

Using spectral libraries to pick a peck of pickled peptides

Ville Koskinen
Matrix Science



Benefits of a spectral library of contaminants

- **Real-life samples contain contaminants**
 - BSA, trypsin autolysis, human keratins, ...
 - Artefactual mods
 - Non-specific cleavage
 - Sample or protocol specific
- **Hard to match these in database search without destroying sensitivity**
- **But if correct match is unavailable, high quality spectra become false positives**

Using spectral libraries to pick a peck of pickled peptides



We added spectral library support in Mascot 2.6. I'll show you how to make good use of the feature by creating a spectral library for contaminants. The topic is wider than pickled foods, but of course you could make use of spectral libraries in food science as well. A peck of peptides would be around 9 litres or two gallons in case you're wondering.

The goal is to have a contaminants library that you can search together with a protein sequence database. Because the library is prepared in advance, it isn't subject to search space restrictions. This means it can contain highly modified or non-specific peptides, for example, without having to increase the database search space.

Why do we care about matching contaminants? Because high-quality contaminant spectra can otherwise become false positive matches in the target database.

Using an older version of Mascot?

- **20% discount offer for 20th anniversary**
 - Order a Mascot Server update from any earlier version to 2.6
 - Offer lasts until 31st July 2018
 - All updates come with Premium Support
- **Visit us at booth 522**

Using spectral libraries to pick a peck of pickled peptides



By the way, if you're on an earlier version and want to try spectral libraries in Mascot, now is a great time to update! 2018 is the 20th anniversary year of Matrix Science Ltd, and to celebrate, we're offering 20% off on all Mascot Server updates. Come talk to us at booth 522 after the presentation and we're happy to send you a quote immediately.

How to create a contaminants library

- **Procedure:**
 1. Identify major contaminant proteins
 2. Search these with many sets of mods
 3. Import high-confidence contaminant matches in the library
- **Works anywhere**
- **Library spectra will be specific to your lab/sample/protocol**

Using spectral libraries to pick a peck of pickled peptides



The library creation procedure is straightforward. First identify major contaminant proteins in your sample using a standard database search. Create a temporary database of the contaminant sequences and search them again using wider search parameters. You may want to do an error tolerant search or include uncommon or suspected modifications, or use a semi-specific enzyme or increase the number of missed cleavages. The point is to identify spectra that you wouldn't normally find for fear of increasing the search space too much. Once you have a set of high-confidence matches to contaminants, import the spectra in a library.

Of course the devil is in the details. The procedure works anywhere, but the resulting library cannot help but be specific to your lab or sample or protocol. There are so many sources of experimental variability: the contaminants you see may depend on the alkylating agent, the protease, sample to substrate ratio, digestion temperature and duration, additional derivatisation steps, isotopic labelling, and so on.

Data set

- **Mycobacterium tuberculosis L7-35 whole-cell lysate (2017)**
 - PRIDE: PXD006117
 - doi:10.3389/fmicb.2017.00795
- **Biological replicates 1 and 3: in-gel tryptic digestion in 6 fractions, each with 2-3 replicate runs of LC-MS/MS**
- **Peak pick raw files in Mascot Distiller**

Using spectral libraries to pick a peck of pickled peptides



For demonstration purposes, I'll use a whole-cell lysate of a strain of *Mycobacterium tuberculosis*. This is from a study published in 2017. The data set is on PRIDE. There are three biological replicates of the whole-cell lysate for this strain. It's an in-gel tryptic digest, where the gel lane is divided in six fractions. Each fraction in turn has 2-3 replicate runs of LC-MS/MS using a Thermo Q-Exactive. We'll build the spectral library based on biological replicate 1 and then try it out on replicate 3.

Step 1: identify major contaminants

- **Determine suitable search parameters for target proteins at 1% FDR**
 - Precursor tolerance 20ppm
 - MS/MS tolerance 0.05Da
 - Trypsin, one missed cleavage
 - Carbamidomethyl (C) as fixed mod
 - No variable mods
 - Instrument ESI-TRAP
 - SwissProt (taxonomy=Mycobacterium)

Using spectral libraries to pick a peck of pickled peptides



First step is to identify the major contaminants. Suitable search parameters can be found by experience or trial and error. The instrument was a Thermo Q-Exactive. I've not included any variable mods at this stage, because the goal is to calibrate the mass tolerances and enzyme settings. The correct parameters are ones that give the highest sensitivity at 1% FDR.

For some reason, the authors used MS/MS tolerance of 0.5Da, but the data is an order of magnitude more accurate.

We'll search against SwissProt with Mycobacterium taxonomy, mainly for the convenience of this presentation. In real life, you would use a complete proteome or a genome database for the strain being analysed.

Step 1: identify major contaminants

- **Search again to find contaminants**
 - Carbamidomethyl (C) as *variable* mod
 - Whole SwissProt
- **Set FDR to 1%**
- **Report Builder: num. of significant unique sequences > 2**
- **Filter out accessions containing “_MYC”**

Using spectral libraries to pick a peck of pickled peptides



Repeat the search, but this time set all fixed mods as variable. Potential contaminants are not necessarily alkylated, for example. Search against the whole of SwissProt to identify all potential contaminants. We don't expect to find exotic contaminants in this sample.

After the search is done, set FDR to 1%. Go to Report Builder and filter out the really minor contaminants by requiring matches to at least two significant sequences. Filter out any protein accessions from tuberculosis. The easiest filter is underscore-MYC.

▼ **Sensitivity and FDR (reversed protein sequences)**

SwissProt Decoy FDR

PSMs 21873 218 1.00% Adjust to

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

[Proteins \(974\)](#) [Report Builder](#) [Unassigned \(97171\)](#) [s_permalink](#)

Protein hits (434 proteins)

► Columns (9 out of 16)

► Filters: "Num. of significant unique sequences" > 2

[Export as CSV](#)

#	Family	M	DB	Accession	Score	Mass	Match(sig)	Seq(sig)	Description
1		1	SwissProt	CH601_MYCBP	27823	56692	697	28	60 kDa chaperonin 1 OS=Mycobacterium bov
2		1	SwissProt	ALBU_BOVIN	21270	69248	752	29	Serum albumin OS=Bos taurus GN=ALB PE=1
3		1	SwissProt	LLDD_MYCTO	20494	45313	644	21	Putative L-lactate dehydrogenase OS=Mycot
4		1	SwissProt	EFTU_MYCBO	16142	43566	450	20	Elongation factor Tu OS=Mycobacterium bov
5		1	SwissProt	DNAK_MYCBO	14642	66790	444	24	Chaperone protein DnaK OS=Mycobacterium
6		1	SwissProt	HBHA_MYCBO	13350	21522	300	9	Heparin-binding hemagglutinin OS=Mycobact
7		1	SwissProt	METE_MYCBO	12555	81531	409	31	5-methyltetrahydropteroyltriglutamate--hom
8		1	SwissProt	ETFA_MYCTO	10562	31672	308	16	Electron transfer flavoprotein subunit alpha C
9		1	SwissProt	FAB2_MYCBO	9970	44250	268	15	3-oxoacyl-[acyl-carrier-protein] synthase 2
9		2	SwissProt	FAB1_MYCBO	4523	43289	139	13	3-oxoacyl-[acyl-carrier-protein] synthase 1
9		3	SwissProt	FAB1_MYCLE	2084	43443	45	6	3-oxoacyl-[acyl-carrier-protein] synthase 1
10		1	SwissProt	GLNA1_MYCBO	9090	53536	274	17	Glutamine synthetase 1 OS=Mycobacterium l
11		1	SwissProt	DBH_MYCBO	6634	22174	289	9	DNA-binding protein HU homolog OS=Mycoba
12		1	SwissProt	ATPB_MYCBO	6389	53061	259	24	ATP synthase subunit beta OS=Mycobacteriu
13		1	SwissProt	Y0148_MYCTU	6326	29760	159	10	Putative short-chain type dehydrogenase/re
14		1	SwissProt	SERA_MYCBO	6323	54521	172	19	D-3-phosphoglycerate dehydrogenase OS=M

Using spectral libraries to pick a peck of pickled peptides



Here's the list of protein hits against all of SwissProt at 1% FDR. Most of the accessions are from various mycobacteria. Their accessions end with MYC-something. We can easily filter them out.

Preferred taxonomy: All entries

▼ **Sensitivity and FDR (reversed protein sequences)**

SwissProt Decoy FDR
 PSMs: above homology 21873 218 1.00% Adjust to 1%

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

[Proteins \(974\)](#) Report Builder [Unassigned \(97171\)](#) [S permalink](#)

Protein hits (9 proteins)

► Columns (9 out of 16)

► Filters: (NOT(Accession CONTAINS "_MYC") AND "Num. of significant unique sequences" > 2)

Export as CSV

Family	M	DB	Accession	Score	Mass	Match(sig)	Seq(sig)	Description
2	1	SwissProt	ALBU_BOVIN	21270	69248	752	29	Serum albumin OS=Bos taurus GN=ALB PE=1 S
31	2	SwissProt	RPOC_LEIXX	618	142285	20	4	DNA-directed RNA polymerase subunit beta' O
51	1	SwissProt	K2C1_HUMAN	2349	65999	96	13	Keratin, type II cytoskeletal 1 OS=Homo sapi
51	2	SwissProt	K2C6B_HUMAN	553	60030	33	7	Keratin, type II cytoskeletal 6B OS=Homo sap
51	3	SwissProt	K22E_HUMAN	365	65393	26	8	Keratin, type II cytoskeletal 2 epidermal OS=
71	1	SwissProt	K1C10_HUMAN	1793	58792	56	12	Keratin, type I cytoskeletal 10 OS=Homo sapi
75	1	SwissProt	TTHY_BOVIN	1732	15717	43	3	Transthyretin OS=Bos taurus GN=TTR PE=1 S
98	1	SwissProt	K1C9_HUMAN	1457	62027	52	11	Keratin, type I cytoskeletal 9 OS=Homo sapi
294	1	SwissProt	TRYP_PIG	423	24394	27	3	Trypsin OS=Sus scrofa PE=1 SV=1

Export as CSV

Not what you expected? Try [the select summary](#).

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

Using spectral libraries to pick a peck of pickled peptides



This leaves a list of nine proteins. The hit in family 31 is similar to an RNA polymerase subunit from *M. tuberculosis*, so it's not a contaminant. The eight proteins remaining are the usual suspects: trypsin, bovine serum albumin, human keratins.

Step 1: identify major contaminants

- **Create a new FASTA file:**
 - Click on Protein View for each contaminant
 - Copy sequence, paste in file
- **Create an auxiliary contaminants database in Database Manager**

Using spectral libraries to pick a peck of pickled peptides



Now that we have a list of potential contaminants, let's make a FASTA file out of them. It's not strictly necessary but will make subsequent searches and filtering easier. Since there are only 8 contaminant proteins, we can copy and paste the protein sequences from Protein View in a new file.

Preferred taxonomy: All entries

▼ **Sensitivity and FDR (reversed protein sequences)**

SwissProt Decoy FDR
 PSMs: above homology 21873 218 1.00% Adjust to 1%

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

[Proteins \(974\)](#) Report Builder [Unassigned \(97171\)](#) [§ permalink](#)

Protein hits (9 proteins)

► Columns (9 out of 16)

► Filters: (NOT(Accession CONTAINS "_MYC") AND "Num. of significant unique sequences" > 2)

Export as CSV

Family	M	DB	Accession	Score	Mass	Match(sig)	Seq(sig)	Description
2	1	SwissProt	ALBU_BOVIN	21270	69248	752	29	Serum albumin OS=Bos taurus GN=ALB PE=1 S
31	2	SwissProt	RPOC_LEIXX	618	142285	20	4	DNA-directed RNA polymerase subunit beta' O
51	1	SwissProt	K2C1_HUMAN	2349	65999	96	13	Keratin, type II cytoskeletal 1 OS=Homo sapie
51	2	SwissProt	K2C6B_HUMAN	553	60030	33	7	Keratin, type II cytoskeletal 6B OS=Homo sap
51	3	SwissProt	K22E_HUMAN	365	65393	26	8	Keratin, type II cytoskeletal 2 epidermal OS=
71	1	SwissProt	K1C10_HUMAN	1793	58792	56	12	Keratin, type I cytoskeletal 10 OS=Homo sapi
75	1	SwissProt	TTHY_BOVIN	1732	15717	43	3	Transthyretin OS=Bos taurus GN=TTR PE=1 S'
98	1	SwissProt	K1C9_HUMAN	1457	62027	52	11	Keratin, type I cytoskeletal 9 OS=Homo sapie
294	1	SwissProt	TRYP_PIG	423	24394	27	3	Trypsin OS=Sus scrofa PE=1 SV=1

Export as CSV

Not what you expected? Try [the select summary](#).

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

Using spectral libraries to pick a peck of pickled peptides



I'll quickly show you how to harvest the sequences. Click on the protein in the accession column...

MATRIX SCIENCE MASCOT Search Results

Protein View: TRYP_PIG

Trypsin OS=Sus scrofa PE=1 SV=1

Database: SwissProt
 Score: 423
 Monoisotopic mass (M_r): 24394
 Calculated pI: 7.00
 Taxonomy: [Sus scrofa](#)

Sequence similarity is available as [an NCBI BLAST search of TRYP_PIG against nr](#).

Search parameters

MS data file: L7_35.mgf
 Enzyme: Trypsin: cuts C-term side of KR unless next residue is P.
 Variable modifications: [Carbamidomethyl \(C\)](#)

Protein sequence coverage: 16%

Matched peptides shown in **bold red**.

```

1  FPTDDDDKIV GGYTCAANSI PYQVSLNSGS HFCGGSLINS QWVVSAAHCY
51  KSRIQVRLGE HNIDVLEGNEPINAAKIIT HPNFNGNTLD NDIMLIKLSS
101 PATLNSRVAT VSLPRSAAA GTECLISGWG NTKSSGSSYP SLLQCLKAPV
151 LSDSSCKSSY PGQI SC VGFLEGGKDS CQGDGGGPVV CNGQLQGIVS
201 WGYGCAQRNK PGVYTR S NWIQQTIAA N
  
```

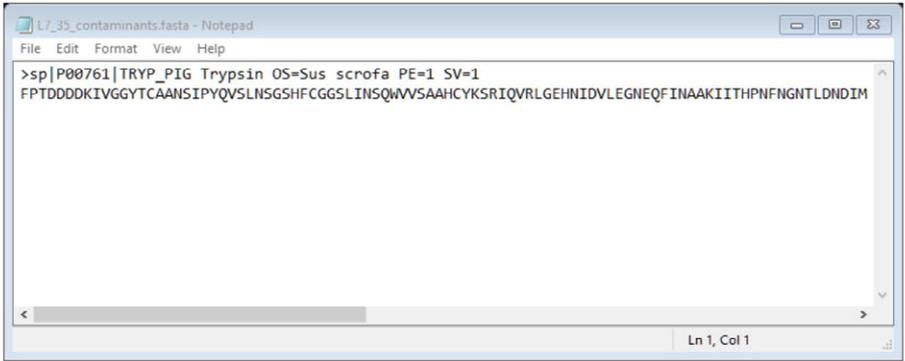
Unformatted sequence string: [231 residues](#) (for pasting into other applications).

Sort by residue number increasing mass decreasing mass
 Show matched peptides only predicted peptides also

Using spectral libraries to pick a peck of pickled peptides 

Here's porcine trypsin. Click on the link to unformatted sequence string.

```
*FPTDDDDKIVGGYTCAANSIPYQVSLNSGSHFCGGSLINSQWVWSAAHCYKSRIQVRLGEHNIDVLEGNEQFINAAKIITHPNFNGNTLDNDIMLIKLSPPATLNSRVATV
SLPRSCAAAGTECLISGWNTKSSGSSYPSSLQCLKAPVLSDSCKSSYPGQITGNMICVGFLEGGKDSQCQSDSGGPVVCNGLQGGIVSWGYGCAQKNKPGVYTKVCMYVNW
IQQTIAAN
>sp|P00761|TRYP_PIG Trypsin OS=Sus scrofa PE=1 SV=1
```



=> L7_35_contaminants in Database Manager

Using spectral libraries to pick a peck of pickled peptides



Here's the unformatted sequence string. Copy-paste into your text editor, swap the lines and remove the leading asterisk from the sequence. Rinse and repeat for the seven other contaminant proteins. This only takes a few minutes.

Creating the database is easy: create new in Database Manager, upload the FASTA file, choose the recommended parse rules and you're done.

Step 2: search wide

- **First error tolerant search**
 - Trypsin/P, 2 missed cleavages
 - Carbamidomethyl (C) as variable mod
 - L7_35_contaminants + SwissProt (Mycobacterium)

Using spectral libraries to pick a peck of pickled peptides



Now we come to the fun part. Let's see what modifications the contaminants carry. Search only the contaminant proteins from the previous step. The purpose is find the most abundant modifications that need to specified as fixed and variable. Allow for more missed cleavages as well.

Step 2: search wide

In the 49k significant PSMs identified:

Modification	Site	Above thr.	ET	Total
Oxidation	M	0	3946	3946
Carbamidomethyl	C	0	3247	3247
Iodo	Y	0	1514	1514
Delta:H(2)C(2)	N-term	0	1080	1080
Label:15N(1)	V	0	998	998
Non-specific cleavage	-	0	569	569
Deamidated	N	0	505	505

(+ more with < 500 instances)

Acetaldehyde adduct?

Deamidation nearby (or 13C)

Using spectral libraries to pick a peck of pickled peptides



The error tolerant search finds about 49000 matches in total in both the contaminants and tuberculosis proteins. Here are the top modifications.

There's definitely oxidised methionine in this sample, and you might as well set carbamidomethylation as fixed mod. Looks reasonable to choose iodination of tyrosine as a variable mod.

The 15N label is hard to believe, since there's no such labelling in the sample. Many of the sequences with this delta have an asparagine or glutamine next to or near the modified residue, so this is more likely to be deamidation – which also appears near the top of the list. Another possibility is that peak detection has chosen the 13C peak rather than 12C.

The more surprising item is Delta:H(2)C(2). This is an acetaldehyde adduct that normally modifies lysine. In this search, it's only found on the N-terminus.

Step 2: search wide

- **Second ET search**
 - Trypsin/P, 2 missed cleavages
 - Fixed: Carbamidomethyl (C)
 - Variable: Oxidation (M), Iodo (Y), Delta:H(2)C(2) (N-term), Deamidation (NQ)
 - #13C=0
 - L7_35_contaminants + SwissProt (Mycobacterium)

Using spectral libraries to pick a peck of pickled peptides



Do a second ET search. This time, choose Carbamidomethyl as a fixed mod and these four as variable mods. This will show which mods occur in combination with the most abundant ones. It will also either confirm or disprove the presence of the acetaldehyde adduct.

I've chosen deamidation rather than the possibility of carbon isotopes. A quick search with 13C set to 1 or 2 found hardly any new matches in the contaminants, so deamidation is the more likely candidate.

Step 2: search wide

In the 47k significant PSMs identified:

- **Delta:H(2)C(2) with +15.99 Da at N-term**
 - Total delta (42.01 Da) = Acetyl (N-term)
- **Diiodo (Y) is plausible**
- **Methyl/dimethyl in trypsin**
- **Semi-specific trypsin autolysis?**
 - Hundreds of differently modified matches to
LGEHNIDVLEGNEQFINAAKIITHPNFNGNTLDNDIMLIK
LSSPATLNSR and its subsequences

Using spectral libraries to pick a peck of pickled peptides



When the search is done, it's important to inspect the high-scoring error tolerant matches in contaminants, especially those with both variable modifications and an error tolerant modification. The error tolerant delta can cancel out the variable modification, or they can sum up to a different, more common variable modification.

The acetaldehyde adduct often occurs in combination with another error tolerant delta of 15.99 Daltons at the N-terminus or N-terminal residue. The combined delta is exactly the same as an acetylated N-terminus, which is a far more common variable modification.

There's clearly iodination in the sample, and diiodo appears near the top of the list, so it's reasonable to include it.

Peptide matches in trypsin show signs of methylation or dimethylation. This isn't surprising, since trypsin is frequently methylated to prevent autolysis. Interestingly, there are hundreds of low-scoring matches to this long peptide and its subsequences, using various permutations of modifications, which suggests semi-specific autolysis.

Step 2: search wide

- **Decoy search base settings**
 - semiTrypsin, 2 missed cleavages
 - Fixed: Carbamidomethyl (C)
 - Variable: Oxidation (M), Deamidation (NQ)
 - L7_35_contaminants + SwissProt (Mycobacterium)
- 1. **Base + Iodo (Y), Acetyl (N-term)**
- 2. **Base + Diiodo (Y), Acetyl (N-term)**
- 3. **Base + Methyl (K), Methyl (N-term)**
- 4. **Base + Dimethyl (K), Dimethyl (N-term)**

Using spectral libraries to pick a peck of pickled peptides



Now we can create the final peptide matches to import in the library. The library import tool will refuse to import error tolerant matches, because there's no way to tell whether the match is "real". It's important to run a decoy search with your chosen set of modifications, so that you can exert quality control.

In real life, you might want to run these searches separately for each contaminant. In the interests of time, I'll do all contaminants together. We'll use semiTrypsin to account for autolysis, and these modifications as the base settings.

We'll run four separate searches with different combinations of modifications. The first two are mainly for iodination in serum albumin, and the next two for methylated lysines in trypsin. We don't expect to see a methyl and a dimethyl on the same peptide, for example, or both iodination and diiodination, so there's no point in putting all the variable mods in one big search.

Step 3: import in spectral library

- **Create a custom spectral library in Database Manager**
- **Specify import filters for matches**
 - Only matches with low expect or high score, matching in contaminants DB
 - Error tolerant matches not allowed
- **Specify which files to process**

Using spectral libraries to pick a peck of pickled peptides



The final step in creating the contaminants library is of course importing the matches. This is straightforward in Database Manager. The two details to address are defining the import filters for peptide matches, and defining which results files are processed. We only want to import high-quality matches, so they should have low expect value or high score. We only want to import peptides that match in the auxiliary contaminants database; this is where it becomes handy.

The screenshot shows a web browser window with the URL `http://yabby/mascot/x-cgi/db_manager.pl?sub=newdbs.new` and the page title "Mascot Database Manager". The main content area is titled "New spectral library definition". On the left, there is a navigation menu with sections: "Database Manager" (containing Databases (42), Parse rules (25), Scheduled updates (0), Running tasks (0), and Settings), "Fasta" (containing Enable predefined definition, Synchronise custom definitions, and Create new), and "Library" (containing Enable predefined definition, Synchronise custom definitions, Create new, and Spectral library filters). The main form includes a "Library name" field with the value "L7_35_contaminants_SL". Below this are three radio button options: "Copy of" (with a dropdown menu), "Use predefined definition template" (with a dropdown menu), and "New custom definition". The "Create from search results" option is selected. A "Next" button is located at the bottom of the form.

Using spectral libraries to pick a peck of pickled peptides

Choose Library, Create New. Type in a name and choose to create from search results.

Browser address bar: http://yabby/mascot/x-cgi/db_manager.pl?dbs.new.config-type= Mascot Database Manager

Create spectral library from search results

Database Manager
Databases (42)
Parse rules (25)
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

Library name:
L7_35_contaminants_SL

Base directory (?)

Library files will be located in the subdirectory *L7_35_contaminants_SL* of the base directory. The new directory will be created if it does not already exist.

Using spectral libraries to pick a peck of pickled peptides



Click Next.

The screenshot shows the 'Mascot Database Manager' web interface. The main heading is 'Create spectral library from search results'. On the left, there is a navigation menu with options like 'Database Manager', 'Databases (42)', 'Parse rules (25)', 'Scheduled updates (0)', 'Running tasks (0)', 'Settings', 'Fasta', 'Library', and 'Spectral library filters'. The main content area contains the following form fields:

- Library name:** L7_35_contaminants_SL
- Sequence directory:** D:/inetpub/mascot/sequence
- Reference database:** SwissProt (selected from a dropdown menu)
- Taxonomy:** (none) (selected from a dropdown menu)
- MS/MS tolerance:** 0.05 Da and 20 ppm (input fields)

At the bottom of the page, there is a footer with the text 'Using spectral libraries to pick a peck of pickled peptides' and the 'MATRIX SCIENCE' logo.

We want to use SwissProt as the reference database. The fragment tolerance in this data set is a nice 20ppm, which you can approximate as 0.05Da.

Database Manager
Databases (43)
Parse rules (25)
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 filters

Database: L7_35_contaminants_SL

Copy Delete

Name
L7_35_contaminants_SL

Database type
Spectral library (created from search results)

Database directory
D:/inetpub/mascot/sequence/L7_35_contaminants_SL/current

Filename pattern
L7_35_contaminants_SL_*.msp

Create MSP file from search results

Peptide match filters
(none)

Edit filters

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.

Import search results

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

Using spectral libraries to pick a peck of pickled peptides

The library definition has been created. Click on Edit Filters.

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

Database name must must not equal L7_35_contaminants OR

AND

Search title must must not contain import this OR

AND

Cancel Test Save

Using spectral libraries to pick a peck of pickled peptides

Based on inspecting the two search reports, matches with expect value below 0.01 and score above 50 are very likely to filter out false positives. We only want matches in the auxiliary contaminants database – this is where having an auxiliary database comes in handy.

If you're very organised, you can use a search title filter here to include specific searches only. I put the string "import this" in the search title.

Database Manager
Databases (43)
Parse rules (25)
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

Database: L7_35_contaminants_SL

Copy Delete

Name
L7_35_contaminants_SL

Database type
Spectral library (created from search results)

Database directory
D:/inetpub/mascot/sequence/L7_35_contaminants_SL/current

Filename pattern
L7_35_contaminants_SL_*.msp

Create MSP file from search results

Peptide match filters
(expect < 0.01 AND score > 50 AND DB = L7_35_contaminants AND COM contains "import this")

Edit filters

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.

Import search results

Using spectral libraries to pick a peck of pickled peptides

Match filters are defined. Now click on Import.

Using spectral libraries to pick a peck of pickled peptides

MATRIX SCIENCE

Here you choose the files to process. It's a good idea to use a narrow date range if all your searches were made today or last week, to avoid having to crawl through the whole search archive.

The default wildcard chooses all searches on the Mascot Server machine. If you have files in a specific directory or with a specific name, you can define the wildcard here. Since my search titles include a special string, I don't need to change the wildcard.

You can always repeat the import step to include files in matching the filters in different directories, so importing can be cumulative.

Click on Add Import Task.

The screenshot displays the Mascot Database Manager interface. On the left, there is a sidebar with links: 'Synchronise custom definitions', 'Create new', and 'Spectral library filters'. The main content area is divided into several sections:

- Database status:** Shows 'In use' with a 'Deactivate' button.
- Scheduled updates:** Explains that updates import results files from the Mascot data directory. It notes '(no schedules defined)' and includes an 'Edit schedule' button.
- Peptide match filters:** Displays a filter expression: '(expect < 0.01 AND score > 50 AND DB = L7_35_contaminants AND COM contains "import this")' with an 'Edit filters' button.
- Audit log:** Contains a link 'Show action and task summary (all revisions)'. Two yellow arrows point to this link.
- Imported files:** Includes a link 'Show file index (revision 0)'.
- Peptide matches:** Includes a link 'Show peptide match import history (revision 0)'. A second yellow arrow points to this link.
- Most recent finished task:** Shows a task from 'Fri May 25 16:13:00 2018' (success) with the description 'Finished importing new peptide matches in L7_35_contaminants_SL'. Below it is an 'Import search results' button.

At the bottom of the page, there is a footer with the text 'Using spectral libraries to pick a peck of pickled peptides' and the Matrix Science logo.

When importing is done, you can check which files were processed and which matches imported by following these audit log links. The audit log contains the complete change history of the spectral library, including a history of the filter settings.

Validation: search replicate 3

- **Search FASTA + contaminants library**
 - Trypsin, 1 missed cleavage
 - Fixed: Carbamidomethyl (C)
 - Variable: Oxidation (M), Acetyl (N-term)
 - L7_35_contaminants_SL + SwissProt (Mycobacterium)

Using spectral libraries to pick a peck of pickled peptides



The contaminants library is ready for use. Remember we built it based on biological replicate 1. Let's search replicate 3 as a validation step. Since all the sample prep is the same, we expect to find most if not all of the same contaminants.

The database search parameters use strict trypsin and only two variable mods.

Validation: search replicate 3

Matches	
FASTA	21,367
Library	1,176
Total	22,539

Modification	Site	Above thr.
Oxidation	M	2511
Acetyl	N-term	1798
Carbamidomethyl	C	1029
Dimethyl	K	132
Deamidated	N	45
Iodo	Y	17
Methyl	K	16
Deamidated	Q	10
Dimethyl	N	9

?Family	Accession	Score	Match(sig)
1	!2::ALBU_BOVIN	35680	947
29	!2::TRYP_PIG	4461	158
95	!2::TTHY_BOVIN	1593	49
242	!2::K2C1_HUMAN	578	16
242	!2::K22E_HUMAN	61	4
940	!2::K1C10_BOVIN	9	1

Using spectral libraries to pick a peck of pickled peptides



There are a number of matches in the contaminants library. These are all high-scoring spectra that might have otherwise become false positive matches in the database search. We can find six of the eight contaminant proteins in the replicate.

Note in particular the range of modifications found. I only chose oxidation and acetylation as variable mods and carbamidomethyl as fixed, yet look at the range of mods found in the contaminants.

The last row, Dimethyl on N, is actually on an N-term residue.

Accession contains Find Clear

▼ 158 peptide matches (25 non-duplicate, 133 duplicate)

Auto-fit to window

ppm	M	Score	Source	Expect	Rank	Peptide
2.03	0	895	SL	5.6e-008	▶1	U K.LSSPATLNSR.V
-0.29	0	375	SL	0.0089	▶1	U R.LGEHNIDVLEGNQFINAAK.I + Dimethyl
-0.28	0	474	SL	0.00091	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Dimethyl
0.47	0	747	SL	1.7e-006	▶1	U R.LGEHNIDVLEGNQFIN.A
-1.01	0	333	SL	0.023	▶1	U R.LGEHNIDVLEGNQFINAA.K
0.52	0	438	SL	0.0021	▶1	U R.LGEHNIDVLEGNQFINAAK.I
1.24	0	717	SL	3.4e-006	▶1	U R.LGEHNIDVLEGNQFINAAK.I + Methyl
0.069	0	963	SL	1.2e-008	▶1	U N.TLDNDIMLIKLSPPATLNSR.V + Dimethyl
0.34	0	868	SL	1e-007	▶1	U R.LGEHNIDVLEGNQFINAAK.I + Dimethyl
1.90	0	877	SL	8.5e-008	▶1	U N.TLDNDIMLIKLSPPATLNSR.V + Dimethyl; Oxidation
0.97	0	789	SL	6.4e-007	▶1	U N.GNTLDNDIMLIKLSPPATLNSR.V + Dimethyl
-0.74	0	441	SL	0.0019	▶1	U N.FNGNTLDNDIMLIKLSPPATLNSR.V + Dimethyl
20.6	0	304	SL	0.046	▶1	U P.NFNGNTLDNDIMLIKLSPPATLNSR.V + 3 Deamidated; Dimethyl
6.04	0	395	SL	0.0056	▶1	U P.NFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; 2 Dimethyl; Oxidation
-1.85	0	731	SL	2.4e-006	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Dimethyl
0.53	0	391	SL	0.0062	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; Dimethyl
1.60	0	888	SL	6.6e-008	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; Dimethyl
2.74	0	833	SL	2.3e-007	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; Dimethyl
-2.81	0	622	SL	3e-005	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Dimethyl; Oxidation
0.63	0	708	SL	4.2e-006	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; Dimethyl; Oxidation
1.19	0	794	SL	5.7e-007	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; Dimethyl; Oxidation
2.89	0	519	SL	0.00032	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; Dimethyl; Oxidation
6.69	0	316	SL	0.035	▶1	U K.IITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + 2 Deamidated; Dimethyl; Oxidation
1.06	0	515	SL	0.00035	▶1	U N.AAKIITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; 2 Dimethyl
.0016	0	383	SL	0.0074	▶1	U N.AAKIITHPNFNGNTLDNDIMLIKLSPPATLNSR.V + Deamidated; 2 Dimethyl; Oxidation

Using spectral libraries to pick a peck of pickled peptides



As an example, here's the list of matches to trypsin. Unmodified, methylated and dimethylated trypsin autolysis products are all taken into account. There are many high-scoring semi-specific sequences in the list.

Summary

- **Creating a contaminants library:**
 1. Identify major contaminant proteins
 2. Search these with many sets of mods
 3. Import high-confidence contaminant matches in the library
- **Questions? 20% discount offer?**
- **Visit us at booth 522**

Using spectral libraries to pick a peck of pickled peptides



I've shown you the basic procedure of creating a spectral library for contaminants. The procedure works anywhere: identify major contaminants; search them with different sets of variable mods; import the results in the library.