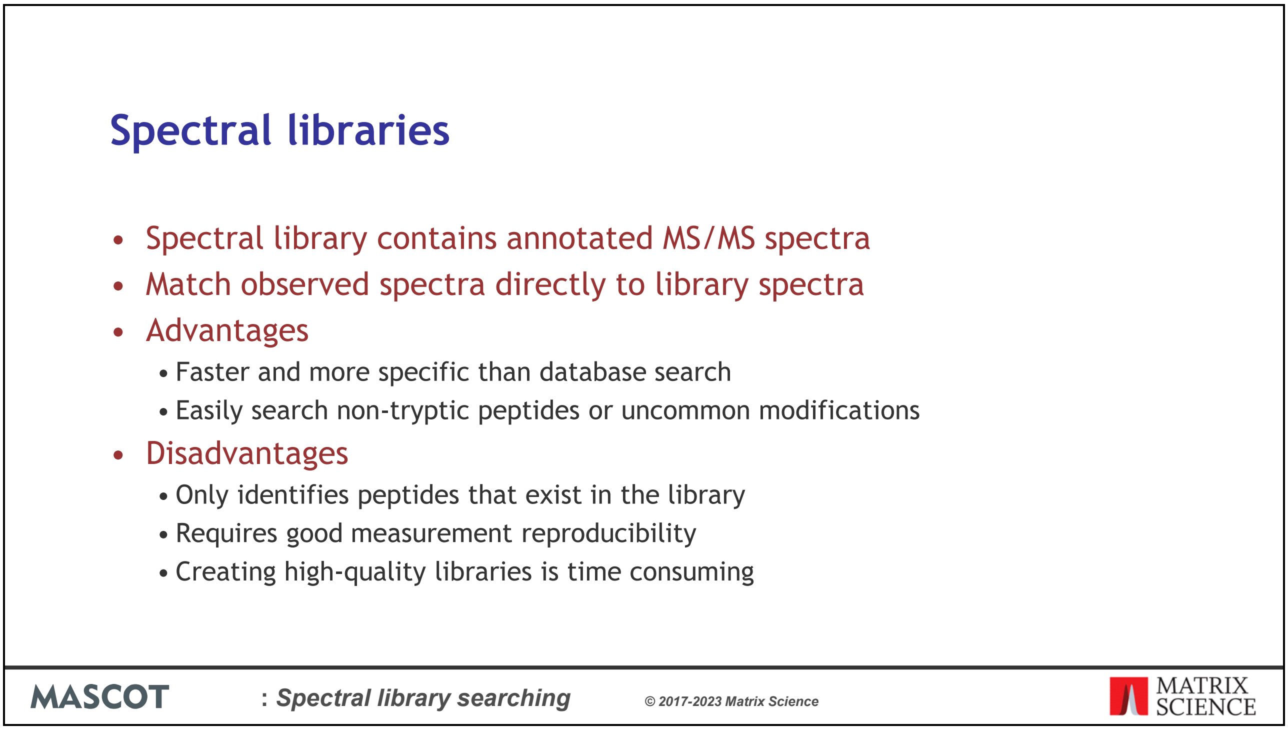
Spectral Library Searching



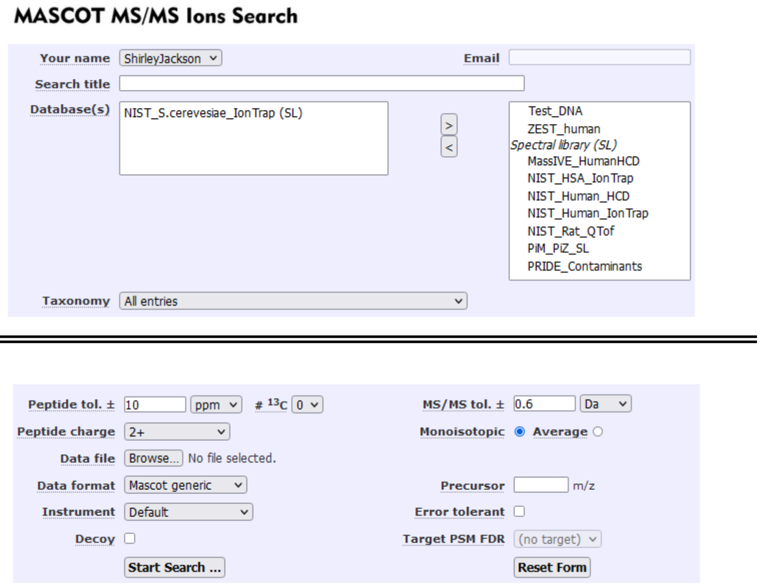
Spectral libraries

- Spectral library contains annotated MS/MS spectra
- Match observed spectra directly to library spectra
- Advantages
  - Faster and more specific than database search
  - Easily search non-tryptic peptides or uncommon modifications
- Disadvantages
  - Only identifies peptides that exist in the library
  - Requires good measurement reproducibility
  - Creating high-quality libraries is time consuming

A spectral library is a collection of annotated MS/MS spectra of peptides. Instead of searching observed spectra against a protein sequence database, you search the observed spectra directly against a spectral library. The observed peaks are compared to the annotated library peaks, then scored in some way based on similarity. Typically, the similarity score takes advantage of peak intensity patterns as well as peak masses. It may also utilise peak annotations, such as giving a higher score to b and y ion matches.

There are several advantages compared to protein sequence databases. A library search is often much faster than a database search, as spectral library typically has orders of magnitude fewer peptides than a tryptic digest of a sequence database. A library search can also be more specific than a database search, for example if it contains previously identified non-tryptic peptides or uncommon variable modifications. Selecting a semi-specific enzyme or including many uncommon variable modifications in a database search greatly increases the search duration. Searching a pre-prepared library of spectra of semi-specific peptides is much faster.

There are no free lunches, so of course spectral library searching has weaknesses. You can only identify peptides for which a spectrum exists in the library. If the library contains a peptide sequence with one missed cleavage, you will not get a match to a peptide with two missed cleavages. A library search also requires decent measurement reproducibility. If peak intensities vary wildly between repeat runs, it is harder to get a good library match. Finally, creating high-quality libraries is time consuming and some care is needed. Fortunately, Mascot ships with several predefined spectral libraries to help you get started.



MASCOT MS/MS Ions Search

Your name: ShirleyJackson
Email:
Search title:
Database(s): NIST_S.cerevesiae_IonTrap (SL)

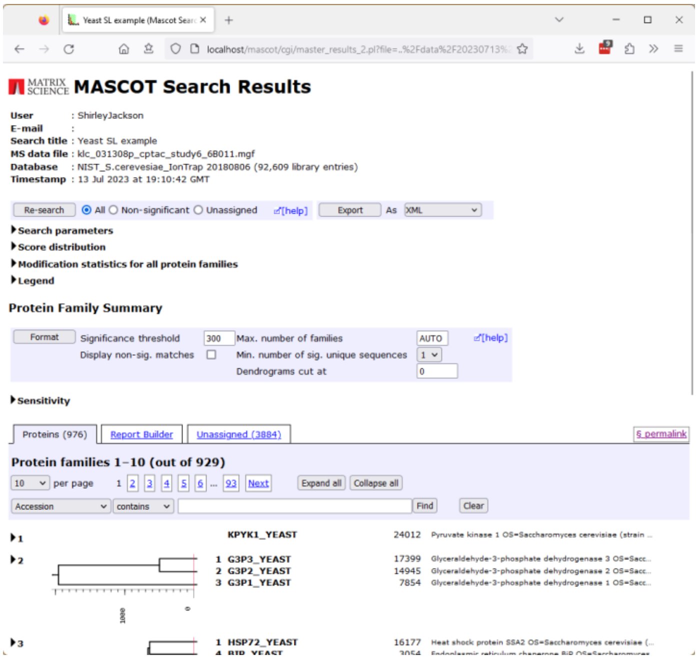
Test_DNA
ZEST_human
Spectral library (SL)
MassIVE_HumanHCD
NIST_HSA_IonTrap
NIST_Human_HCD
NIST_Human_IonTrap
NIST_Rat_QTof
PiM_PiZ_SL
PRIDE_Contaminants

Taxonomy: All entries

Peptide tol. ± 10 ppm # ^{13}C 0
MS/MS tol. ± 0.6 Da
Peptide charge: 2+
Monoisotopic ● Average ○
Data file: Browse... No file selected.
Data format: Mascot generic
Precursor: m/z
Instrument: Default
Error tolerant □
Decoy □
Target PSM FDR: (no target)

Mascot Server can search spectral libraries using MSPepSearch from Steven Stein’s group at NIST. When submitting a search, any combination of amino acid FASTA or nucleic acid FASTA databases, and spectral libraries 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.



Yeast SL example (Mascot Searc ×
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20230713%

MASCOT Search Results

User : ShirleyJackson
E-mail :
Search title : Yeast SL example
MS data file : klc_031308p_cptac_study6_6B011.mgf
Database : NIST_S.cerevesiae_IonTrap 20180806 (92,609 library entries)
Timestamp : 13 Jul 2023 at 19:10:42 GMT

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

- Search parameters
- Score distribution
- Modification statistics for all protein families
- Legend

Protein Family Summary

Significance threshold 300
Display non-sig. matches □
Max. number of families AUTO
Min. number of sig. unique sequences 1
Dendrograms cut at 0

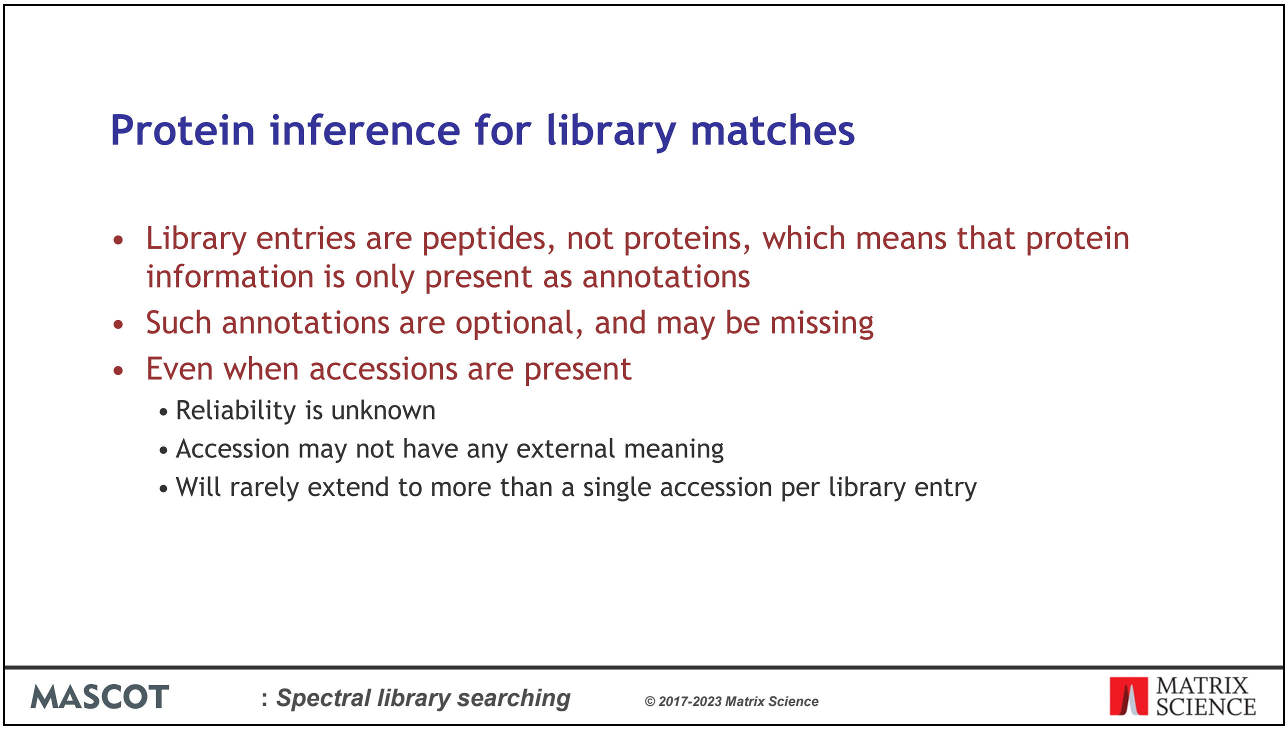
[help]
Sensitivity
Proteins (976) | Report Builder | Unassigned (3884) | § permalink

Protein families 1–10 (out of 929)

10 per page 1 2 3 4 5 6 ... 93 Next Expand all Collapse all
Accession contains Find Clear
1000
0

1		KPYK1_YEAST	24012	Pyruvate kinase 1 OS=Saccharomyces cerevisiae (strain ...
2	1	G3P3_YEAST	17399	Glyceraldehyde-3-phosphate dehydrogenase 3 OS=Sacc...
	2	G3P2_YEAST	14945	Glyceraldehyde-3-phosphate dehydrogenase 2 OS=Sacc...
	3	G3P1_YEAST	7854	Glyceraldehyde-3-phosphate dehydrogenase 1 OS=Sacc...
3	1	HSP72_YEAST	16177	Heat shock protein SSA2 OS=Saccharomyces cerevisiae (...
	4	BIP_YEAST	3054	Endoplasmic reticulum chaperone BiP OS=Saccharomyces

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.

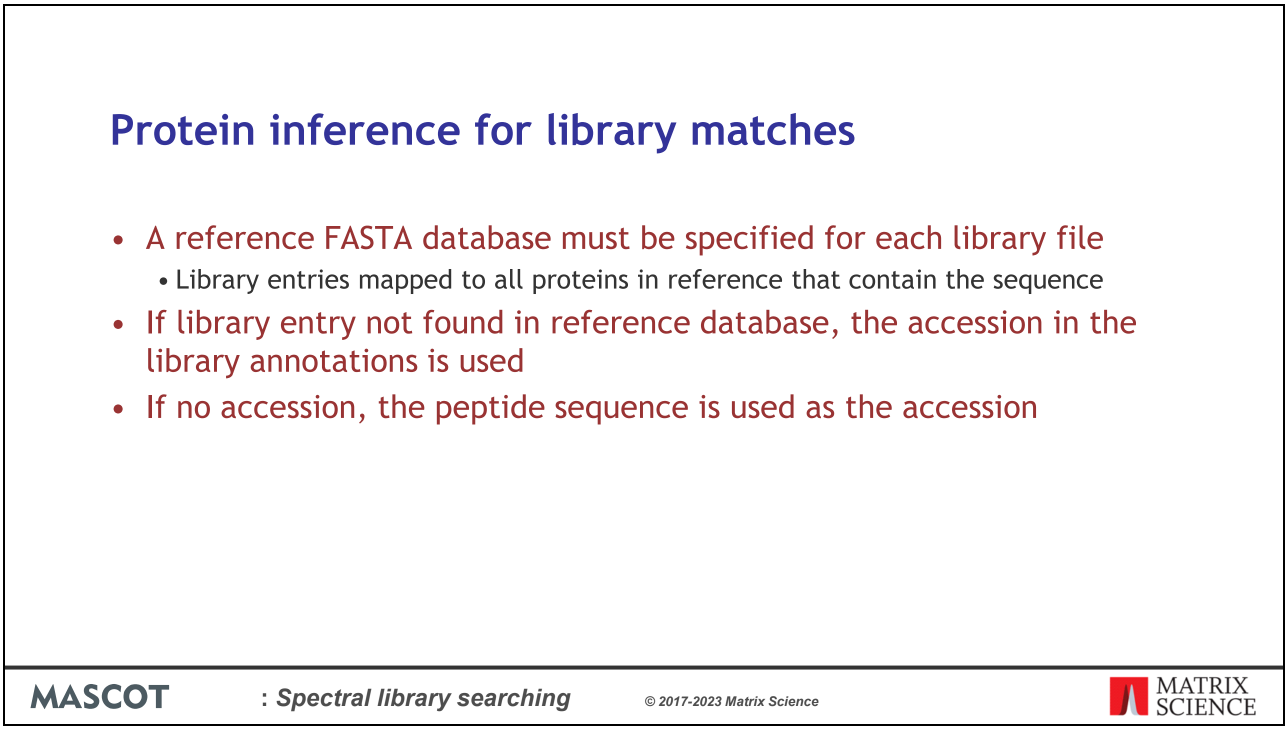


Protein inference for library matches

- 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

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.

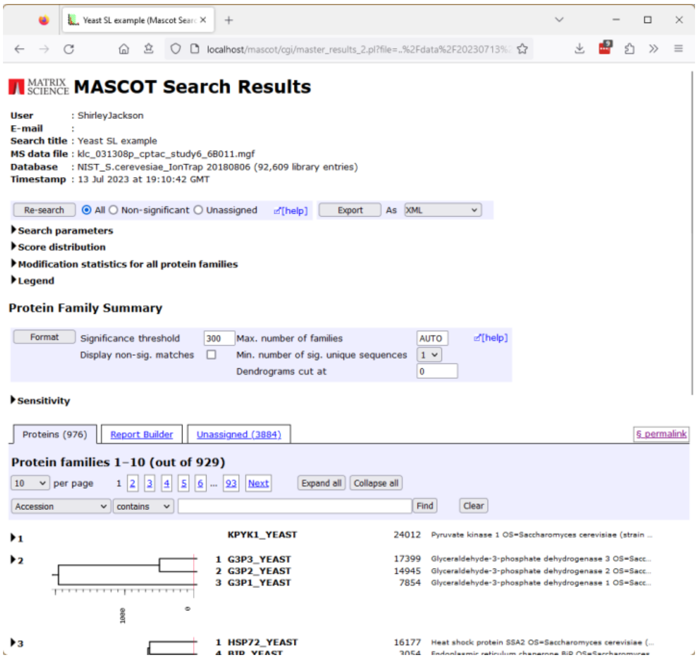
Even when present, the reliability is unknown. The accession could be a meaningless number or string. And, I've never seen a library with more than a single accession per library entry, so protein inference will be inaccurate for shared peptides.



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 database 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.



Yeast SL example (Mascot Searc ×
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20230713%

MASCOT Search Results

User : ShirleyJackson
E-mail :
Search title : Yeast SL example
MS data file : klc_031308p_cptac_study6_6B011.mgf
Database : NIST_S.cerevesiae_IonTrap 20180806 (92,609 library entries)
Timestamp : 13 Jul 2023 at 19:10:42 GMT

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

Search parameters
Score distribution
Modification statistics for all protein families
Legend

Protein Family Summary

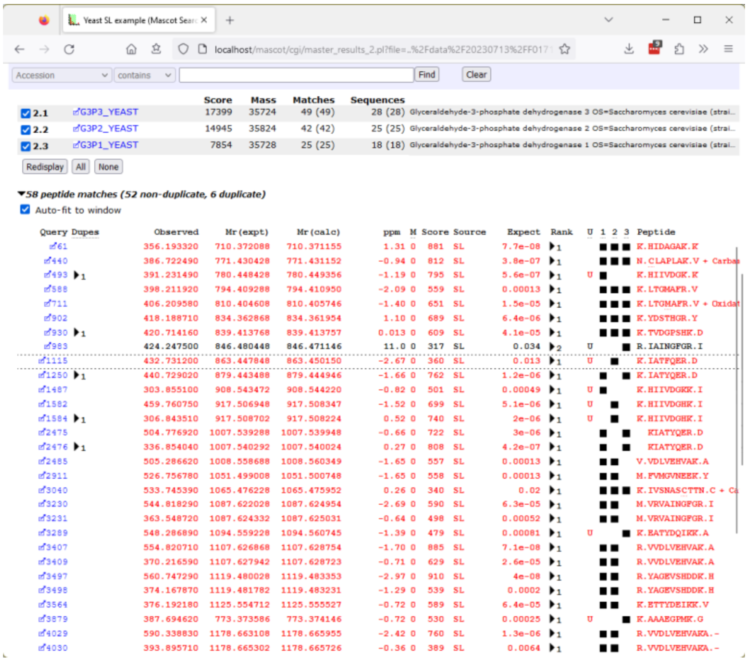
Significance threshold 300
Display non-sig. matches □
Max. number of families AUTO
Min. number of sig. unique sequences 1
Dendrograms cut at 0

Sensitivity

Protein families 1–10 (out of 929)

```
1       KPYK1_YEAST     24012  Pyruvate kinase 1 OS=Saccharomyces cerevisiae (strain ...
2     1 G3P3_YEAST      17399  Glyceraldehyde-3-phosphate dehydrogenase 3 OS=Sacc...
      2 G3P2_YEAST      14945  Glyceraldehyde-3-phosphate dehydrogenase 2 OS=Sacc...
      3 G3P1_YEAST       7854  Glyceraldehyde-3-phosphate dehydrogenase 1 OS=Sacc...
3     1 HSP72_YEAST     16177  Heat shock protein SSA2 OS=Saccharomyces cerevisiae (...
      4 BIP_YEAST        3054  Endoplasmic reticulum chaperone BiP OS=Saccharomyces
```

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 database matches.


Yeast SL example (Mascot Searc ×
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20230713%2FF0171
Accession
contains
Find
Clear

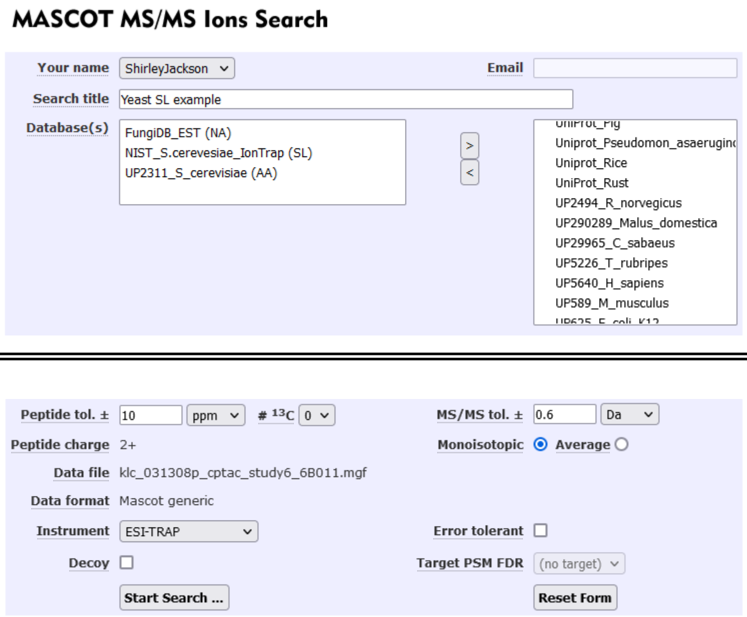
		Score	Mass	Matches	Sequences	
☑ 2.1	G3P3_YEAST	17399	35724	49 (49)	28 (28)	Glyceraldehyde-3-phosphate dehydrogenase 3 OS=Saccharomyces cerevisiae (strai...
☑ 2.2	G3P2_YEAST	14945	35824	42 (42)	25 (25)	Glyceraldehyde-3-phosphate dehydrogenase 2 OS=Saccharomyces cerevisiae (strai...
☑ 2.3	G3P1_YEAST	7854	35728	25 (25)	18 (18)	Glyceraldehyde-3-phosphate dehydrogenase 1 OS=Saccharomyces cerevisiae (strai...

▼58 peptide matches (52 non-duplicate, 6 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Source	Expect	Rank	U	1	2	3	Peptide
61		356.193320	710.372088	710.371155	1.31	0	881	SL	7.7e-08	▶1		■	■	■	K.HIDAGAK.K
440		386.722490	771.430428	771.431152	-0.94	0	812	SL	3.8e-07	▶1		■	■	■	N.CLAPLAK.V + Carba
493	▶1	391.231490	780.448428	780.449356	-1.19	0	795	SL	5.6e-07	▶1	U	■			K.HIIVDGK.K
588		398.211920	794.409288	794.410950	-2.09	0	559	SL	0.00013	▶1		■	■	■	K.LTGMAFR.V
711		406.209580	810.404608	810.405746	-1.40	0	651	SL	1.5e-05	▶1		■	■	■	K.LTGMAFR.V + Oxidat
902		418.188710	834.362868	834.361954	1.10	0	689	SL	6.4e-06	▶1		■	■	■	K.YDSTHGR.Y
930	▶1	420.714160	839.413768	839.413757	0.013	0	609	SL	4.1e-05	▶1		■	■	■	K.TVDGPSHK.D
983		424.247500	846.480448	846.471146	11.0	0	317	SL	0.034	▶2	U			■	R.IAINGFGR.I
1115		432.731200	863.447848	863.450150	-2.67	0	360	SL	0.013	▶1	U		■		K.IATFQER.D
1250	▶1	440.729020	879.443488	879.444946	-1.66	0	762	SL	1.2e-06	▶1		■		■	K.IATYQER.D
1487		303.855100	908.543472	908.544220	-0.82	0	501	SL	0.00049	▶1	U	■			K.HIIVDGKK.I
1582		459.760750	917.506948	917.508347	-1.52	0	699	SL	5.1e-06	▶1	U		■		K.HIIVDGHK.I
1584	▶1	306.843510	917.508702	917.508224	0.52	0	740	SL	2e-06	▶1	U		■		K.HIIVDGHK.I
2475		504.776920	1007.539288	1007.539948	-0.66	0	722	SL	3e-06	▶1		■		■	KIATYQER.D
2476	▶1	336.854040	1007.540292	1007.540024	0.27	0	808	SL	4.2e-07	▶1		■		■	KIATYQER.D
2485		505.286620	1008.558688	1008.560349	-1.65	0	557	SL	0.00013	▶1		■	■		V.VDLVEHVAK.A
2911		526.756780	1051.499008	1051.500748	-1.65	0	558	SL	0.00013	▶1		■	■		M.FVMGVNEEK.Y
3040		533.745390	1065.476228	1065.475952	0.26	0	340	SL	0.02	▶1		■	■	■	K.IVSNASCTTN.C + Ca
3230		544.818290	1087.622028	1087.624954	-2.69	0	590	SL	6.3e-05	▶1		■	■		M.VRVAINGFGR.I
3231		363.548720	1087.624332	1087.625031	-0.64	0	498	SL	0.00052	▶1		■	■		M.VRVAINGFGR.I
3289		548.286890	1094.559228	1094.560745	-1.39	0	479	SL	0.00081	▶1	U			■	K.EATYDQIKK.A
3407		554.820710	1107.626868	1107.628754	-1.70	0	885	SL	7.1e-08	▶1		■	■		R.VVDLVEHVAK.A
3409		370.216590	1107.627942	1107.628723	-0.71	0	629	SL	2.6e-05	▶1		■	■		R.VVDLVEHVAK.A
3497		560.747290	1119.480028	1119.483353	-2.97	0	910	SL	4e-08	▶1		■	■		R.YAGEVSHDDK.H
3498		374.167870	1119.481782	1119.483231	-1.29	0	539	SL	0.0002	▶1		■	■		R.YAGEVSHDDK.H
3564		376.192180	1125.554712	1125.555527	-0.72	0	589	SL	6.4e-05	▶1		■	■		K.ETTYDEIKK.V
3879		387.694620	773.373586	773.374146	-0.72	0	530	SL	0.00025	▶1	U			■	K.AAAEGPMK.G
4029		590.338830	1178.663108	1178.665955	-2.42	0	760	SL	1.3e-06	▶1		■	■		R.VVDLVEHVAKA.-
4030		393.895710	1178.665302	1178.665726	-0.36	0	389	SL	0.0064	▶1		■	■		R.VVDLVEHVAKA.-

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.

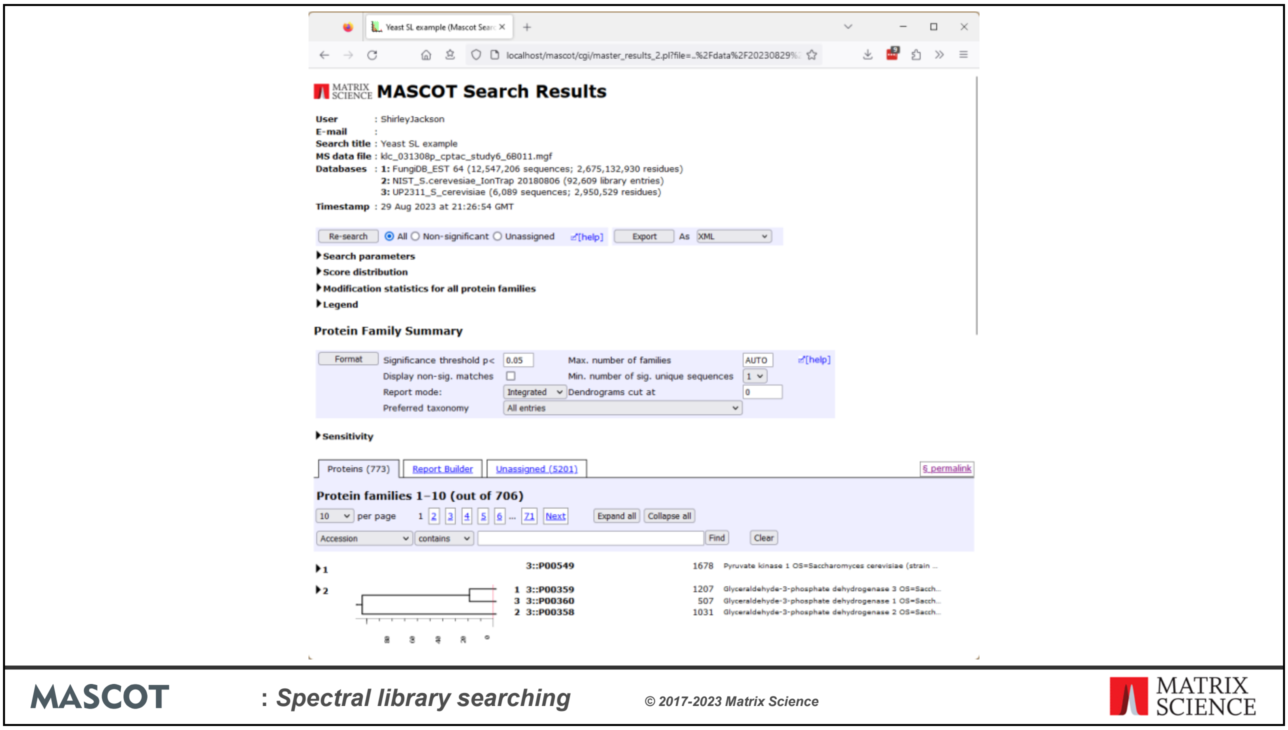


MASCOT MS/MS Ions Search

Your name	ShirleyJackson	Email	
Search title	Yeast SL example		
Database(s)	FungiDB_EST (NA) NIST_S.cerevesiae_IonTrap (SL) UP2311_S_cerevisiae (AA)		UniProt_Fig Uniprot_Pseudomon_asaerugino Uniprot_Rice UniProt_Rust UP2494_R_norvegicus UP290289_Malus_domestica UP29965_C_sabaeus UP5226_T_rubripes UP5640_H_sapiens UP589_M_musculus UP625_E_coli_K12

Peptide tol. ±	10 ppm # ^{13}C 0	MS/MS tol. ±	0.6 Da
Peptide charge	2+	Monoisotopic	● Average ○
Data file	klc_031308p_cptac_study6_6B011.mgf		
Data format	Mascot generic		
Instrument	ESI-TRAP	Error tolerant	□
Decoy	□	Target PSM FDR	(no target)

Let’s expand this example and search a mixture of amino acid and nucleic acid databases with a spectral library.


Yeast SL example (Mascot Searc
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20230829%

MATRIX SCIENCE **MASCOT Search Results**

User : ShirleyJackson
E-mail :
Search title : Yeast SL example
MS data file : klc_031308p_cptac_study6_6B011.mgf
Databases : **1:** FungiDB_EST 64 (12,547,206 sequences; 2,675,132,930 residues)
2: NIST_S.cerevesiae_IonTrap 20180806 (92,609 library entries)
3: UP2311_S_cerevisiae (6,089 sequences; 2,950,529 residues)
Timestamp : 29 Aug 2023 at 21:26:54 GMT

Search parameters
Score distribution
Modification statistics for all protein families
Legend

Protein Family Summary

Significance threshold p< 0.05
Display non-sig. matches □
Report mode: Integrated
Preferred taxonomy All entries
Max. number of families AUTO
Min. number of sig. unique sequences 1
Dendrograms cut at 0

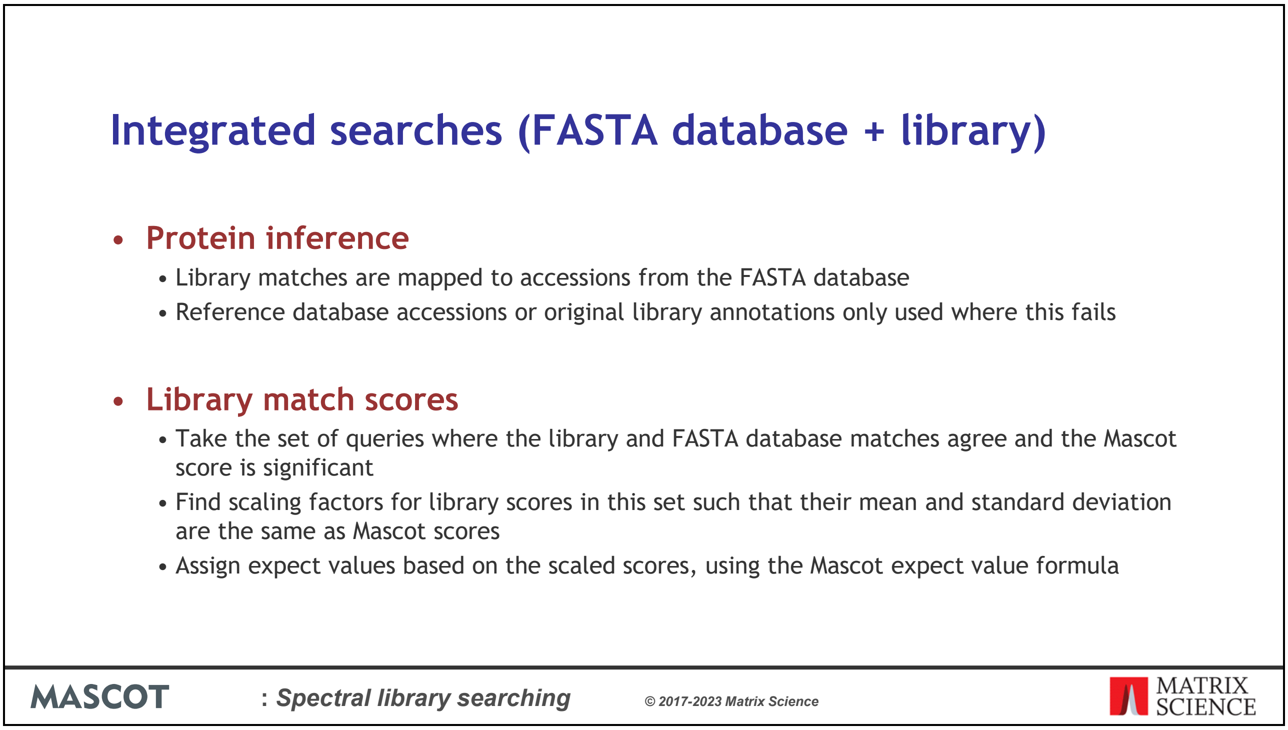
Sensitivity

Protein families 1–10 (out of 706)

1		3::P00549	1678	Pyruvate kinase 1 OS=Saccharomyces cerevisiae (strain ...
2	1	3::P00359	1207	Glyceraldehyde-3-phosphate dehydrogenase 3 OS=Sacch...
	3	3::P00360	507	Glyceraldehyde-3-phosphate dehydrogenase 1 OS=Sacch...
	2	3::P00358	1031	Glyceraldehyde-3-phosphate dehydrogenase 2 OS=Sacch...

MASCOT : *Spectral library searching* *© 2017-2023 Matrix Science* MATRIX SCIENCE

Here is the report from the search. 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 database + library)

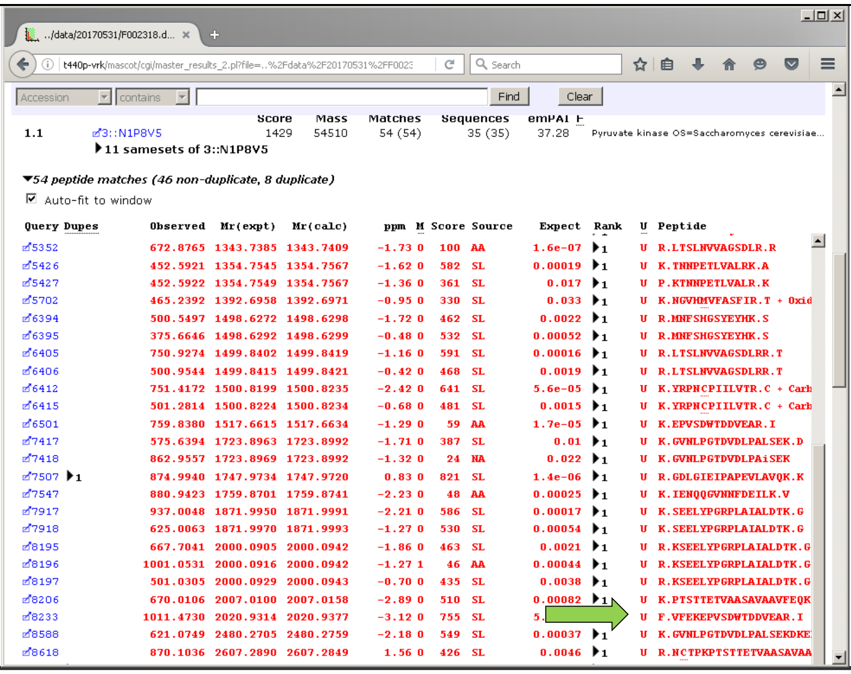
- **Protein inference**
  - Library matches are mapped to accessions from the FASTA database
  - Reference database accessions or original library annotations only used where this fails

- **Library match scores**
  - Take the set of queries where the library and FASTA database 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 : *Spectral library searching* © 2017-2023 Matrix Science MATRIX SCIENCE

For protein inference, if the peptide sequence can be mapped to one of the FASTA databases 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 database 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 database 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	F	
1.1	3::N1P8V5	1429	54510	54 (54)	35 (35)	37.28		Pyruvate kinase OS=Saccharomyces cerevisiae...

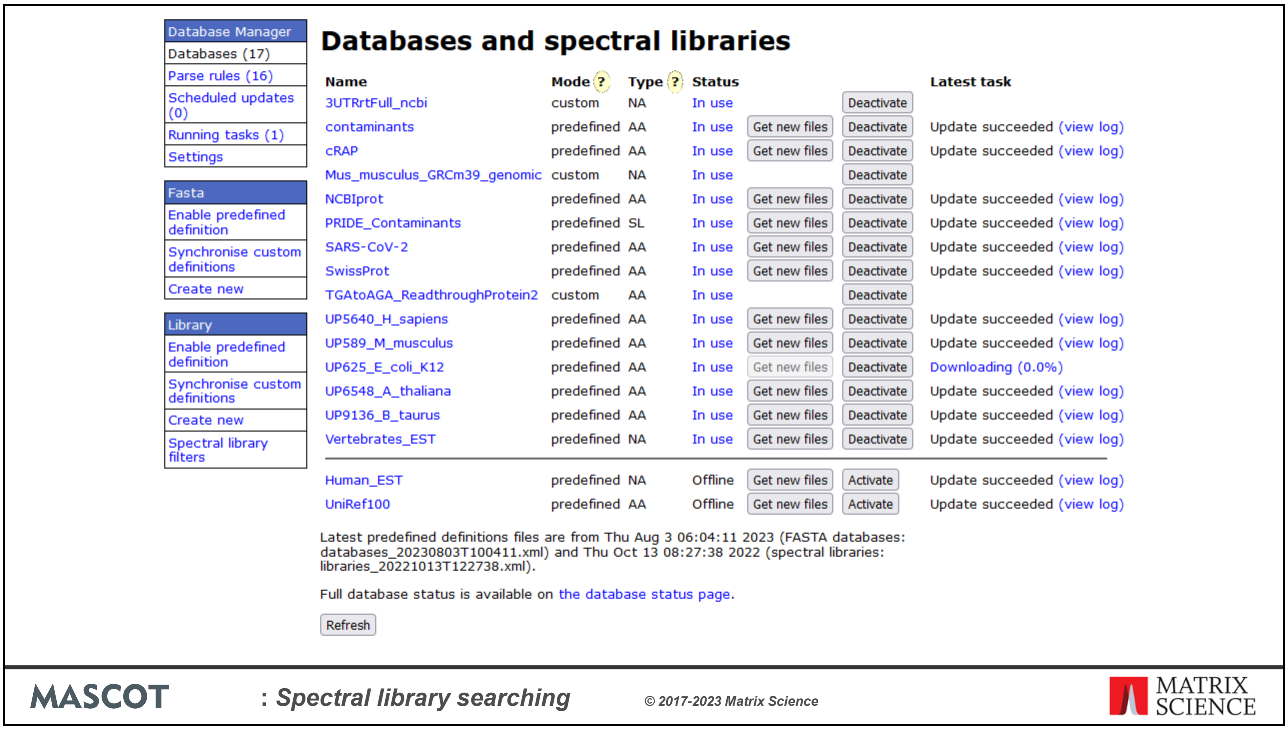
▶11 samesets of 3::N1P8V5

▼*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

Here, the top hit has been expanded. You can see that the top ranking PSMs come from both library and FASTA database. 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 database because the enzyme for the search was strict trypsin.



Databases and spectral libraries

Name	Mode ?	Type ?	Status			Latest task
3UTRrtFull_ncbi	custom	NA	In use		Deactivate	
contaminants	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
cRAP	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
Mus_musculus_GRCm39_genomic	custom	NA	In use		Deactivate	
NCBIprot	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
PRIDE_Contaminants	predefined	SL	In use	Get new files	Deactivate	Update succeeded (view log)
SARS-CoV-2	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
SwissProt	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
TGAtoAGA_ReadthroughProtein2	custom	AA	In use		Deactivate	
UP5640_H_sapiens	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
UP589_M_musculus	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
UP625_E_coli_K12	predefined	AA	In use	Get new files	Deactivate	Downloading (0.0%)
UP6548_A_thaliana	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
UP9136_B_taurus	predefined	AA	In use	Get new files	Deactivate	Update succeeded (view log)
Vertebrates_EST	predefined	NA	In use	Get new files	Deactivate	Update succeeded (view log)
Human_EST	predefined	NA	Offline	Get new files	Activate	Update succeeded (view log)
UniRef100	predefined	AA	Offline	Get new files	Activate	Update succeeded (view log)

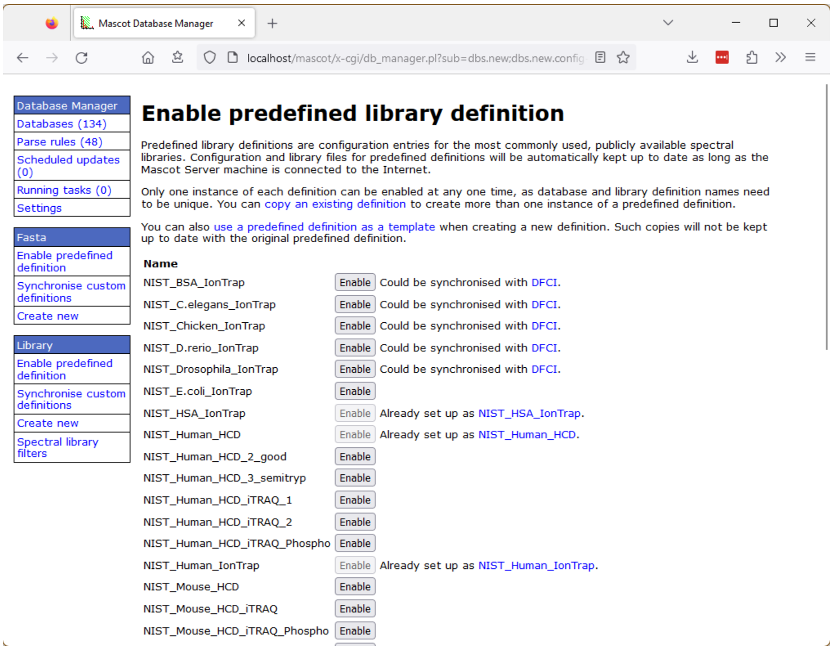
Latest predefined definitions files are from Thu Aug 3 06:04:11 2023 (FASTA databases: databases_20230803T100411.xml) and Thu Oct 13 08:27:38 2022 (spectral libraries: libraries_20221013T122738.xml).

Full database status is available on the database status page.

Refresh

MASCOT : *Spectral library searching* © 2017-2023 Matrix Science MATRIX SCIENCE

Let’s turn our attention to administration aspects. Library files in NIST 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 databases and libraries 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
localhost/mascot/x-cgi/db_manager.pl?sub=dbs.new;dbs.new.config
Database Manager
Databases (134)
Parse rules (48)
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

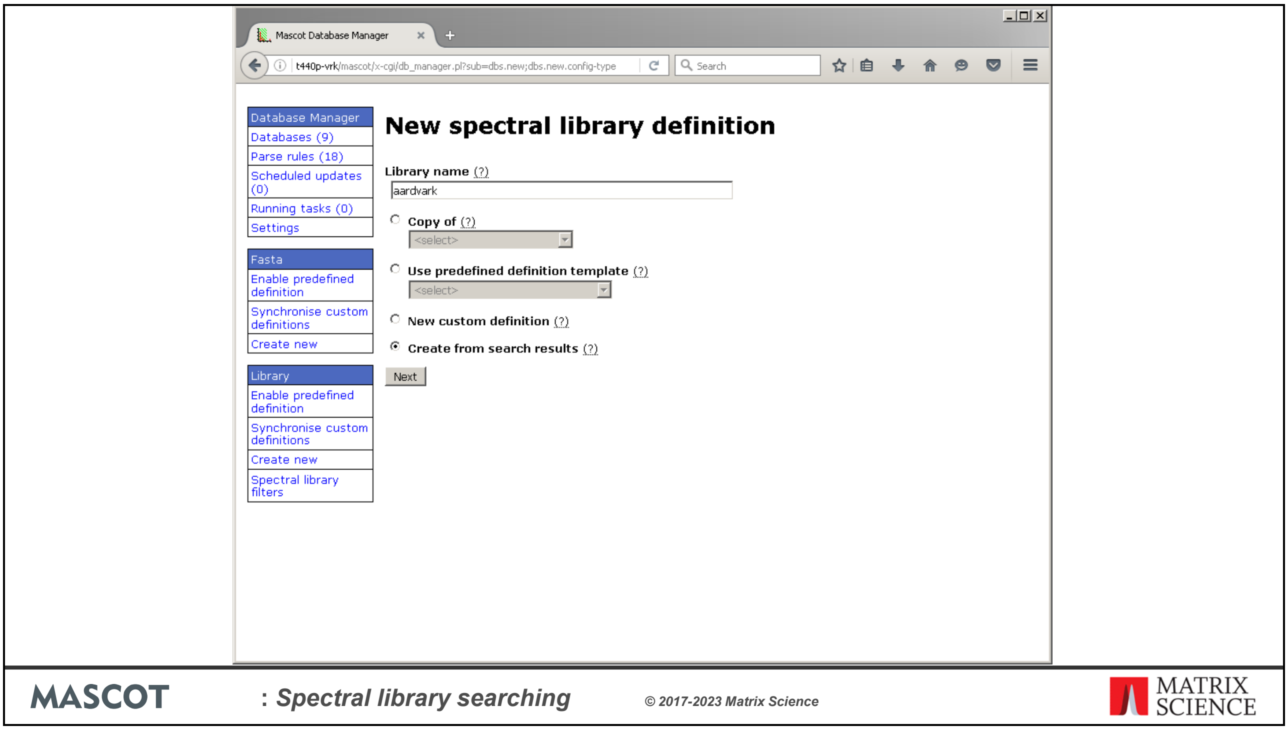
Predefined library definitions are configuration entries for the most commonly used, publicly available spectral libraries. Configuration and library files for predefined definitions will be automatically kept up to date as long as the Mascot Server machine is connected to the Internet.

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	Could be synchronised with DFCI.
NIST_C.elegans_IonTrap	Enable	Could be synchronised with DFCI.
NIST_Chicken_IonTrap	Enable	Could be synchronised with DFCI.
NIST_D.rerio_IonTrap	Enable	Could be synchronised with DFCI.
NIST_Drosophila_IonTrap	Enable	Could be synchronised with DFCI.
NIST_E.coli_IonTrap	Enable	
NIST_HSA_IonTrap	Enable	Already set up as NIST_HSA_IonTrap.
NIST_Human_HCD	Enable	Already set up as NIST_Human_HCD.
NIST_Human_HCD_2_good	Enable	
NIST_Human_HCD_3_semitryp	Enable	
NIST_Human_HCD_iTRAQ_1	Enable	
NIST_Human_HCD_iTRAQ_2	Enable	
NIST_Human_HCD_iTRAQ_Phospho	Enable	
NIST_Human_IonTrap	Enable	Already set up as NIST_Human_IonTrap.
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
Next

New spectral library definition

Library name (?)
aardvark

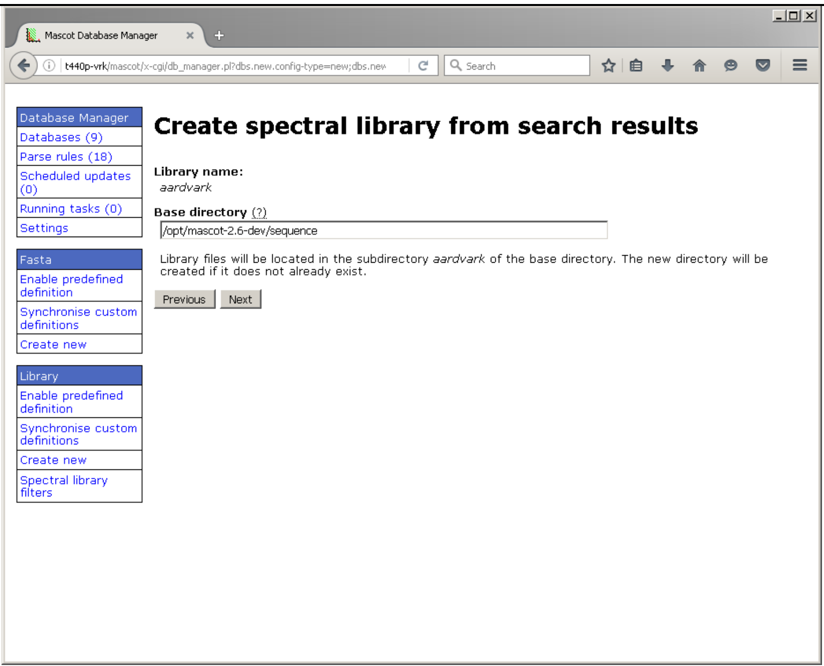
○ Copy of (?) <select>

○ Use predefined definition template (?) <select>

○ New custom definition (?)

● Create from search results (?)

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’.


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

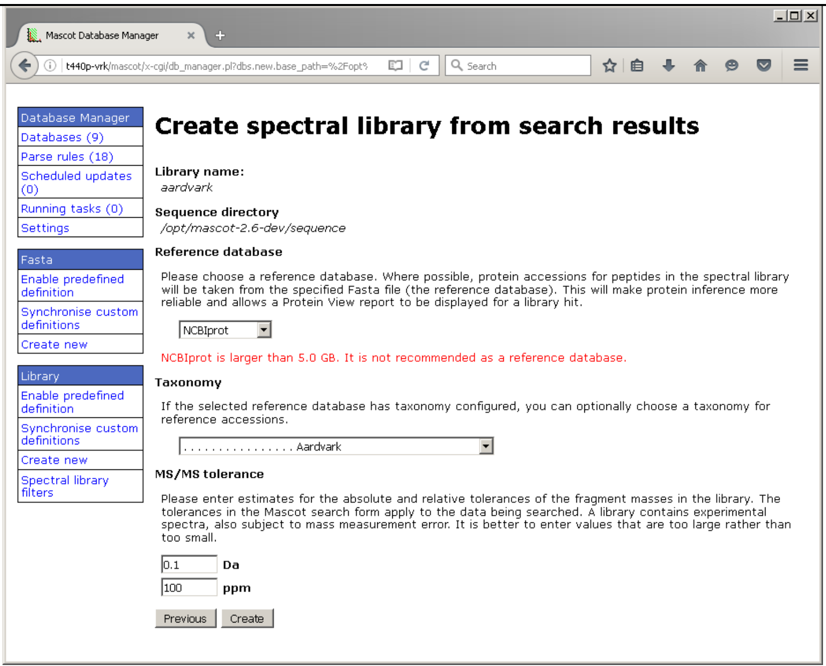
Create spectral library from search results

Library name:
aardvark

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.

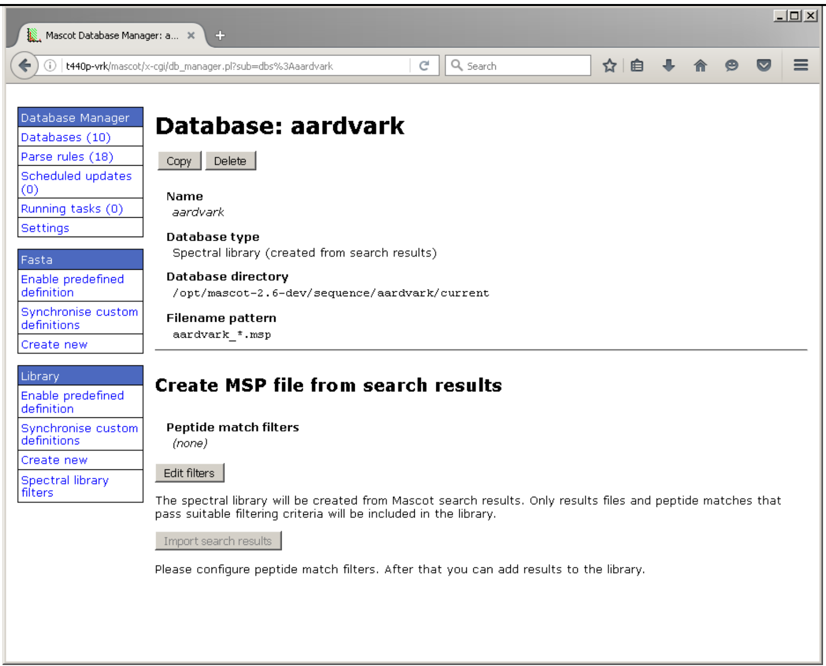
. . . . Aardvark ▾

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`

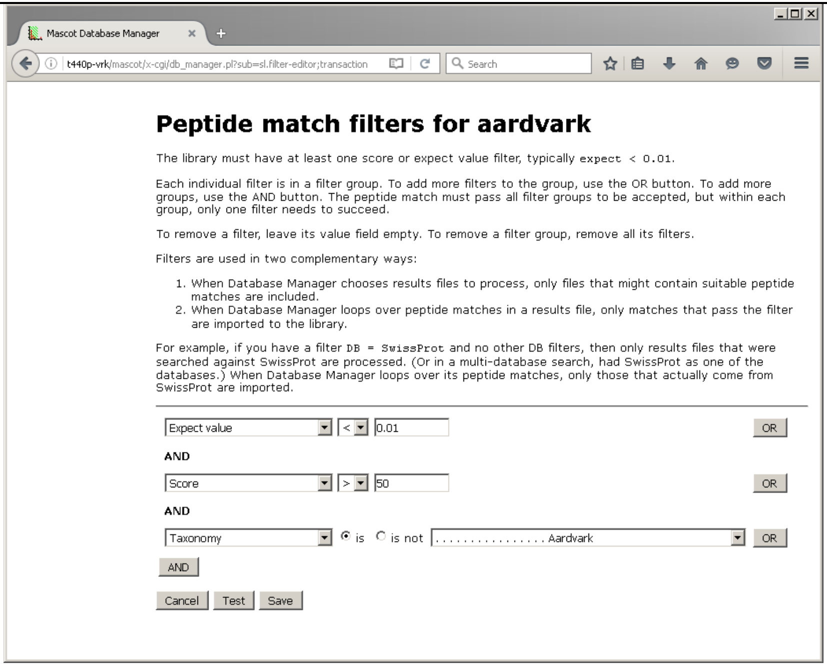
Create MSP file from search results

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 OR

AND

Score > 50 OR

AND

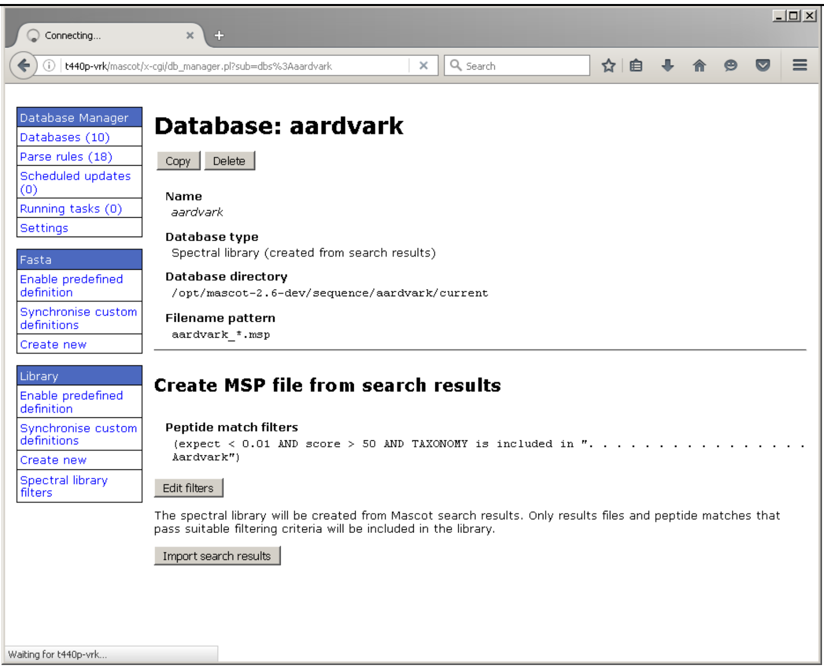
Taxonomy ● is ○ is not .... Aardvark OR

AND

Cancel Test Save

MASCOT : *Spectral library searching* *© 2017-2023 Matrix Science* MATRIX SCIENCE

There is a lot of flexibility here. 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`

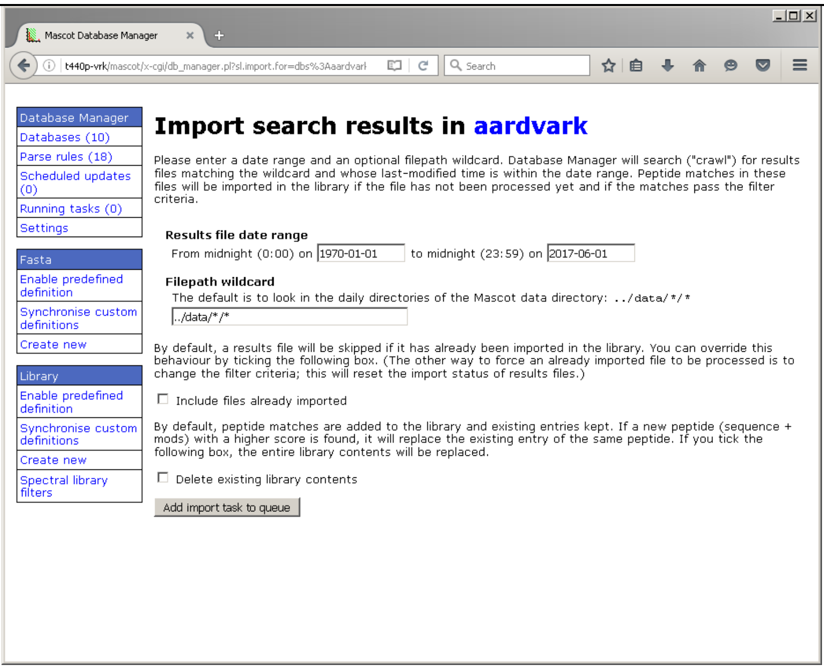
Create MSP file from search results

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.

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

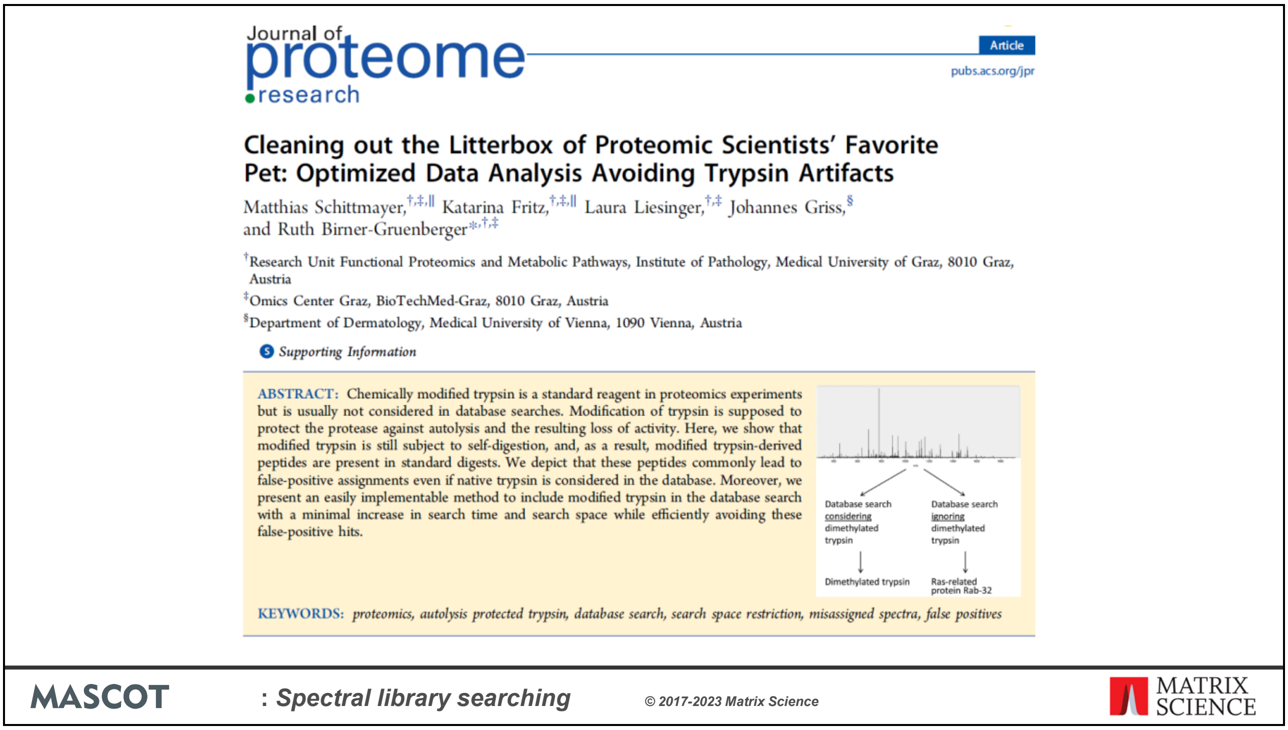
By default, peptide matches are added to the library and existing entries kept. If a new peptide (sequence + mods) with a higher score is found, it will replace the existing entry of the same peptide. If you tick the following box, the entire library contents will be replaced.

□ Delete existing library contents

Add import task to queue

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

S *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.

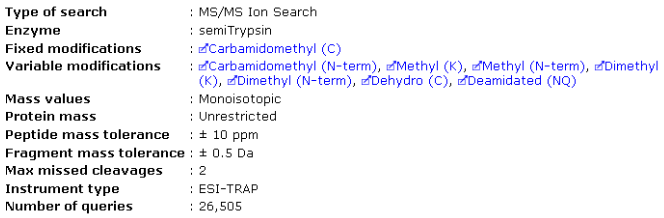
Database search considering dimethylated trypsin
Database search ignoring dimethylated trypsin
Dimethylated trypsin
Ras-related protein Rab-32

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 JPR paper 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.

Creating a trypsin library

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

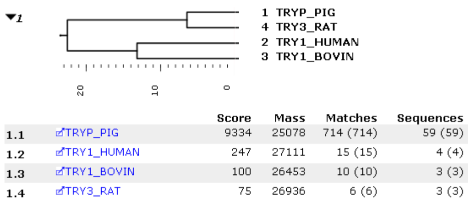


```
Type of search          : MS/MS Ion Search
Enzyme                  : semiTrypsin
Fixed modifications     : Carbamidomethyl (C)
Variable modifications  : Carbamidomethyl (N-term), Methyl (K), Methyl (N-term), Dimethyl
                          (K), Dimethyl (N-term), Dehydro (C), Deamidated (NQ)
Mass 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
```

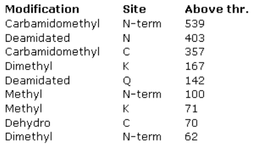
We downloaded the raw files for one of the data sets in this study from PRIDE and tried a variety of error tolerant searches to discover exactly what was present. Based on these results, we chose these search settings. The enzyme specificity was semiTrypsin because peptides show very extensive C-terminal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

Creating a trypsin library

- Large search space, low sensitivity, but many matches to Trypsin
- Import TRYP_PIG matches as new spectral library “Trypsin”



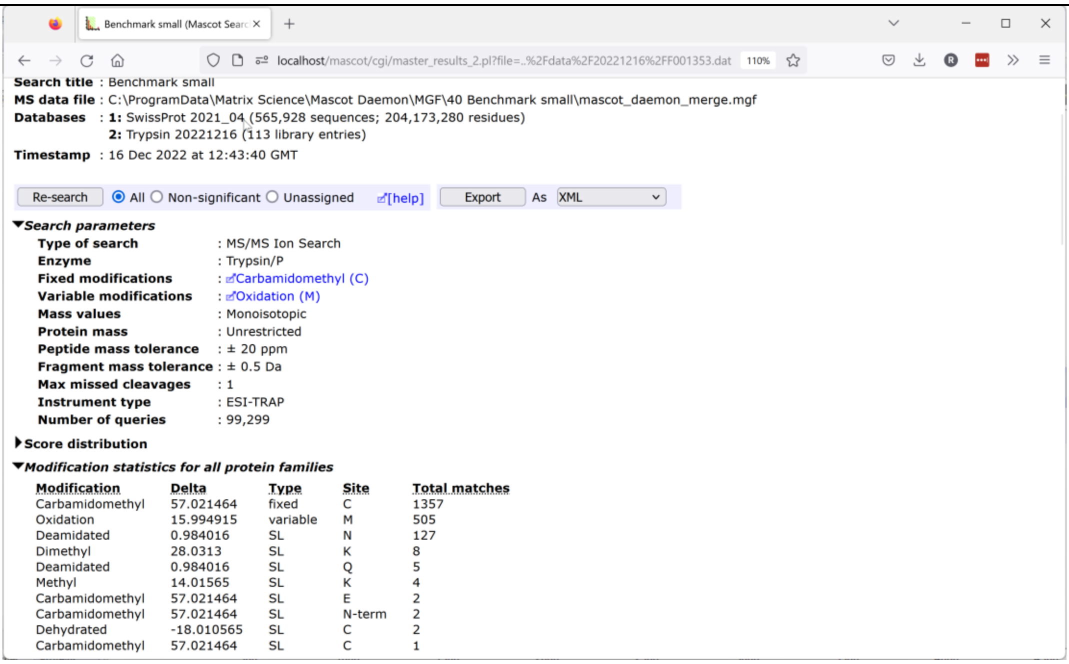
		Score	Mass	Matches	Sequences
1.1	TRYP_PIG	9334	25078	714 (714)	59 (59)
1.2	TRY1_HUMAN	247	27111	15 (15)	4 (4)
1.3	TRY1_BOVIN	100	26453	10 (10)	3 (3)
1.4	TRY3_RAT	75	26936	6 (6)	3 (3)



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

This makes the search space very large, but we do get many matches to trypsin and many modified peptides. 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.


Benchmark small (Mascot Searc
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20221216%2FF001353.dat
110%

Search title : Benchmark small
MS data file : C:\ProgramData\Matrix Science\Mascot Daemon\MGF\40 Benchmark small\mascot_daemon_merge.mgf
Databases : **1:** SwissProt 2021_04 (565,928 sequences; 204,173,280 residues)
2: Trypsin 20221216 (113 library entries)
Timestamp : 16 Dec 2022 at 12:43:40 GMT

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

Search parameters

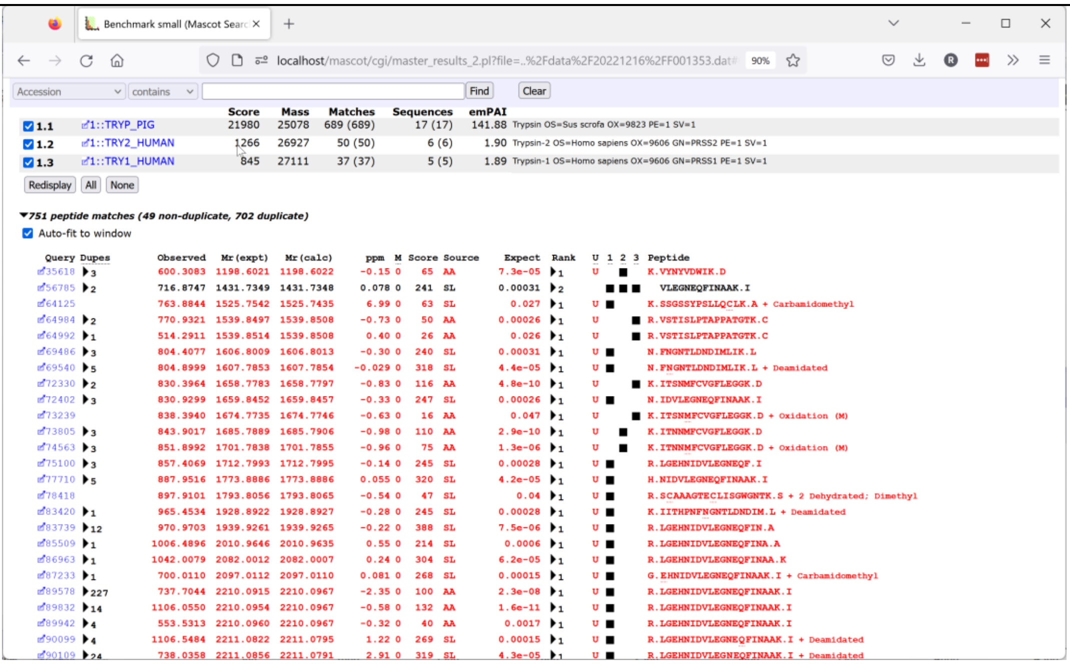
Type of search : MS/MS Ion Search
Enzyme : Trypsin/P
Fixed modifications : Carbamidomethyl (C)
Variable modifications : Oxidation (M)
Mass values : Monoisotopic
Protein mass : Unrestricted
Peptide mass tolerance : ± 20 ppm
Fragment mass tolerance : ± 0.5 Da
Max missed cleavages : 1
Instrument type : ESI-TRAP
Number of queries : 99,299

Score distribution

Modification statistics for all protein families

Modification	Delta	Type	Site	Total matches
Carbamidomethyl	57.021464	fixed	C	1357
Oxidation	15.994915	variable	M	505
Deamidated	0.984016	SL	N	127
Dimethyl	28.0313	SL	K	8
Deamidated	0.984016	SL	Q	5
Methyl	14.01565	SL	K	4
Carbamidomethyl	57.021464	SL	E	2
Carbamidomethyl	57.021464	SL	N-term	2
Dehydrated	-18.010565	SL	C	2
Carbamidomethyl	57.021464	SL	C	1

Here, we search Swissprot 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.


Benchmark small (Mascot Searc
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20221216%2FF001353.dat
90%
Accession
contains
Find
Clear
Redisplay
All
None

			Score	Mass	Matches	Sequences	emPAI	
☑	1.1	1::TRYP_PIG	21980	25078	689 (689)	17 (17)	141.88	Trypsin OS=Sus scrofa OX=9823 PE=1 SV=1
☑	1.2	1::TRY2_HUMAN	1266	26927	50 (50)	6 (6)	1.90	Trypsin-2 OS=Homo sapiens OX=9606 GN=PRSS2 PE=1 SV=1
☑	1.3	1::TRY1_HUMAN	845	27111	37 (37)	5 (5)	1.89	Trypsin-1 OS=Homo sapiens OX=9606 GN=PRSS1 PE=1 SV=1

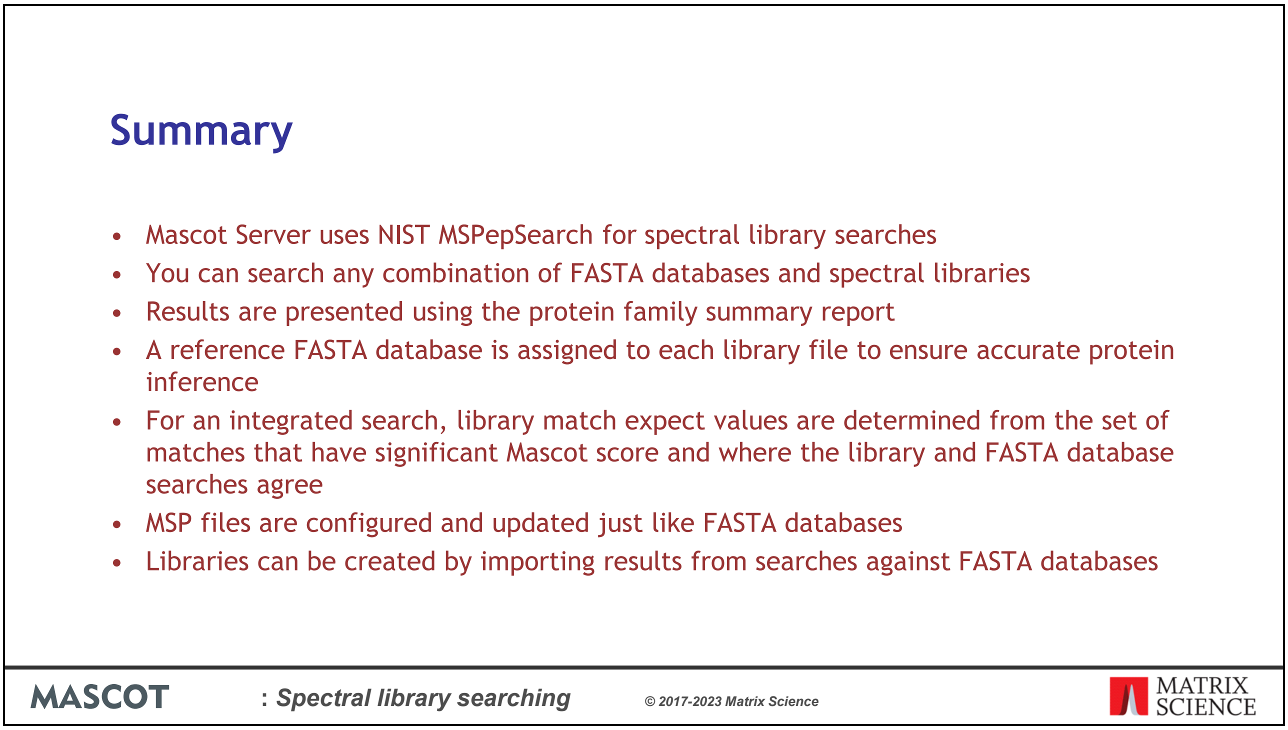
▼751 peptide matches (49 non-duplicate, 702 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Source	Expect	Rank	U	1	2	3	Peptide
35618	▶3	600.3083	1198.6021	1198.6022	-0.15	0	65	AA	7.3e-05	▶1	U		■		K.VYNYVDWIK.D
56785	▶2	716.8747	1431.7349	1431.7348	0.078	0	241	SL	0.00031	▶2		■	■	■	VLEGNEQFINAAK.I
64125		763.8844	1525.7542	1525.7435	6.99	0	63	SL	0.027	▶1	U	■			K.SSGSSYPSLLQCLK.A + Carbamidomethyl
64984	▶2	770.9321	1539.8497	1539.8508	-0.73	0	50	AA	0.00026	▶1	U			■	R.VSTISLPTAPPATGTK.C
64992	▶1	514.2911	1539.8514	1539.8508	0.40	0	26	AA	0.026	▶1	U			■	R.VSTISLPTAPPATGTK.C
69486	▶3	804.4077	1606.8009	1606.8013	-0.30	0	240	SL	0.00031	▶1	U	■			N.FNGNTLDNDIMLIK.L
69540	▶5	804.8999	1607.7853	1607.7854	-0.029	0	318	SL	4.4e-05	▶1	U	■			N.FNGNTLDNDIMLIK.L + Deamidated
72330	▶2	830.3964	1658.7783	1658.7797	-0.83	0	116	AA	4.8e-10	▶1	U			■	K.ITSNMFCVGFLEGGK.D
72402	▶3	830.9299	1659.8452	1659.8457	-0.33	0	247	SL	0.00026	▶1	U	■			N.IDVLEGNEQFINAAK.I
73239		838.3940	1674.7735	1674.7746	-0.63	0	16	AA	0.047	▶1	U			■	K.ITSNMFCVGFLEGGK.D + Oxidation (M)
73805	▶3	843.9017	1685.7889	1685.7906	-0.98	0	110	AA	2.9e-10	▶1	U		■		K.ITNNMFCVGFLEGGK.D
74563	▶3	851.8992	1701.7838	1701.7855	-0.96	0	75	AA	1.3e-06	▶1	U		■		K.ITNNMFCVGFLEGGK.D + Oxidation (M)
75100	▶3	857.4069	1712.7993	1712.7995	-0.14	0	245	SL	0.00028	▶1	U	■			R.LGEHNIDVLEGNEQF.I
77710	▶5	887.9516	1773.8886	1773.8886	0.055	0	320	SL	4.2e-05	▶1	U	■			H.NIDVLEGNEQFINAAK.I
78418		897.9101	1793.8056	1793.8065	-0.54	0	47	SL	0.04	▶1	U	■			R.SCAAAGTECLISGWGNTK.S + 2 Dehydrated; Dimethyl
83420	▶1	965.4534	1928.8922	1928.8927	-0.28	0	245	SL	0.00028	▶1	U	■			K.IITHPNFNGNTLDNDIM.L + Deamidated
83739	▶12	970.9703	1939.9261	1939.9265	-0.22	0	388	SL	7.5e-06	▶1	U	■			R.LGEHNIDVLEGNEQFIN.A
85509	▶1	1006.4896	2010.9646	2010.9635	0.55	0	214	SL	0.0006	▶1	U	■			R.LGEHNIDVLEGNEQFINA.A
86963	▶1	1042.0079	2082.0012	2082.0007	0.24	0	304	SL	6.2e-05	▶1	U	■			R.LGEHNIDVLEGNEQFINAA.K
87233	▶1	700.0110	2097.0112	2097.0110	0.081	0	268	SL	0.00015	▶1	U	■			G.EHNIDVLEGNEQFINAAK.I + Carbamidomethyl
89578	▶227	737.7044	2210.0915	2210.0967	-2.35	0	100	AA	2.3e-08	▶1	U	■			R.LGEHNIDVLEGNEQFINAAK.I
89832	▶14	1106.0550	2210.0954	2210.0967	-0.58	0	132	AA	1.6e-11	▶1	U	■			R.LGEHNIDVLEGNEQFINAAK.I
89942	▶4	553.5313	2210.0960	2210.0967	-0.32	0	40	AA	0.0017	▶1	U	■			R.LGEHNIDVLEGNEQFINAAK.I
90099	▶4	1106.5484	2211.0822	2211.0795	1.22	0	269	SL	0.00015	▶1	U	■			R.LGEHNIDVLEGNEQFINAAK.I + Deamidated
90109	▶24	738.0358	2211.0856	2211.0791	2.91	0	319	SL	4.3e-05	▶1	U	■			R.LGEHNIDVLEGNEQFINAAK.I + Deamidated

This removes 776 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 uses NIST MSPepSearch for spectral library searches
- You can search any combination of FASTA databases and spectral libraries
- Results are presented using the protein family summary report
- A reference FASTA database 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 database searches agree
- MSP files are configured and updated just like FASTA databases
- Libraries can be created by importing results from searches against FASTA databases

To summarise.