Technical Overview

Agilent

Trusted Answers

Processing Agilent AutoMSMS Proteomic Data Using Mascot Online

Introduction

One of the key challenges of proteomics is the fast and accurate identification of proteins. This process uses spectral matching of AutoMSMS data to known protein sequences, together with an Agilent LC/Q-TOF system using Agilent MassHunter software and Mascot MS/MS Ions Search (Matrix Science). These tools enable the creation of a proteomic workflow.

Acquiring good quality data is important. Data should be checked for chromatography, mass accuracy, fragmentation, and confounding artifacts such as excessive sodium adducts or undigested protein. This can be done using Agilent MassHunter Qualitative Analysis software. The trypsin digestion protocol should be checked for robustness and reproducibility. Data should be acquired as centroid to transition smoothly to the data extraction step.

Data extraction produces a reduced data set in the form of an MS/MS peak list text file with additional information, including precursor *m/z* and retention time. The aim is to export signal and reduce noise. The size of the extracted file will affect the proteomic search time. MassHunter Qualitative Analysis software version 10.0 can be used. Vendor-neutral data extraction software is available, such as MSConvert from ProteoWizard. Care should be taken to optimize data extraction parameters using an iterative approach.

The database search will first try to match tryptic peptide masses to a protein sequence database, including any modifications selected. It will then try to match the acquired extracted MS/MS data to a theoretical fragmentation of the matched peptide sequences and assign a score. This is the search space, which puts a high demand on computational resources. Care should be taken to optimize parameters to reduce search time without impacting data quality. This can be done by optimizing the size of the data file for signal-to-noise and reducing the search space by using both a smaller database and a smaller MS ppm error window.

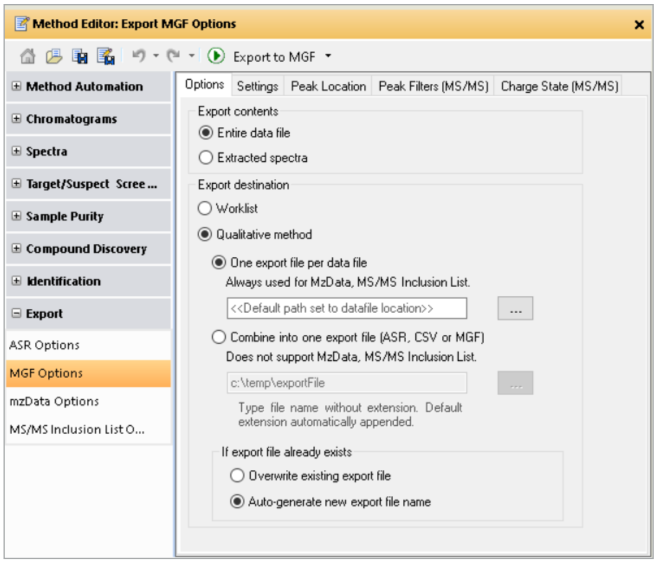
Results should be checked for correct modifications, including: cysteine with and without modification, number of unexpected missed tryptic cleavages, mass accuracy, score, and percent sequence coverage. The use of a system check sample (similar to your target) is highly recommended. Reporting can be as simple as copying and pasting, or exporting the results to a .CSV file.

Workflow overview of Mascot (online)

Mascot online from Matrix Science is a free version of the Mascot Server 3.0 and related products (note that the free version is limited in terms of function and upload file size). It is frequently used as an introduction to proteomic data processing due to its easy-to-use interface and online tutorials. Agilent AutoMSMS proteomic data can be searched using the Mascot MS/MS Ions Search page.

For data extraction using Agilent MassHunter Qualitative Analysis 10.0 through a mascot generic format file (.MGF), choose the following:

1. Select **Configuration** and click **Show Advanced Settings**.
2. Start from a Default.m method and save as DA_Qual_MFG_Export.m.
3. Using the **Compounds** view, navigate to **Method Editor Export MGF options**. Extract the entire data file (Figure 1). This tends to give the highest Mascot score. If the file is too large, it can be preprocessed with **Method Editor Compound Discovery Find By Auto MS/MS** using the default parameters (make sure that **Results Extract Complete Result Set Automatically** is unticked and that **Highlight First Compound** is selected to reduce processing time). If the file is still too large, the **Limit to the largest 10,000 Compounds** function can be used. If a preprocess is used, then select the **Export the Extracted Spectra** function (Figure 1).



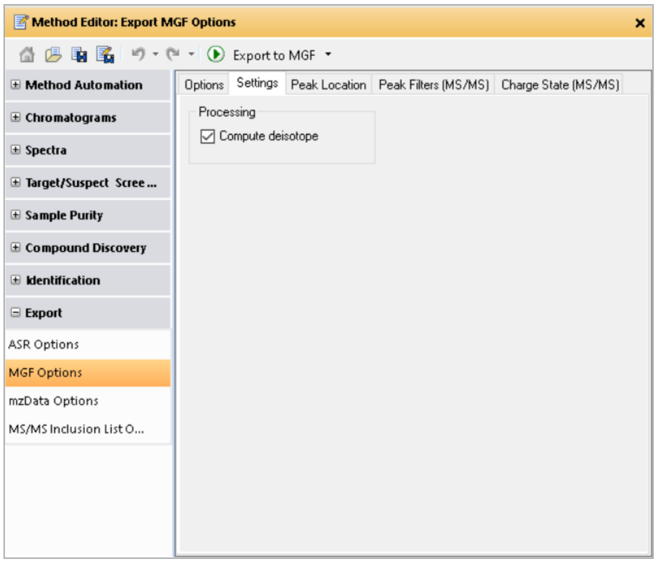
Export contents
● Entire data file
○ Extracted spectra

Export destination
○ Worklist
● Qualitative method
● One export file per data file
Always used for MzData, MS/MS Inclusion List.
<<Default path set to datafile location>> ...
○ Combine into one export file (ASR, CSV or MGF)
Does not support MzData, MS/MS Inclusion List.
c:\temp\exportFile ...
Type file name without extension. Default extension automatically appended.

If export file already exists
○ Overwrite existing export file
● Auto-generate new export file name

Figure 1. Method Editor: Export MGF Options > MGF Options > Options.

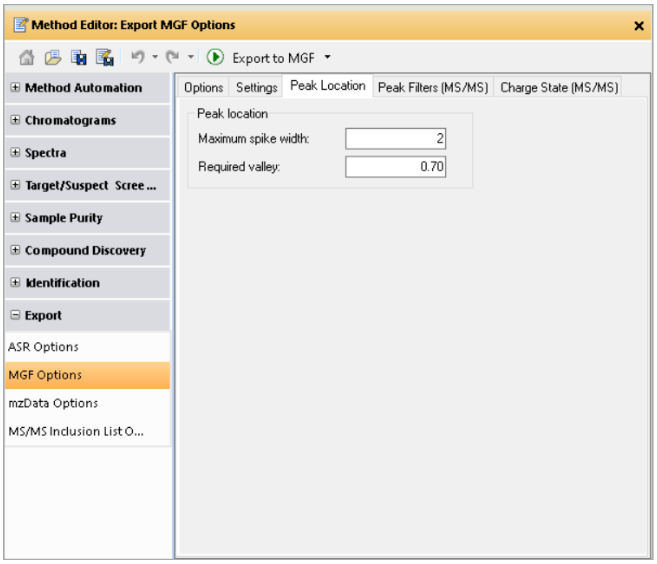
4. Under **Method Editor Export MGF options**, select **Compute deisotope** (Figure 2).



Processing
☑ Compute deisotope

Figure 2. Method Editor: Export MGF Options > MGF Options > Settings.

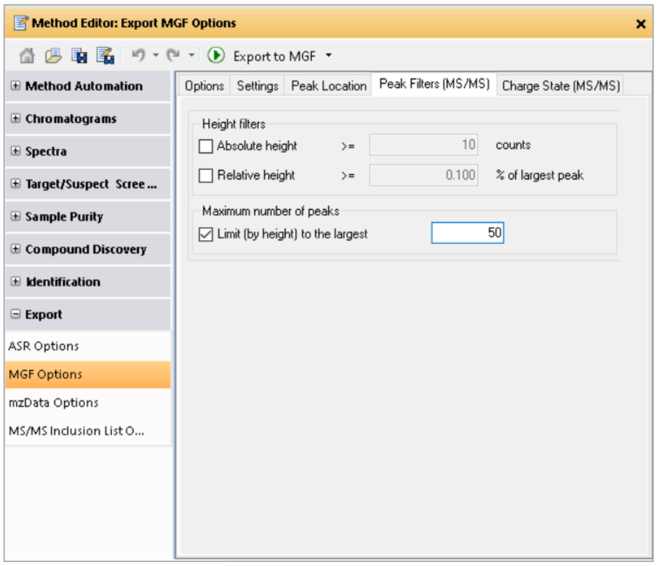
5. For **Peak Location**, leave as default (Figure 3).



Peak location
Maximum spike width: 2
Required valley: 0.70

Figure 3. Method Editor: Export MGF Options > MGF Options > Peak Location.

6. For **Peak Filters (MS/MS)**, deselect both height filters, and set Maximum number of peaks to 50 to export signal and reduce noise (Figure 4).



Method Editor: Export MGF Options

Export to MGF

⊞ Method Automation
⊞ Chromatograms
⊞ Spectra
⊞ Target/Suspect Scree...
⊞ Sample Purity
⊞ Compound Discovery
⊞ Identification
⊟ Export
ASR Options
MGF Options
mzData Options
MS/MS Inclusion List O...

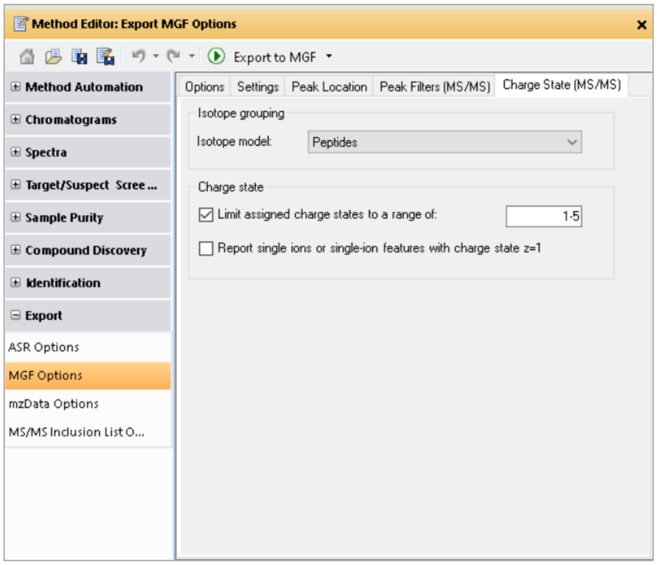
Options | Settings | Peak Location | Peak Filters (MS/MS) | Charge State (MS/MS)

Height filters
□ Absolute height >= 10 counts
□ Relative height >= 0.100 % of largest peak

Maximum number of peaks
☑ Limit (by height) to the largest 50

Figure 4. Method Editor: Export MGF Options > MGF Options > Peak Filters (MS/MS).

7. For **Charge State (MS/MS)**, select **Peptides** as the isotope model and a charge state of 1-5 (Figure 5).



Method Editor: Export MGF Options

Export to MGF

⊞ Method Automation
⊞ Chromatograms
⊞ Spectra
⊞ Target/Suspect Scree...
⊞ Sample Purity
⊞ Compound Discovery
⊞ Identification
⊟ Export
ASR Options
MGF Options
mzData Options
MS/MS Inclusion List O...

Options | Settings | Peak Location | Peak Filters (MS/MS) | Charge State (MS/MS)

Isotope grouping
Isotope model: Peptides

Charge state
☑ Limit assigned charge states to a range of: 1-5
□ Report single ions or single-ion features with charge state z=1

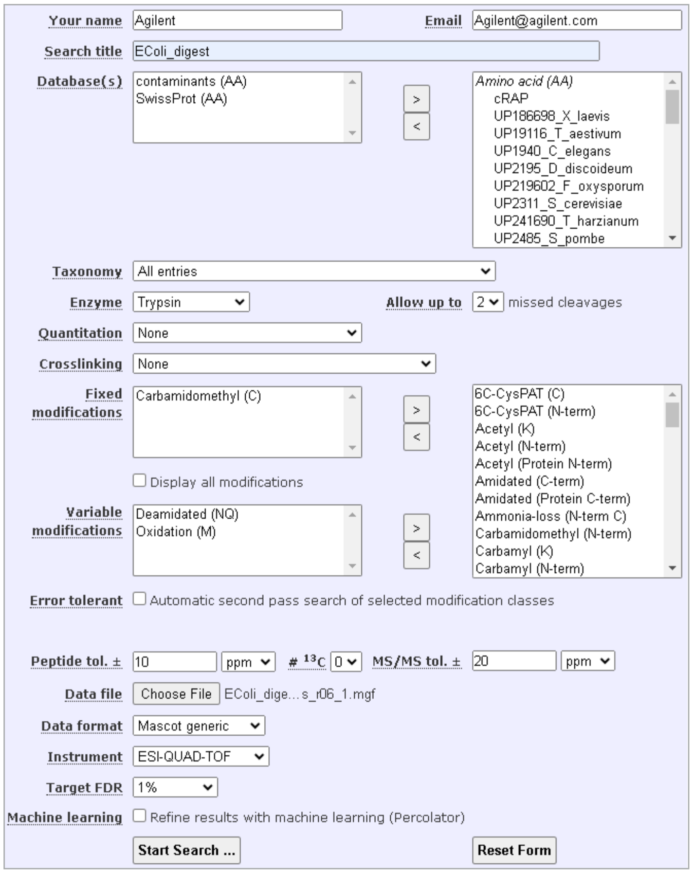
Figure 5. Method Editor: Export MGF Options > MGF Options > Charge State (MS/MS).

8. Select the output file location and click **Export to MGF**.

Data extraction tips and tricks

- Monitor the .MGF file size in Microsoft Windows File Explorer. Similar samples should have similar size files.
- Use an iterative extraction/search approach to optimize parameters.
- Extracting the entire data file will improve the Mascot score, but creates a larger .MGF file.
- Using **Find by AutoMSMS** instead of **Entire Data File** will give a lower Mascot score and a smaller .MGF file.
- Using **Deisotope** improves Mascot scores and reduces the .MGF file size.
- Setting **Maximum Number of Peaks** to 50 will give a higher Mascot score and a smaller .MGF file.
- Setting **Maximum Number of Peaks** to 100 will create an .MGF file approximately twice as large as using 50, but may result in more identifications.

Using the Mascot MS/MS Ions Search page



Your name	Agilent	Email	Agilent@agilent.com
Search title	EColi_digest		
Database(s)	contaminants (AA) SwissProt (AA)		
Taxonomy	All entries		
Enzyme	Trypsin	Allow up to	2 missed cleavages
Quantitation	None		
Crosslinking	None		
Fixed modifications	Carbamidomethyl (C)		
	□ Display all modifications		
Variable modifications	Deamidated (NQ) Oxidation (M)		
Error tolerant	□ Automatic second pass search of selected modification classes		
Peptide tol. ±	10 ppm	# ^{13}C	0
MS/MS tol. ±	20 ppm		
Data file	Choose File EColi_dige...s_r06_1.mgf		
Data format	Mascot generic		
Instrument	ESI-QUAD-TOF		
Target FDR	1%		
Machine learning	□ Refine results with machine learning (Percolator)		

Figure 6. The Mascot MS/MS Ions Search page.

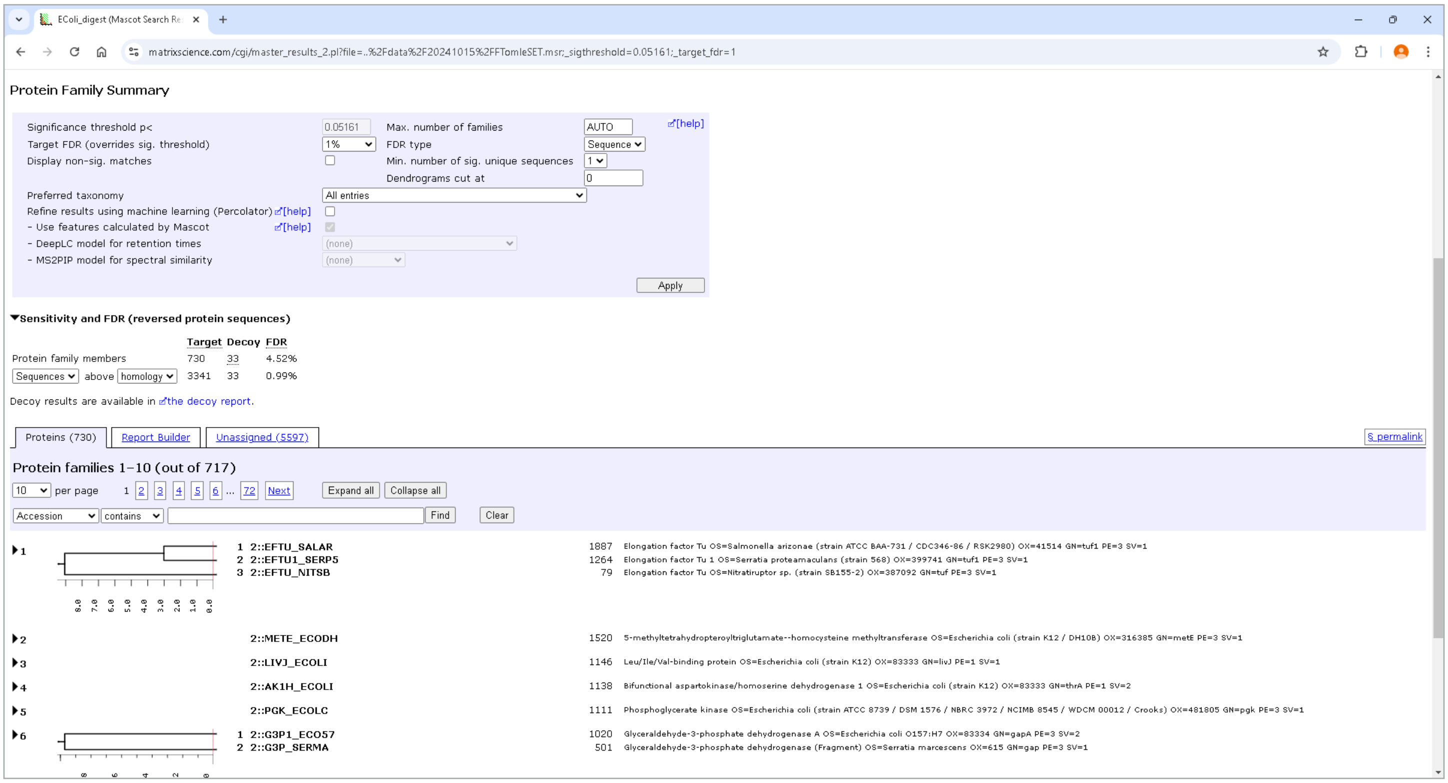
1. Fill in the information as requested.
2. Select a database. SwissProt is a good start as it is a nonredundant, manually curated, and relatively small database. Because Mascot scoring thresholds are related to database size, larger databases may not increase the number of matches, but will increase search space. An easy way to reduce search space is to select the correct taxonomy within that database. It is good practice to leave the "Contaminate database" in the search.
3. Select the digestion enzyme (this will typically be Trypsin) and allow up to two missed cleavages. Including more than two missed cleavages will increase the search space and slow the process.
4. Quantitation and Crosslinking are advanced features beyond the scope of this technical overview.
5. Fixed modifications would include the alkylating reagent used to block reduced disulfide bonds, usually carbamidomethylation (iodoacetamide) or carboxymethylation (iodoacetic acid). It is recommended to periodically check for incomplete alkylation either by omitting this fixed modification or adding it as a variable modification.
6. Variable modifications as a minimum add oxidation (M) and deamidated (NQ) as the most common. The use of an excessive number of variable modifications greatly increases the search space, slows the process and may increase the FDR.
7. Deselect **Error Tolerant**. This is an advanced feature used to account for artifacts in the data (refer to the Mascot help page for more information).
8. Use Peptide Tol ± 10 ppm MS/MS Tol ± 20 ppm.
From a small molecule point of view, these parameters seem excessively wide, but due to the unique way the data are processed, there is little downside and the mass accuracy can be checked in the results. As the peptide tolerance range increases, the search space increases; 20 ppm is the maximum recommended value. The # ^{13}C is typically set to zero, because the data have been deisotoped as part of data extraction.
9. Select the .MGF data file and upload it to the server center (data will be transferred through the internet). The .MGF data format is the most reliable way to get the extracted data into the Mascot MS/MS Ions Search page. The instrument should be set to ESI-QUAD-TOF. This tells Mascot what type of fragmentation product ions to expect. This cannot be edited in the free online version. If the full product is purchased, you should add the immonium and a-series ions for improved scores.
10. Set Target FDR to 1%, and deselect **Machine Learning**.
11. Click **Start Search**.

The search time online is very fast due to uploading the data to a dedicated server center. The above parameters can be set and saved as a cookie on the web page.[4]

Mascot MS/MS Ions Search page tips and tricks

- Use the smallest database and taxonomy possible, as this will reduce the search space.
- Try three missed cleavages to test for enzyme digestion completeness.
- Review the reported ppm errors and adjust Peptide Tol and MS/MS Tol, if required.
- Use an iterative approach to optimize parameters.

Results are delivered as a series of web pages, from which results can be copied and pasted (Figures 7 and 8). The protein results are downloadable as a .CSV file from the Report Builder (Figure 9). Graphics, including MS/MS spectra, can be downloaded as an .SVG file. Mascot provides complete online training for its software. Agilent can provide end-to-end proteomic and biopharmaceutical workflows and training.


EColi_digest (Mascot Search Re
matrixscience.com/cgi/master_results_2.pl?file=..%2Fdata%2F20241015%2FFTomleSET.msr;_sigthreshold=0.05161;_target_fdr=1
[help]
Apply
Expand all
Collapse all
Find
Clear
§ permalink
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
8
6
4
2
0

Protein Family Summary

Significance threshold p< 0.05161
Target FDR (overrides sig. threshold) 1%
Display non-sig. matches □
Max. number of families AUTO
FDR type Sequence
Min. number of sig. unique sequences 1
Dendrograms cut at 0
Preferred taxonomy All entries
Refine results using machine learning (Percolator) [help] □
- Use features calculated by Mascot [help] ☑
- DeepLC model for retention times (none)
- MS2PIP model for spectral similarity (none)

▼Sensitivity and FDR (reversed protein sequences)

	Target	Decoy	FDR
Protein family members	730	33	4.52%
Sequences above homology	3341	33	0.99%

Decoy results are available in the decoy report.

Proteins (730) | Report Builder | Unassigned (5597)

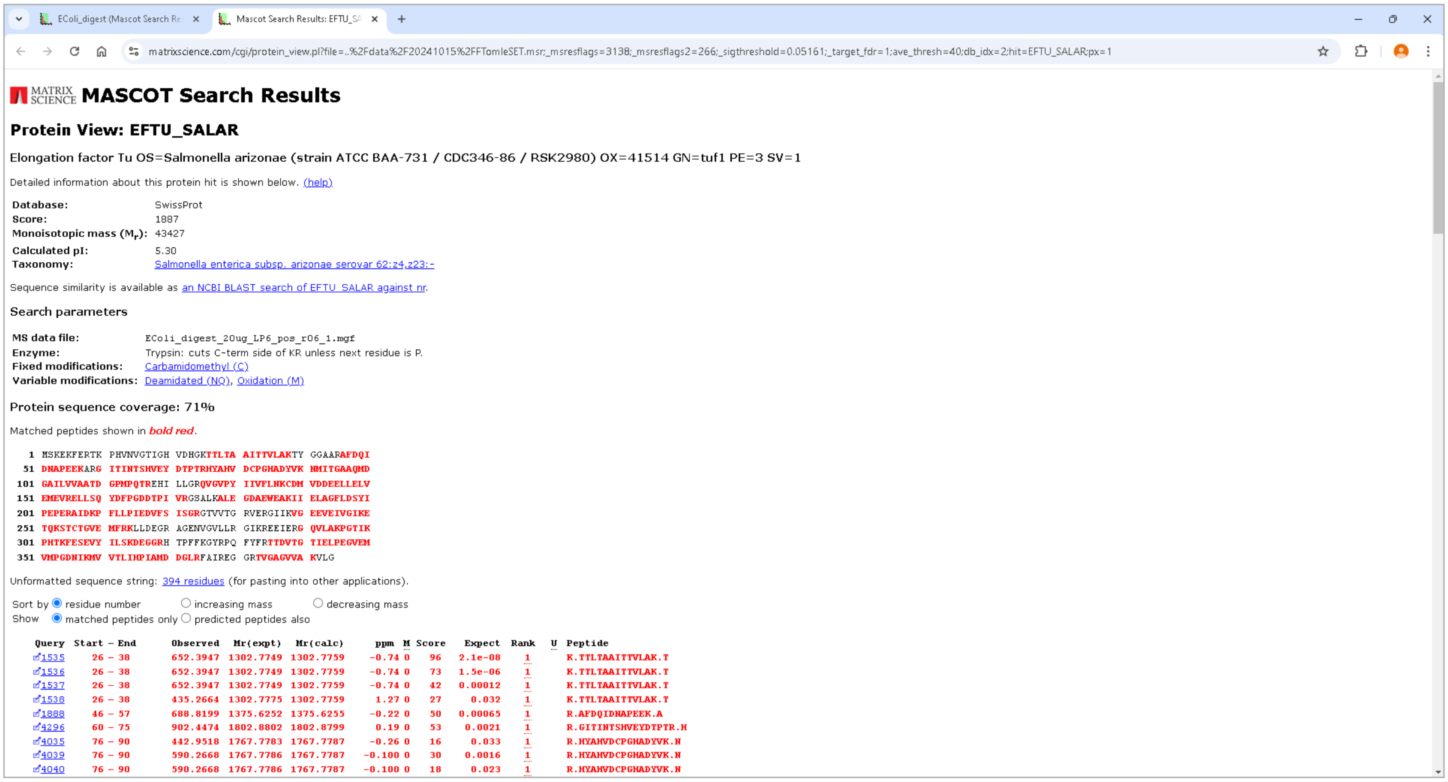
Protein families 1–10 (out of 717)

10 per page 1 2 3 4 5 6 ... 72 Next

Accession contains

▶1	1	2::EFTU_SALAR	1887	Elongation factor Tu OS=Salmonella arizonae (strain ATCC BAA-731 / CDC346-86 / RSK2980) OX=41514 GN=tuf1 PE=3 SV=1
	2	2::EFTU1_SERP5	1264	Elongation factor Tu 1 OS=Serratia proteamaculans (strain 568) OX=399741 GN=tuf1 PE=3 SV=1
	3	2::EFTU_NITSB	79	Elongation factor Tu OS=Nitratiruptor sp. (strain SB155-2) OX=387092 GN=tuf PE=3 SV=1
▶2		2::METE_ECODH	1520	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase OS=Escherichia coli (strain K12 / DH10B) OX=316385 GN=metE PE=3 SV=1
▶3		2::LIVJ_ECOLI	1146	Leu/Ile/Val-binding protein OS=Escherichia coli (strain K12) OX=83333 GN=livJ PE=1 SV=1
▶4		2::AK1H_ECOLI	1138	Bifunctional aspartokinase/homoserine dehydrogenase 1 OS=Escherichia coli (strain K12) OX=83333 GN=thrA PE=1 SV=2
▶5		2::PGK_ECOLC	1111	Phosphoglycerate kinase OS=Escherichia coli (strain ATCC 8739 / DSM 1576 / NBRC 3972 / NCIMB 8545 / WDCM 00012 / Crooks) OX=481805 GN=pgk PE=3 SV=1
▶6	1	2::G3P1_ECO57	1020	Glyceraldehyde-3-phosphate dehydrogenase A OS=Escherichia coli O157:H7 OX=83334 GN=gapA PE=3 SV=2
	2	2::G3P_SERMA	501	Glyceraldehyde-3-phosphate dehydrogenase (Fragment) OS=Serratia marcescens OX=615 GN=gap PE=3 SV=1

Figure 7. Mascot search results overview of all proteins.


EColi_digest (Mascot Search Re
Mascot Search Results: EFTU_SA
matrixscience.com/cgi/protein_view.pl?file=..%2Fdata%2F20241015%2FFTomleSET.msr;_msresflags=3138;_msresflags2=266;_sigthreshold=0.05161;_target_fdr=1;ave_thresh=40;db_idx=2;hit=EFTU_SALAR;px=1
MATRIX SCIENCE

MASCOT Search Results

Protein View: EFTU_SALAR

Elongation factor Tu OS=Salmonella arizonae (strain ATCC BAA-731 / CDC346-86 / RSK2980) OX=41514 GN=tuf1 PE=3 SV=1

Detailed information about this protein hit is shown below. (help)

Database:	SwissProt
Score:	1887
Monoisotopic mass (M_r):	43427
Calculated pI:	5.30
Taxonomy:	Salmonella enterica subsp. arizonae serovar 62:z4,z23:-

Sequence similarity is available as an NCBI BLAST search of EFTU_SALAR against nr.

Search parameters

MS data file:	EColi_digest_20ug_LP6_pos_r06_1.mgf
Enzyme:	Trypsin: cuts C-term side of KR unless next residue is P.
Fixed modifications:	Carbamidomethyl (C)
Variable modifications:	Deamidated (NQ), Oxidation (M)

Protein sequence coverage: 71%

Matched peptides shown in ***bold red***.

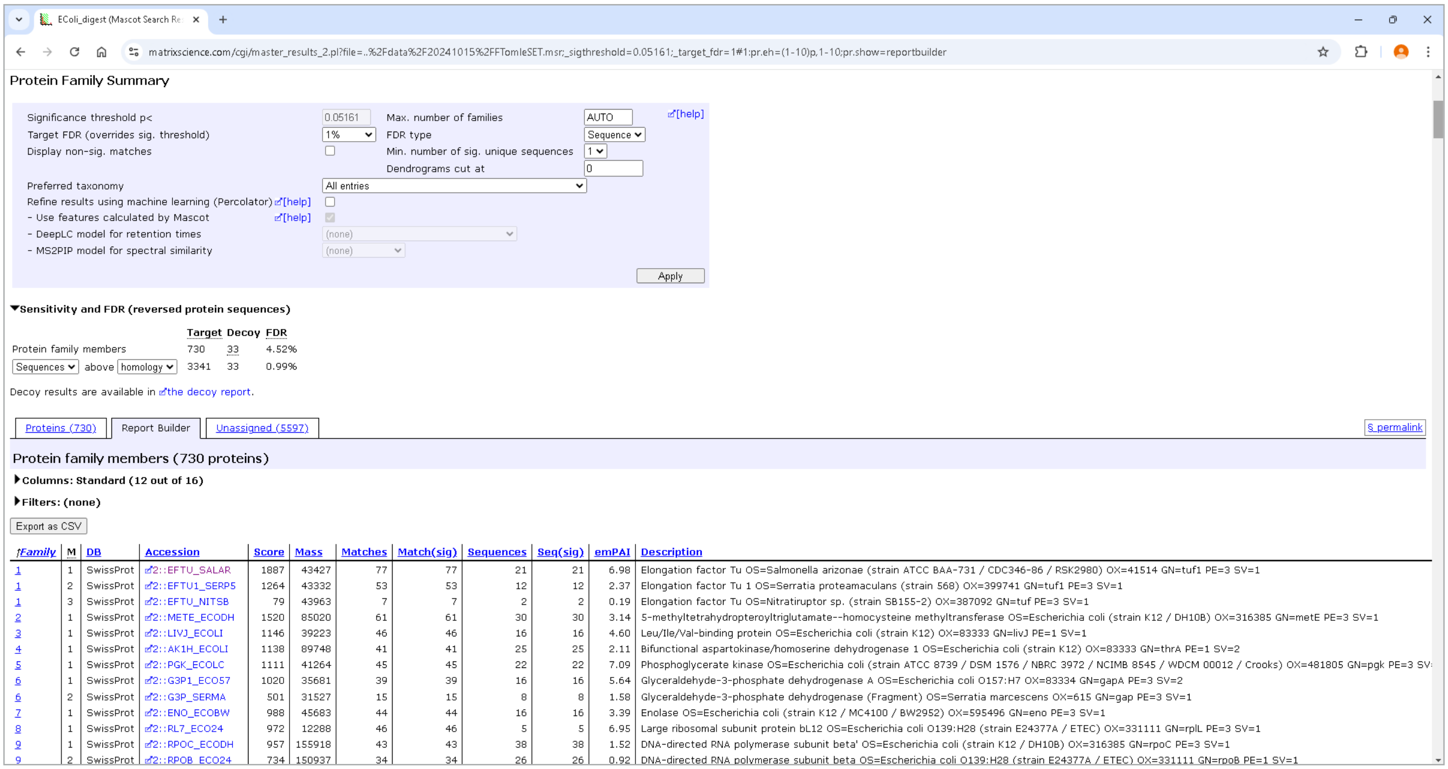
```
  1 MSKEKFERTK PHVNVGTIGH VDHGKTTLTA AITTVLAKTY GGAARAFDQI
 51 DNAPEEKARG ITINTSHVEY DTPTRHYAHV DCPGHADYVK NMITGAAQMD
101 GAILVVAATD GPMPQTREHI LLGRQVGVPY IIVFLNKCDM VDDEELLELV
151 EMEVRELLSQ YDFPGDDTPI VRGSALKALE GDAEWEAKII ELAGFLDSYI
201 PEPERAIDKP FLLPIEDVFS ISGRGTVVTG RVERGIIKVG EEVEIVGIKE
251 TQKSTCTGVE MFRKLLDEGR AGENVGVLLR GIKREEIERG QVLAKPGTIK
301 PHTKFESEVY ILSKDEGGRH TPFFKGYRPQ FYFRTTDVTG TIELPEGVEM
351 VMPGDNIKMV VTLIHPIAMD DGLRFAIREG GRTVGAGVVA KVLG
```

Unformatted sequence string: 394 residues (for pasting into other applications).

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

Query	Start – End	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
1535	26 – 38	652.3947	1302.7749	1302.7759	-0.74	0	96	2.1e-08	1		K.TTLTAAITTVLAK.T
1536	26 – 38	652.3947	1302.7749	1302.7759	-0.74	0	73	1.5e-06	1		K.TTLTAAITTVLAK.T
1537	26 – 38	652.3947	1302.7749	1302.7759	-0.74	0	42	0.00012	1		K.TTLTAAITTVLAK.T
1538	26 – 38	435.2664	1302.7775	1302.7759	1.27	0	27	0.032	1		K.TTLTAAITTVLAK.T
1888	46 – 57	688.8199	1375.6252	1375.6255	-0.22	0	50	0.00065	1		R.AFDQIDNAPEEK.A
4296	60 – 75	902.4474	1802.8802	1802.8799	0.19	0	53	0.0021	1		R.GITINTSHVEYDTPTR.H
4035	76 – 90	442.9518	1767.7783	1767.7787	-0.26	0	16	0.033	1		R.HYAHVDCPGHADYVK.N
4039	76 – 90	590.2668	1767.7786	1767.7787	-0.100	0	30	0.0016	1		R.HYAHVDCPGHADYVK.N
4040	76 – 90	590.2668	1767.7786	1767.7787	-0.100	0	18	0.023	1		R.HYAHVDCPGHADYVK.N

Figure 8. Mascot search results: single protein details.


EColi_digest (Mascot Search Re
matrixscience.com/cgi/master_results_2.pl?file=..%2Fdata%2F20241015%2FFTomleSET.msr;_sigthreshold=0.05161;_target_fdr=1#1:pr.eh=(1-10)p,1-10;pr.show=reportbuilder

Protein Family Summary

Significance threshold p<	0.05161	Max. number of families	AUTO	[help]
Target FDR (overrides sig. threshold)	1%	FDR type	Sequence	
Display non-sig. matches	☐	Min. number of sig. unique sequences	1	
		Dendrograms cut at	0	
Preferred taxonomy	All entries			
Refine results using machine learning (Percolator) [help]	☐			
- Use features calculated by Mascot [help]	☑			
- DeepLC model for retention times	(none)			
- MS2PIP model for spectral similarity	(none)			

Apply

Sensitivity and FDR (reversed protein sequences)

	Target	Decoy	FDR
Protein family members	730	33	4.52%
Sequences above homology	3341	33	0.99%

Decoy results are available in the decoy report.

Proteins (730) | Report Builder | Unassigned (5597) | § permalink

Protein family members (730 proteins)

Columns: Standard (12 out of 16)

Filters: (none)

Export as CSV

↑Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	SwissProt	2::EFTU_SALAR	1887	43427	77	77	21	21	6.98	Elongation factor Tu OS=Salmonella arizonae (strain ATCC BAA-731 / CDC346-86 / RSK2980) OX=41514 GN=tuf1 PE=3 SV=1
1	2	SwissProt	2::EFTU1_SERP5	1264	43332	53	53	12	12	2.37	Elongation factor Tu 1 OS=Serratia proteamaculans (strain 568) OX=399741 GN=tuf1 PE=3 SV=1
1	3	SwissProt	2::EFTU_NITSB	79	43963	7	7	2	2	0.19	Elongation factor Tu OS=Nitratiruptor sp. (strain SB155-2) OX=387092 GN=tuf PE=3 SV=1
2	1	SwissProt	2::METE_ECODH	1520	85020	61	61	30	30	3.14	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase OS=Escherichia coli (strain K12 / DH10B) OX=316385 GN=metE PE=3 SV=1
3	1	SwissProt	2::LIVJ_ECOLI	1146	39223	46	46	16	16	4.60	Leu/Ile/Val-binding protein OS=Escherichia coli (strain K12) OX=83333 GN=livJ PE=1 SV=1
4	1	SwissProt	2::AK1H_ECOLI	1138	89748	41	41	25	25	2.11	Bifunctional aspartokinase/homoserine dehydrogenase 1 OS=Escherichia coli (strain K12) OX=83333 GN=thrA PE=1 SV=2
5	1	SwissProt	2::PGK_ECOLC	1111	41264	45	45	22	22	7.09	Phosphoglycerate kinase OS=Escherichia coli (strain ATCC 8739 / DSM 1576 / NBRC 3972 / NCIMB 8545 / WDCM 00012 / Crooks) OX=481805 GN=pgk PE=3 SV
6	1	SwissProt	2::G3P1_ECO57	1020	35681	39	39	16	16	5.64	Glyceraldehyde-3-phosphate dehydrogenase A OS=Escherichia coli O157:H7 OX=83334 GN=gapA PE=3 SV=2
6	2	SwissProt	2::G3P_SERMA	501	31527	15	15	8	8	1.58	Glyceraldehyde-3-phosphate dehydrogenase (Fragment) OS=Serratia marcescens OX=615 GN=gap PE=3 SV=1
7	1	SwissProt	2::ENO_ECOBW	988	45683	44	44	16	16	3.39	Enolase OS=Escherichia coli (strain K12 / MC4100 / BW2952) OX=595496 GN=eno PE=3 SV=1
8	1	SwissProt	2::RL7_ECO24	972	12288	46	46	5	5	6.95	Large ribosomal subunit protein bL12 OS=Escherichia coli O139:H28 (strain E24377A / ETEC) OX=331111 GN=rplL PE=3 SV=1
9	1	SwissProt	2::RPOC_ECODH	957	155918	43	43	38	38	1.52	DNA-directed RNA polymerase subunit beta' OS=Escherichia coli (strain K12 / DH10B) OX=316385 GN=rpoC PE=3 SV=1
9	2	SwissProt	2::RPOB_ECO24	734	150937	34	34	26	26	0.92	DNA-directed RNA polymerase subunit beta OS=Escherichia coli O139:H28 (strain E24377A / ETEC) OX=331111 GN=rpoB PE=1 SV=1

Figure 9. Report building, export to .CSV file.

Conclusion

Agilent AutoMSMS proteomic data can be extracted using Agilent MassHunter Qualitative Analysis software, version 10.0 and searched with Mascot online from Matrix Science through an .MGF file. The results can be exported as graphics and as a .CSV file.

References

1. https://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS
2. https://proteowizard.sourceforge.io/tools/msconvert.html
3. https://www.matrixscience.com/blog/back-to-basics-optimize-your-search-parameters.html#comments
4. https://www.matrixscience.com/search_form_select.html

www.agilent.com

DE-004600

This information is subject to change without notice.

Printed in the USA, March 12, 2025
5994-8099EN