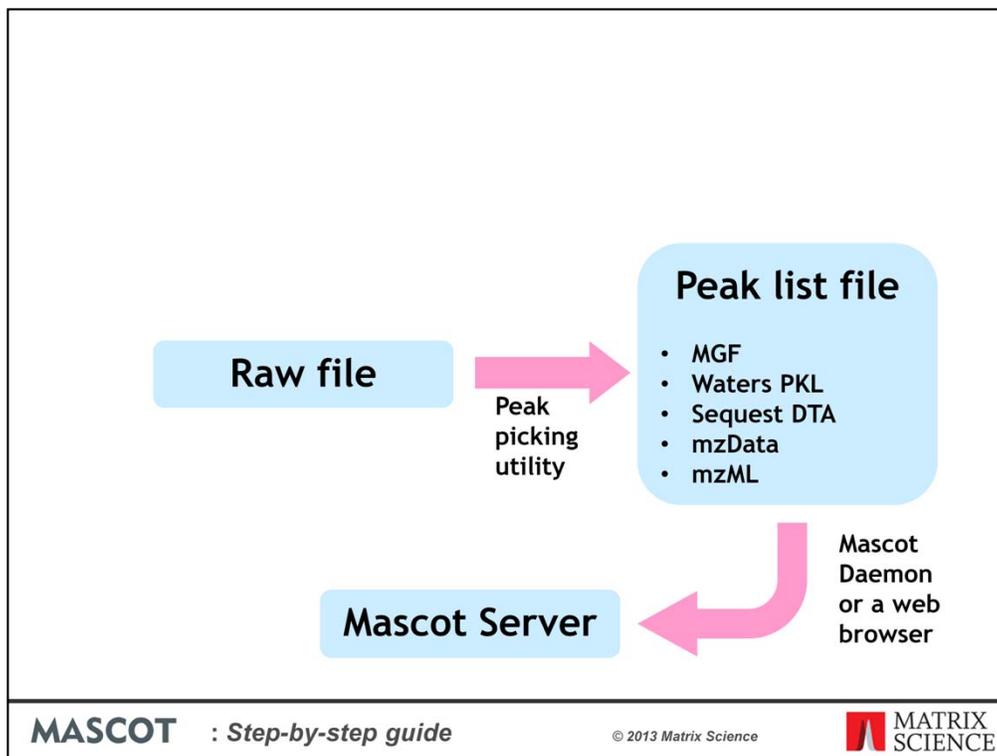


Step-by-step guide to searching MS/MS data with Mascot

MASCOT

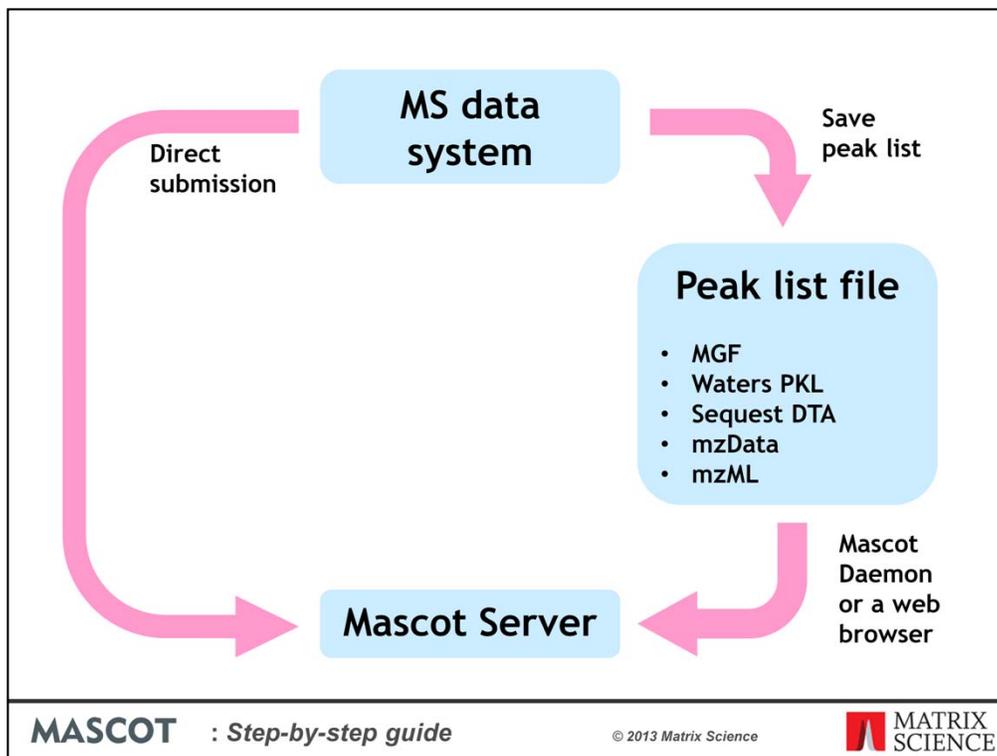


We are painfully aware that we don't have a good, introductory tutorial for Mascot on our web site. Its something that has come up in discussions many times, and we always resolve to do something but then get sidetracked. This talk is a dry run for a brief tutorial on searching MS/MS data. If you are new to database searching, I hope you will find it informative. If you are an experienced user, please give us your feedback on anything that is confusing or missing.

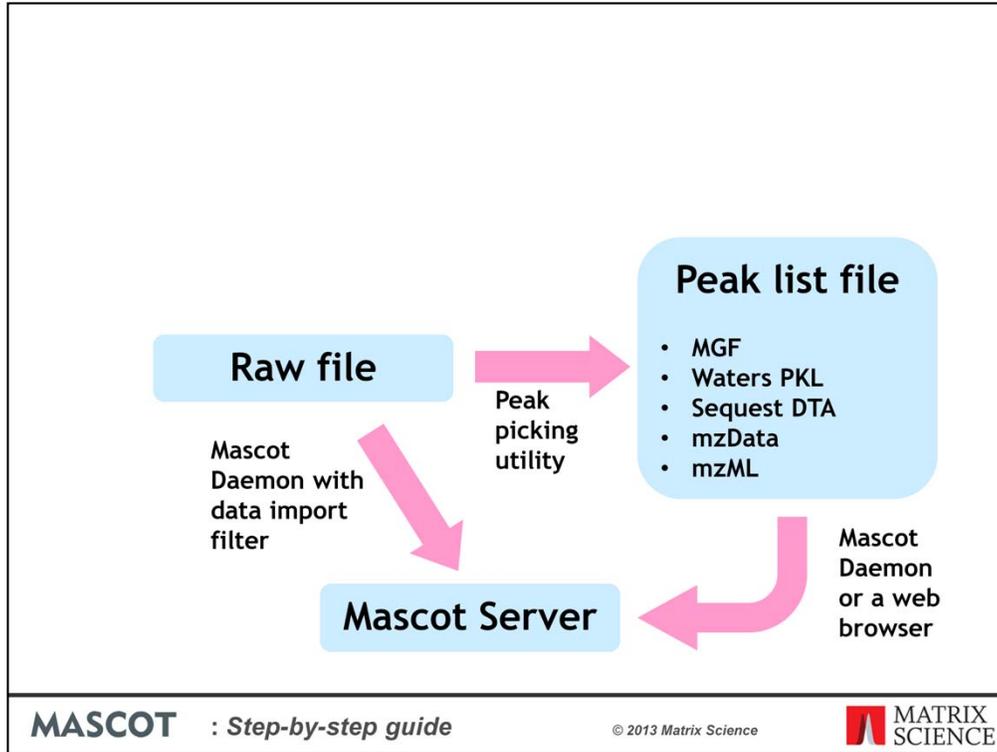


The first requirement for database searching with Mascot is a peak list; you cannot upload a raw data file. Raw data is converted into a peak list by a process called peak picking or peak detection.

Peak lists are text files and come in various different formats. If you have a choice, MGF is recommended, but you can also use any of the ones listed, plus a few others that are less widely used. Be careful with mzML, because this may contain either raw data or a peak list.

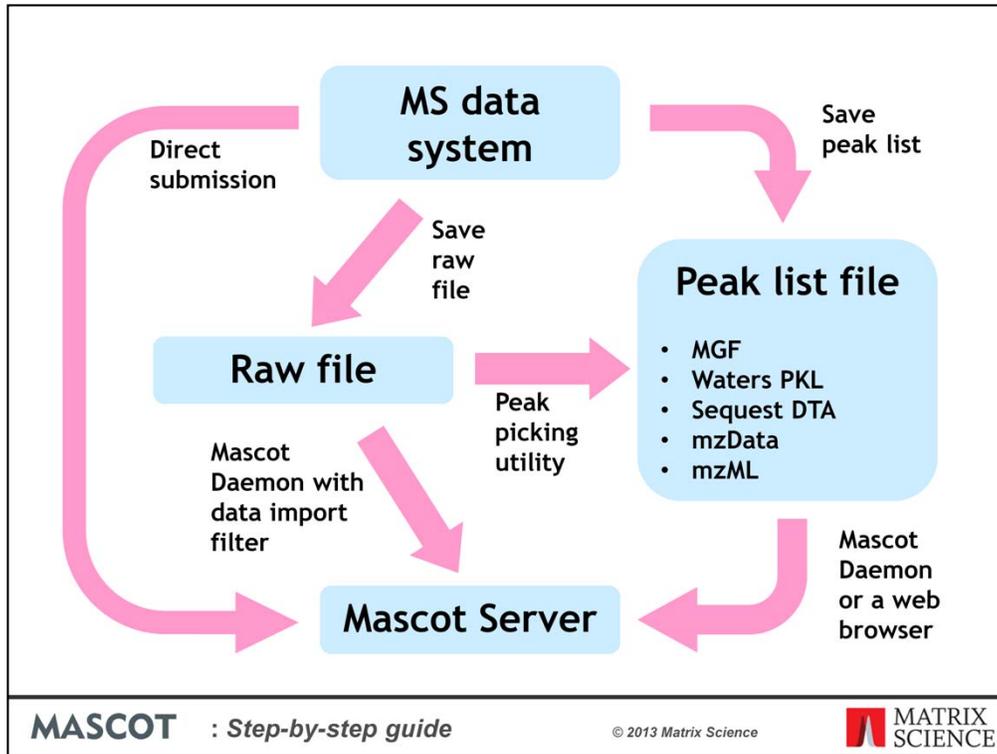


Often, the instrument data system takes care of peak picking, and you can submit a Mascot search directly from the data system or save a peak list to a disk file for submission using the web browser search form.

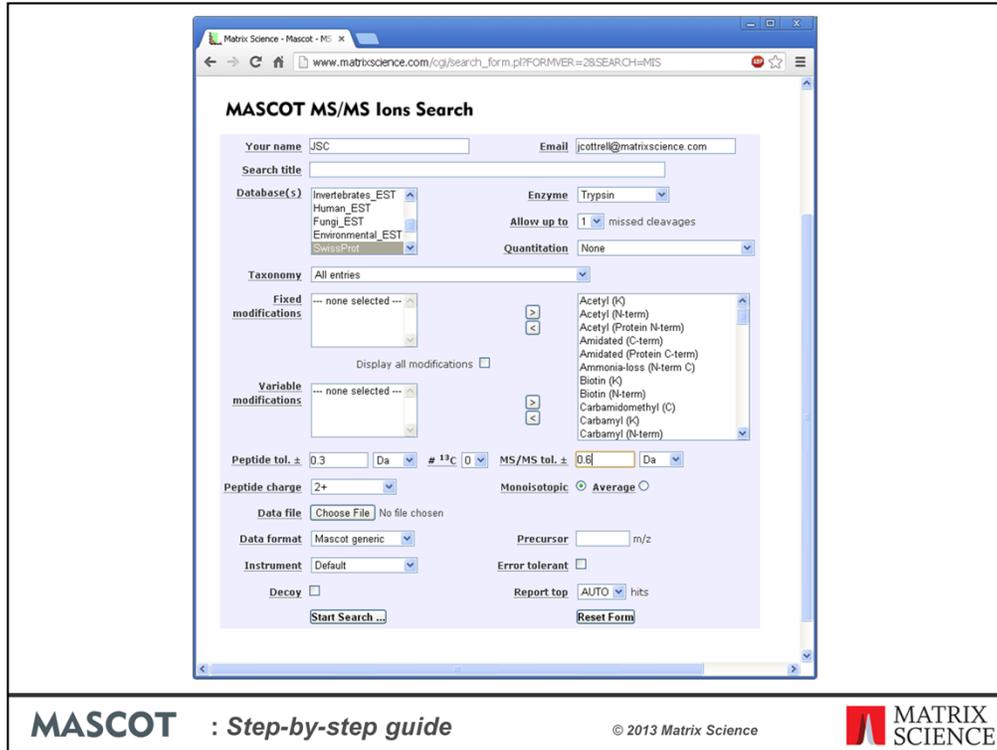


If not, or if you have a raw data file and no access to the data system, you'll need to find a utility to convert it into a peak list.

If you have Mascot in-house, you can also use Mascot Daemon to batch automate peak picking and search submission.



So, lots of ways to submit searches.



By itself, a peak list is not sufficient. There are also a number of search parameters that must be set appropriately. Here, we see the web browser search form for the current version of Mascot, 2.4. The labels for each control are also links to help topics.

The form looks much the same whether you have your own Mascot server, in-house, or whether you are connected to the free, public Mascot Server. If you are using the public Mascot Server, there are some restrictions, one of which is that you have to provide a name and email address so that we can email a link to your search results if the connection is broken. A more important restriction is that searches are limited to a maximum of 1200 spectra.

Whether you enter a search title is your choice. It is displayed at the top of the result report, and can be a useful way of identifying the search at some later time.

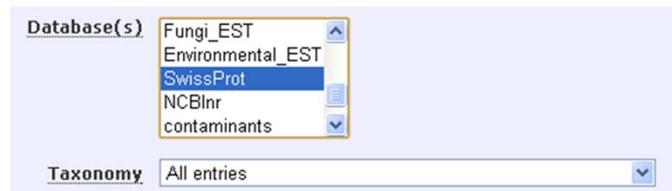
Golden Rule #1

- **If at all possible, run a standard sample and use this to set all the search parameters.**



This brings us to our first 'golden rule'. If at all possible, run a standard sample and use this to set all the search parameters. By standard sample, I mean something like a BSA digest, which will give strong matches and where you know what the answer is supposed to be. Trying to set search parameters on an unknown is much more difficult, especially if the sample was lost somewhere during the work-up or if the instrument has developed a fault. I know this advice will often be ignored, but it is probably the single most important message of this talk.

Database



The screenshot shows a web interface for selecting a database. The 'Database(s)' dropdown menu is open, displaying a list of options: Fungi_EST, Environmental_EST, SwissProt (which is highlighted in blue), NCBItr, and contaminants. Below this, the 'Taxonomy' dropdown menu is set to 'All entries'.

The first choice you have to make, and one of the more complicated, is which database to search. The free public Mascot Server has just a few of the more popular public databases, but an in-house server may have a hundred or more. Some databases contain sequences from a single organism. Others contain entries from multiple organisms, but often include the taxonomy for each entry, so that entries for a specific organism can be selected during a search using a taxonomy filter.

If you're not sure what is in the sample, Swiss-Prot is a good starting point. The entries are all high quality and well annotated. Because Swiss-Prot is non-redundant, it is relatively small. The size of the database is one factor in the size of the search space - the number of peptide sequences that are compared with a spectrum to see which gives the best match. The smaller the search space, the easier it is to get a statistically significant match. This is a very important concept and other factors that affect the size of the search space will be highlighted as we come to them.

If you think you know what is in the sample, you may want to search an organism specific database. But, you can never rule out contaminants. This can be a severe problem if you only have a handful of spectra. You are interested in a human protein, so you search a human database, but your spectrum is for a peptide from a contaminant, so you get no match or a misleading match.

Golden Rule #2

- **When searching entries for a single organism, always include a database of common contaminants.**



So, our second 'golden rule' is, when searching entries for a single organism, always include a database of common contaminants. This is important, even if you have a large dataset and no interest in proteins from anything other than your target organism. You may end up reporting your sample is full of human serum albumin when its really BSA or mouse keratin when its really sheep keratin from a sweater.

In the web browser search form, select your target database then hold down the control key to select an additional database of contaminants. If your search uses a taxonomy filter, that's not a problem because taxonomy is not configured for the contaminants database. All the entries will be searched, whatever the taxonomy setting.

Database

The screenshot shows the NCBI Taxonomy Browser interface. The search bar contains 'Citrus sinensis' and the results are displayed in a table. The table has three columns: Database name, Subtree links, and Direct links. The data is as follows:

Database name	Subtree links	Direct links
Nucleotide	1,629	1,626
Nucleotide EST	214,598	214,598
Nucleotide GSS	3	3
Protein	758	758
Genome	1	1
Popset	26	26
GEO Datasets	108	108
UniGene	15,422	15,422
UniSTS	142	142
PubMed Central	357	357
Gene	140	140
SEA Experiments	15	15
Probe	14	14
Bio Project	25	25
Bio Sample	25	25
Protein Clusters	70	70
Taxonomy	2	1

MASCOT : Step-by-step guide

© 2013 Matrix Science



If your target organism is well characterised, such as human or mouse or yeast or arabidopsis, there may be no need to look beyond Swiss-Prot. You can get a sense of how well your organism is represented by going to <http://www.uniprot.org/> and looking at the Swiss-Prot release notes, which list the 250 best represented species. On the other hand, if you are interested in a bacterium or a plant, you may find that it is poorly represented. The next place to look is one of the comprehensive protein databases, which aim to include all known protein sequences. The two best known are NCBIInr and UniRef100. If the genome of your organism hasn't been sequenced, you may still be out of luck, and your best hope may be to search an EST database (Expressed Sequence Tags are relatively short nucleic acid sequences).

Here, we can see the entry in the NCBI taxonomy browser for orange, the citrus fruit. Just 94 entries for orange in Swiss_prot and only 758 in the whole of NCBIInr. If this is your organism of interest, you'll probably want to search the ESTs

Database

```
Num sequences for taxonomy : All entries=540052
Num sequences for taxonomy : Archaea (Archaeobacteria)=19010
Num sequences for taxonomy : Eukaryota (eucaryotes)=175860
Num sequences for taxonomy : Alveolata (alveolates)=1040
Num sequences for taxonomy : Plasmodium falciparum (malaria parasite)=300
Num sequences for taxonomy : Other Alveolata=740
Num sequences for taxonomy : Metazoa (Animals)=102656
Num sequences for taxonomy : Caenorhabditis elegans=3425
Num sequences for taxonomy : Drosophila (fruit flies)=5515
Num sequences for taxonomy : Chordata (vertebrates and relatives)=83807
Num sequences for taxonomy : bony vertebrates=83207
Num sequences for taxonomy : lobe-finned fish and tetrapod clade=78127
Num sequences for taxonomy : Mammalia (mammals)=66146
Num sequences for taxonomy : Primates=26973
Num sequences for taxonomy : Homo sapiens (human)=20256
Num sequences for taxonomy : Other primates=6717
Num sequences for taxonomy : Rodentia (Rodents)=26139
Num sequences for taxonomy : Mus.=16670
Num sequences for taxonomy : Mus musculus (house mouse)=16619
Num sequences for taxonomy : Rattus=7867
Num sequences for taxonomy : Other rodentia=1602
Num sequences for taxonomy : Other mammalia=13034
Num sequences for taxonomy : Xenopus laevis (African clawed frog)=3365
Num sequences for taxonomy : Other lobe-finned fish and tetrapod clade=8616
Num sequences for taxonomy : Actinopterygii (ray-finned fishes)=5080
Num sequences for taxonomy : Takifugu rubripes (Japanese Pufferfish)=173
```

MASCOT : *Step-by-step guide*

© 2013 Matrix Science



Don't choose a narrow taxonomy without looking at the counts of entries and understanding the classification. In the current SwissProt, for example, there are 26,139 entries for rodentia, of which all but 1,602 are for mouse and rat. So, even if your target organism is hamster, it isn't a good idea to choose 'other rodentia'. Better to search rodentia and hope to get matches to homologous proteins from mouse and rat.

Swiss-Prot is a non-redundant database, where sequences that are very similar have been collapsed into a single entry. This means that the database entry will often differ slightly from the protein you analysed. Standard database searching requires the exact peptide sequence, so you may miss some matches due to SNPs and other variants. This would be another reason to search a large, comprehensive database. However, NCBIInr is 50 times the size of Swiss-Prot, so searches take proportionally longer and the search space is proportionally larger, meaning that you need higher quality data to get a significant match.

Enzyme

Enzyme

Allow up to missed cleavages

MASCOT : *Step-by-step guide* © 2013 Matrix Science **MATRIX SCIENCE**

If your protein was digested using an enzyme, always choose this enzyme. Choosing a semi-specific enzyme or non-specific cleavage greatly increases the search time and the search space, which will almost certainly cause a net reduction in the number of matches. The error tolerant search, to be discussed shortly, is a much better way of finding non-specific peptides. If you are studying endogenous peptides, such as MHC peptides, you have no choice, and enzyme 'None' will look for matches in all sub-sequences of all proteins. If you are doing top-down, and analysing the intact protein, choose NoCleave. Note that NoCleave is not the same as None; in some ways, it is the exact opposite.

The number of missed cleavages should be set empirically, by running a standard and looking at the proportions of matches with missed cleavages. Setting this value higher than necessary just increases the size of the search space, which I hope we are now coming to recognise is a 'bad thing'.

Golden Rule #3

- If your protein was digested using an enzyme, always choose this enzyme.



We'll be flagging this up as another golden rule in the summary at the end

Modifications

Fixed modifications Carbamidomethyl (C)

Display all modifications

Variable modifications Oxidation (M)

mTRAQ:13C(6)15N(2) (N-term)
mTRAQ:13C(6)15N(2) (Y)
NIPCAM (C)
Oxidation (HW)
Phospho (ST)
Phospho (Y)
Propionamide (C)
Pyridylethyl (C)
Pyro-carbamidomethyl (N-term C)
Sulfo (STY)
TMT2plex (K)

MASCOT : Step-by-step guide © 2013 Matrix Science **MATRIX SCIENCE**

Modifications in database searching are handled in two ways. First, there are the fixed modifications. An example would be the efficient alkylation of cysteine. Since all cysteines are modified, this is effectively just a change in the mass of cysteine. It carries no penalty in terms of search speed or specificity.

In contrast, most post-translational modifications do not apply to all instances of a residue. For example, phosphorylation might affect just one serine in a peptide containing many serines. These variable modifications are expensive, in the sense that they increase the search space because the software has to permute out all the possible arrangements of modified and unmodified residues that fit to the peptide molecular mass. As more and more modifications are considered, the number of combinations and permutations increases geometrically. We get a so-called combinatorial explosion.

Golden Rule #4

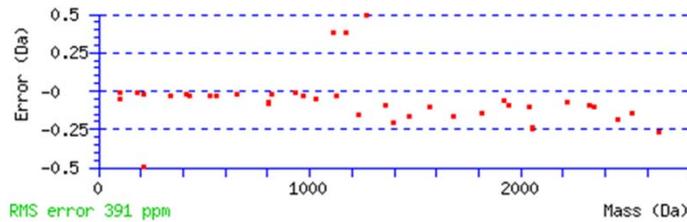
- Only specify very abundant modifications as variable.



Hence, it is very important to be as sparing as possible with variable modifications. If the aim of the search is to identify as many proteins as possible, the best advice is to use a minimum of variable modifications. Most post-translational modifications, such as phosphorylation, are rare and it is much more efficient to use an error tolerant search to find them.

Mass accuracy

Peptide tol. \pm 0.3 Da # ¹³C 0 MS/MS tol. \pm 0.6 Da
Monoisotopic Average



MASCOT : Step-by-step guide

© 2013 Matrix Science



Making an estimate of the mass accuracy doesn't have to be a guessing game. The Mascot result reports include graphs of mass errors, like the one shown here. Just run a standard and look at the error graphs for the strong matches. You'll normally see some kind of trend. Add on a safety margin and this is your error estimate. The graph for precursor mass error is in the Protein View report and the graph for MS/MS fragment mass error is in the Peptide View report

Sometimes, peak picking chooses the ¹³C peak rather than the ¹²C, so the mass is out by 1 Da. In extreme cases, it may pick the ¹³C₂ peak. The #¹³C control allows for this, allowing you to use a tight mass tolerance and still get a match.

Most modern instruments produce monoisotopic mass values. You will only have average masses if the entire isotope distribution has been centroided into a single peak, which usually implies very low resolution. (If you get this setting wrong, the mass errors will be very large and show a strong trend, because the difference between an average and a monoisotopic mass for peptides and proteins is approximately 0.06%.)

Peptide charge

Peptide charge

```
BEGIN IONS
TITLE=1337: Scan 5413 (rt=39.3462)
PEPMASS=591.73309 291522.19
CHARGE=2+
SCANS=5413
RAWSCANS=sn5413
RTINSECONDS=2360.7723
255.11138 37.37178
256.11293 21.135261
513.22442 26.917156
272.13556 229.38011
273.13798 144.60458
547.27148 85.617325
290.1655 26.233281
291.16756 17.040373
295.11934 15.032945
599.37675 27.27575
```

MASCOT : *Step-by-step guide*

© 2013 Matrix Science



Peptide charge is a default, only used if no charge is specified in the peak list. Most peak lists always specify a charge state, so this default is never used.

Instrument

Instrument **ESI-TRAP**

Ion series	Default	ESI TOF	MALDI TOF	ESI TRAP	ESI QUAD	ESI FTICR	MALDI TOF	ESI 4SECTOR	FTMS ECD	ETD TRAP	MALDI TOF	MALDI TOF	MALDI TOF	CID+ETD	FSD
1+	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2+ (precursor>2+)	X	X		X	X	X			X	X	X				X
2+ (precursor>3+)															
immonium			X				X	X			X	X			
a	X						X	X			X	X		X	
a+	X		X				X				X			X	
a0			X				X				X			X	
b	X	X	X	X	X	X	X	X			X	X		X	
b+	X	X	X	X	X	X	X	X			X	X		X	
b0	X	X	X	X	X	X	X	X			X	X		X	
c									X	X				X	X
x															
y	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
y+	X	X	X	X	X	X	X				X	X		X	
y0		X		X	X	X	X				X	X		X	
z								X							
z+								X	X		X	X			
z0								X	X		X	X			
y must be significant															
y must be highest score															
z+1									X	X				X	
d							X								
v							X								
w							X							X	
z+2									X	X				X	
Minimum mass	700	700	700	700	700	700	700	700	700	700	700	700	700	700	700
Max mass	Delete Edit														

MASCOT : Step-by-step guide

© 2013 Matrix Science



The Instrument setting determines which fragment ion series will be considered in the search. Choose the description that best matches your instrument. If you follow the label link to the help, you'll see that many of the instruments are very similar. The only serious problem is if you choose CID for ETD data or vice versa.

Report

Report top hits

MASCOT : *Step-by-step guide* © 2013 Matrix Science 

Report determines the maximum number of hits displayed in a search results report. Always choose *AUTO* to display all the protein hits containing one or more significant peptide matches. There is absolutely no point setting this to a very high value. You are just listing junk proteins, for which there is no evidence.

Decoy

Decoy

The screenshot shows the Mascot web interface with the 'Decoy' checkbox checked. The 'Protein Family Summary' section includes filter options for significance threshold p< (0.00645), max. number of families (AUTO), ions score or expect cut-off (0), and dendrograms cut at (0). The 'Decoy search summary (reversed protein sequences)' section shows peptide matches: 1503 in target vs 13 in decoy (0.86% FDR) above identity threshold, and 1869 in target vs 18 in decoy (0.96% FDR) above identity or homology threshold. Below this, protein families 1-10 are listed, including 2::TRY1_BOVIN, 1::GRP78_MOUSE, 3::HSZ1_MOUSE, 2::HSF7C_MOUSE, 1::CYB5_MOUSE, and 1::CBZCT_MOUSE, each with associated accession numbers and descriptions.

MASCOT : *Step-by-step guide*

© 2013 Matrix Science



The decoy checkbox enables you to estimate the peptide false discovery rate as recommended by most journals. Mascot repeats the search, using identical search parameters, against a database in which the sequences have been reversed. You do not expect to get any real matches from the decoy database. So, the number of matches that are obtained in the decoy database is an excellent estimate of the number of false positives in the results from the target database. The result report gains a control that adjusts the significance threshold to achieve a peptide FDR of 5% or 1% or whatever you believe is appropriate for your work.

Error tolerant search

Error tolerant

red	Hr(expt)	Hr(calc)	ppm	H Score	Expect	Rank	U	I	2	Peptide
328	973.3911	973.3906	0.53	48		1	U			K.ADEGI.SFR.G + [+79.9663 at 56]
595	979.5244	979.5239	0.48	19	0.059	1	U			K.SHPAPRFK.A
773	984.4401	984.4400	0.074	49		1	U			K.SLSLSDYK.G + [+44.9851 at 57]
362	1106.5979	1106.5972	0.67	49	0.0001	1	U			R.TIAQDFYVLR.A
775	1107.5804	1107.5812	-0.70	51		1	U			R.TIAQDFYVLR.A + [+8.9848 at 04]
158	1124.6171	1124.6190	-1.43	68		1	U			Q.IIVVDLPPVGR.S + [ModX] (H-term)
981	1131.5817	1131.5822	-0.49	46		1	U			R.TIAQDFYVLR.A + [+44.9851 at 56]
393	1163.5641	1163.5645	-0.38	46	0.00022	1	U			K.ATAVHFDGQFK.D
321	1164.5497	1164.5485	1.03	39		1	U			K.ATAVHFDGQFK.D + [+8.9848 at 09]
154	1179.5562	1179.5594	-2.69	36		1	U			K.ATAVHFDGQFK.D + [+15.9949 at 05]
265	1193.6385	1193.6405	-1.61	63	1.7e-05	1	U			R.QIVVDLPPVGR.S + Glu->pyro-Glu (H-term 0)
176	1194.6506	1194.6243	-3.23	51		1	U			R.QIVVDLPPVGR.S + Glu->pyro-Glu (H-term 0) + [+8.9848 at 05]
194	1185.6242	1185.6237	0.40	69	2.1e-06	1	U			R.LVQAFQFTDK.H
108	1196.6071	1196.6077	-0.57	70		1	U			R.LVQAFQFTDK.H + [+8.9848 at 03]
114	1196.6082	1196.6077	0.42	63		1	U			R.LVQAFQFTDK.H + [+8.9848 at 06]
106	1210.6667	1210.6670	-0.23	61	6.4e-05	1	U			R.QIVVDLPPVGR.S
321	1211.6496	1211.6510	-1.15	54		1	U			R.QIVVDLPPVGR.S + [+8.9848 at H-term 0]
327	1211.6509	1211.6518	-0.100	53		1	U			R.QIVVDLPPVGR.S + [+8.9848 at 05]
243	1212.6341	1212.6350	-0.79	43		1	U			R.QIVVDLPPVGR.S + 2 [+8.9848 at 01,05]
100	1262.7014	1262.6983	2.47	11	1.2	2	U			K.RTIAGQYVLR.A
312	1275.5879	1275.5901	-1.71	43		1	U			R.LVQAFQFTDK.H + [+79.9663 at 78]
184	1358.7933	1358.7922	0.82	49	0.00026	1	U			R.GLFIDDEKILR.Q
384	1400.8022	1400.8140	-0.43	71		1	U			R.GLFIDDEKILR.Q + [+47.0218 at 08]
333	1508.8680	1508.8603	5.12	50		1	U			R.QIVVDLPPVGR.S + [+278.1933 at H-term]
390	1663.9435	1663.9733	-3.02	1	0.21	1	U			R.GILRQIVVDLPPVGR.S
149	1663.9729	1663.9733	-0.24	64		1	U			R.GILRQIVVDLPPVGR.S + [+14.8157 at 04]

MASCOT : Step-by-step guide

© 2013 Matrix Science



As mentioned several times already, an error tolerant search is the most efficient way to discover most post-translational modifications, as well as non-specific peptides and sequence variants. This is a two pass search, the first pass being a simple search of the entire database with minimal modifications. The protein hits found in the first pass search are then selected for an exhaustive second pass search, during which we look for all possible modifications, sequence variants, and non-specific cleavage products. Because only a small number of entries are being searched, search time is not an issue. The matches from the first pass search, in a limited search space, are the evidence for the presence of the proteins, while the matches from the second pass search give increased coverage.

Running an error tolerant search couldn't be easier; just check the box. The hard work is in studying the report, and deciding which of the modifications you believe, because there will often be a choice.

If you see a very abundant modification, best to add this as a variable modification and then search again, because the error tolerant search only catches peptides with a single unsuspected modification.

Error tolerant searching is not so useful for very heavily modified proteins, such as histones, or where there is only one peptide per protein, such as endogenous peptides

Reasons why a spectrum fails to match

- The exact peptide sequence isn't in the database
- The peptide is modified in an unexpected way
- Non-specific enzyme cleavage
- The precursor m/z or charge is wrong
- The spectrum is very weak or noisy

If possible, you should search a peak list containing data for as many peptides as possible. This slide lists some of the many reasons why any one spectrum may fail to give a match.

If you search one spectrum and don't get a matches, you can only resort to changing the search parameters by trial and error, which is time consuming and carries the risk of ending up with a false positive. If you search many spectra, you have a much better chance that some of them match, and the search parameters can be modified systematically, or even automatically, in an error tolerant search.

Golden rules



1. Mascot requires peak lists; you cannot upload raw data
2. Search parameters are critical and should be set on a well characterised standard
3. When searching entries for a single organism, always include a database of common contaminants
4. Only specify very abundant modifications as variable.
5. If the protein was digested with an enzyme, choose this enzyme
6. Use an error tolerant search to find post-translational modifications, SNPs, and non-specific cleavage products
7. For important work, run a target-decoy search, set the peptide FDR to 1%, and filter the proteins by requiring significant matches to 2 distinct sequences

This slide summarises the key points. If you follow these guidelines, you shouldn't go far wrong.