

Search Parameters

MASCOT

MATRIX
SCIENCE

Search Parameters

MASCOT Peptide Mass Fingerprint

Your name: _____ Email: _____

Search title: _____

Database(s): MSIPF_mouse, NCBI, SwissProt, TrEMBL, UnRef100

Enzyme: Trypsin/P

Allow up to: 1 missed cleavages

Taxonomy: All entries

Fixed modifications: Carbamidomethyl (C)

Variable modifications: Oxidation (M)

Protein mass: _____ kDa

Mass values: MH⁺ M_n M-H⁺

Peptide tol. s: 20 ppm

Peptide charge: 2+ and 3+

Peptide mass: _____

MS/MS tol. s: 0.1 Da

Data file: Choose File | No file chosen

Query: NB Contents of this field are ignored if a data file is specified.

Decay: Report top: AUTO hits

Start Search... Reset Form

MASCOT MS/MS Ions Search

Your name: _____ Email: _____

Search title: _____

Database(s): MSIPF_mouse, NCBI, SwissProt, TrEMBL, UnRef100

Enzyme: Trypsin/P

Allow up to: 1 missed cleavages

Taxonomy: All entries

Fixed modifications: Carbamidomethyl (C)

Variable modifications: Oxidation (M)

Peptide tol. s: 20 ppm

Peptide charge: 2+ and 3+

MS/MS tol. s: 0.1 Da

Data format: Mascot generic

Instrument: ESI-TRAP

Decay: Report top: AUTO hits

Precursor: _____ m/z

Error tolerant:

Start Search... Reset Form

MASCOT : Search Parameters

© 2007-2010 Matrix Science



In this presentation, we will describe each of the Mascot search parameters.

If you submit a search from a web browser, you have a choice of three different search forms. All three forms submit to the same search engine, but they have been optimised for three different types of search. The form for a peptide mass fingerprint is shown on the left, and the form for a search of uninterpreted MS/MS data on the right. Most of the controls are common to both.

Search Parameters

The screenshot shows the MASCOT Sequence Query interface. Key elements include:

- Search title:** A text input field.
- Database(s):** A dropdown menu with selected options: MSP1_human, MSP1_mouse, NCBItrf, MASCOT, and Trembl.
- Enzyme:** Trypsin/P.
- Allow up to:** 1 missed cleavages.
- Quantitation:** None.
- Taxonomy:** All entries.
- Fixed modifications:** Carbamidomethyl (C).
- Variable modifications:** Oxidation (M).
- Peptide tol.:** 20 ppm.
- # 13C:** 0.
- MS/MS tol.:** 0.1 Da.
- Peptide charge:** 2+ and 3+.
- Instrument:** ESI-TRAP.
- Decoy:** A checkbox.
- Query:** A large text area for entering search terms.
- Buttons:** Start Search and Reset Form.

MASCOT : Search Parameters

© 2007-2010 Matrix Science



The third form is for a sequence query, such as a sequence tag search. The controls on this form are very similar to those on the MS/MS form. The main difference is that we have a text area to type in the queries, rather than a data file upload control.

Help

PMF ✓ SQ ✓ MS/MS ✓

The screenshot shows the Mascot Peptide Mass Fingerprint search interface. The main window has a search form with fields for 'Your name', 'Email', 'Search title', 'Database(s)', 'Enzyme', 'Allow up to', 'Taxonomy', 'Fixed modifications', 'Variable modifications', 'Protein mass', 'Mass values', 'Data file', 'Peptide tol.', 'Monoisotopic', 'Average', 'Decay', and 'Report top'. A help window is open over the 'Fixed modifications' section, titled 'Matrix Science - Help - Mascot Search Fields - Microsoft Internet Explorer'. The help window contains the following text:

Modifications

Select any known or suspected **modifications**.

Multiple selections can be made by means of the shift and control keys (platform dependent). N.B. If you are using Internet Explorer 3.x, you may find that one modification is selected by default when the form loads. This is a bug in IE3. To clear an item, hold down the control key while clicking on the selected item.

Mascot supports two types of modification. Fixed modifications are applied universally, to every instance of the specified residue(s) or terminus. There is no computational overhead associated with a fixed modification, it is simply equivalent to using a different mass for the modified residue(s) or terminus. For example, selecting Carboxymethyl (C) means that all calculations will use 161 Da as the mass of cysteine.

Variable modifications are those which may or may not be present. Mascot tests all possible arrangements of variable modifications to find the best match. For example, if Oxidation (M) is selected, and a peptide contains 3 methionines, Mascot will test for a match with the experimental data for that peptide containing 0, 1, 2, or 3 oxidised methionine residues.

At the bottom of the help window, there is a yellow button that says 'Click on any link for help'.

MASCOT : Search Parameters

© 2007-2010 Matrix Science



At the top of each slide, there is a key to show which search parameter applies to which type of search.

The labels on the search form are hyperlinks. Just click on them to get detailed help

User details and title

PMF✓ SQ✓ MS/MS✓

Your name	<input type="text" value="Expert User"/>	Email	<input type="text" value="smartie@matrixscience.com"/>
Search title	<input type="text" value="Arabidopsis sample #3476"/>		

- Search form will 'remember' user name and email address in cookie
- If Mascot security is enabled, then this information taken from user database
- Email address used for sending results
- Search title is shown in report, can be searched for in the search log, and (in 2.1 and later) appears in the status screens.

MASCOT : Search Parameters

© 2007-2010 Matrix Science



At the top of the form are a couple of fields for user information. The name and email are saved as a browser cookie when a search is submitted, so you don't need to complete them every time.

If you have an in-house server, and Mascot security is enabled, these fields will be populated automatically with the details of the user who is logged in

When you use the Matrix Science public web site, you have to supply a name and email address. This is to allow the results of a search to be returned by email. Usually, search results are returned promptly to your browser window. However, if your connection to the web site is broken before the search is complete, they will be emailed to the supplied address. If you have an in-house server, you can enable this if you wish. It is turned off by default

The search title is free text. You don't have to enter anything. However, it is a good idea to fill in all of these fields, because it makes it much easier to find your old search results in the search log.


Database PMF ✓ SQ ✓ MS/MS ✓

Database(s)

- MSIPL_mouse
- NCBIInr
- SwissProt
- Trembl
- UniRef100

Choose the right database

- Swiss-Prot good for PMF
- NCBIInr or UniRef100 may be better for MS/MS
- ESTs for MS/MS if no match from protein database

MASCOT : Search Parameters © 2007-2010 Matrix Science 

Choosing the right database is so important that there will be a complete presentation on this topic.

Very briefly, for a peptide mass fingerprint, search a comprehensive, non-redundant database, like SwissProt. If the data are any good, it won't matter if one or two mass values fail to find matches. The advantage of searching a small database is that the search is fast and the reports are concise.

In MS/MS, the most interesting protein in the mixture might be at a very low level and only represented by a single spectrum. So, you don't want to miss a single peptide. You need a non-identical database, where every single peptide is explicitly represented, such as NCBIInr or UniRef100.

The EST databases are huge. Worth trying with high quality MS/MS data if a good match could not be found in a protein database. Not advisable for PMF, because many sequences correspond to protein fragments.

In Mascot 2.3, you can select multiple databases for a search. This is particularly useful when you want to search a single organism database and include the sequences of common contaminants, such as BSA and trypsin. One restriction is that you cannot mix AA and DNA databases.

Taxonomy

PMF ✓ SQ ✓ MS/MS ✓

 All entries

- Speeds up the search
- Simplifies the result report
- The drop-down list is easily configurable.
- Make sure that the taxonomy indexes are kept up to date.

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

If a database contains taxonomy information, we can use this to restrict the search to entries for a particular organism or family. This speeds up the search because, in effect, it makes the database smaller.

Limiting the taxonomy simplifies the result report, because you don't see all the homologous proteins from other species.

The drop down list in the search form is configurable. If you are working on a particular organism, you can easily add this to the list

It is important that the taxonomy is as accurate as possible, which means keeping the indexes up to date

Taxonomy

PMF ✓ SQ ✓ MS/MS ✓

Tax IDs	Count
0	6711
1	3394815
2	6185474
3	429656
4	360699
5	79587
6	109694
7	32185
8	45698
9	22370
10	23009
11	11026
12	14284
13	6859
14	9592

MASCOT : Search Parameters

© 2007-2010 Matrix Science



From time to time, its a good idea to go to the database status page and check the stats file for each database. The stats file contains lots of useful information, like whether entries contain illegal characters or whether an entry is too long.

It also tells you how good your taxonomy is. Here are the numbers for the nr database on our web site in April 2010. There are 10.8 million entries, and 6711 have no taxonomy. In other words, better than 99.9% of the entries have a taxonomy assigned. If you look at your stats file and see that (say) 10% of the entries have no taxonomy, that's 10% of the entries that are going to be missed whenever you do a search with taxonomy specified.

Taxonomy

In most cases, if the correct protein is not in the database, you'd like to see the closest match ... whatever the species

PMF ✓ SQ ✓ MS/MS ✓

```
Time files compressed : Sun Mar 28 06:04:15 2010
Time files compressed (int) : 1569782655
Time / date of fasta file : Tue Mar 23 14:49:00 2010
Time of fasta files (int) : 1569355740
Number of residues : 181677051
Number of sequences : 516081
Number with invalid residues : 0
Number of sequences too long : 0
Length of longest sequence : 35213
Maximum accession length : 11
Version of Mascot : 2.2.105
Version of this file : 4
Seqs with invalid taxon tree : 3
Num sequences for taxonomy : All entries=516081
Num sequences for taxonomy : Archaea (Archaeobacteria)=18197
Num sequences for taxonomy : Eukaryota (eucaryotes)=159476
Num sequences for taxonomy : Alveolata (alveolates)=905
Num sequences for taxonomy : Plasmodium falciparum (malaria parasite)=279
Num sequences for taxonomy : Other Alveolata=656
Num sequences for taxonomy : Metazoa (Animals)=97227
Num sequences for taxonomy : Caenorhabditis elegans=3286
Num sequences for taxonomy : Drosophila (fruit flies)=5141
Num sequences for taxonomy : Chordata (vertebrates and relatives)=81184
Num sequences for taxonomy : bony vertebrates=80606
Num sequences for taxonomy : lobe-finned fish and tetrapod clade=75902
Num sequences for taxonomy : Mammalia (mammals)=64900
Num sequences for taxonomy : Primates=28926
Num sequences for taxonomy : Homo sapiens (human)=20280
Num sequences for taxonomy : Other primates=6646
Num sequences for taxonomy : Rodentia (Rodents)=25329
Num sequences for taxonomy : Mus.=16297
Num sequences for taxonomy : Mus musculus (house mouse)=16246
Num sequences for taxonomy : Rattus=9505
Num sequences for taxonomy : Other rodentia=1527
Num sequences for taxonomy : Other mammalia=12645
Num sequences for taxonomy : Xenopus laevis (African clawed frog)=3227
Num sequences for taxonomy : Other lobe-finned fish and tetrapod clade=7775
Num sequences for taxonomy : Actinopterygii (ray-finned fishes)=4704
Done
```

MASCOT : Search Parameters

© 2007-2010 Matrix Science



A word of warning. Don't specify a very narrow taxonomy in a search.

Think carefully about what you are trying to achieve when you do this.

If the correct protein from the correct species is not in the database, wouldn't you want to see a good match to a protein from a similar species?

This is especially important for poorly represented species. For example, look at these numbers for the Swiss-Prot 2010_04: half a million entries; 25 thousand entries for rodents, but only 1500 are not either mouse or rat. So, even if you are studying hamster or porcupine, you probably don't want to choose 'Other rodentia'.

Enzyme

PMF✓ SQ✓ MS/MS✓

Enzyme

Allow up to missed cleavages

- First choice should normally be the enzyme actually used, and 1 missed cleavage
- Large number of missed cleavages, try increasing to 2
- Use semi-trypsin rather than no enzyme
- No enzyme only in exceptional cases, and never for PMF.

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

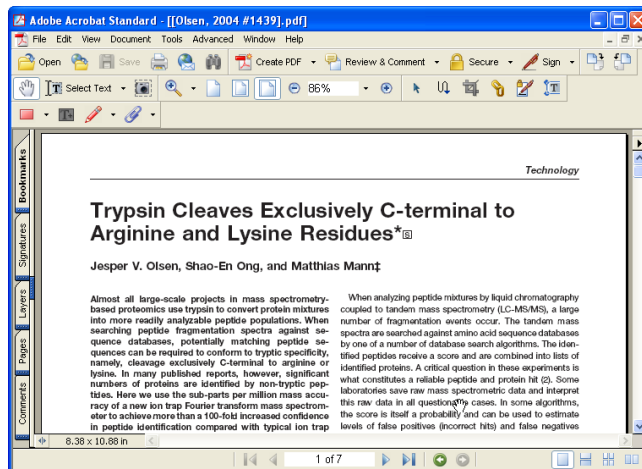
All the search forms have a drop down list for choosing an enzyme. If your peptides come from an enzyme digest, you need to know what the enzyme was and then choose it from the list.

Setting the number of allowed missed cleavage sites to zero simulates a limit digest. If you are confident that your digest is perfect, with no partial fragments present, this will give maximum discrimination and the highest score for a peptide mass fingerprint.

If experience shows that your digest mixtures usually include some partials, that is, peptides with missed cleavage sites, you should choose a setting of 1, or maybe 2 missed cleavage sites. Don't specify a higher number without good reason, because each additional level of missed cleavages increases the number of calculated peptide masses to be matched against the experimental data. In other words, the missed cleavage parameter should be set by looking at some successful search results to see how complete your digests really are.

Enzyme

PMF ✓ SQ ✓ MS/MS ✓



Olsen, J. V., Ong, S.-E. and Mann, M., *Mol. and Cellular Proteomics*, 3, 608-14 (2004)

MASCOT : Search Parameters

© 2007-2010 Matrix Science



Although some people like to perform searches without enzyme specificity, and then gain confidence that a match is correct if the match is tryptic, this isn't a good idea. If there is evidence for a lot of non-specific cleavage, then a semi-specific enzyme allows one end of the peptide to be non-specific, but not both. Only abandon enzyme specificity if you have no other choice, such as when searching endogenous peptides.

You cannot perform a no-enzyme peptide mass fingerprint. It simply won't work, even if you have good mass accuracy

There is some controversy over the level of non-specific peptides that can be expected in a tryptic digest. Our experience is that the levels of non-specific peptides are very low, less than 3%, unless there is something seriously wrong with the trypsin or the protocol.

Why do we advise so strongly against no-enzyme searches?

Enzyme

PMF✓ SQ✓ MS/MS✓

plc dataset on dual processor 2.8 GHz P4				
CLE	peptides tested	minutes	identity matches	average threshold
trypsin	7.5E+07	10	399	41
semi-trypsin	1.2E+09	127	379	53
none	1.0E+10	1067	299	62

MASCOT : Search Parameters

© 2007-2010 Matrix Science



Here are some numbers for a typical ion trap dataset when we search using loose trypsin, semi-specific trypsin, and no enzyme specificity

As you can see, the no enzyme search takes a lot longer and we get fewer reliable matches.

The reason is simple, the search space for a no-enzyme search is much, much larger than for a tryptic search. This means that the thresholds are higher and we lose marginal matches. Unless you have a high level of non-specific peptides, you lose more than you gain.

So, doing a no-enzyme search in Mascot is not a good idea unless there is a very high level of non-specific peptides. Semi-trypsin is almost always a better choice if the peptides came from a tryptic digest. Only use no enzyme if the peptides are not the products of a deliberate enzyme digest, e.g. MHC peptides or endogenous peptides.

Enzyme

PMF ✓ SQ ✓ MS/MS ✓

Title	Sense	Cleave at	Restrict	Independent	Semispecific	
Trypsin	C-Term	KR	P	no	no	Edit Delete
Arg-C	C-Term	R	P	no	no	Edit Delete
Asp-N	N-Term	BD		no	no	Edit Delete
Asp-N_ambic	N-Term	DE		no	no	Edit Delete
Chymotrypsin	C-Term	FLWY	P	no	no	Edit Delete
CNBr	C-Term	M		no	no	Edit Delete
CNBr+Trypsin	C-Term	M		no	no	Edit Delete
	C-Term	KR	P			
Formic_add	N-Term	D		no	no	Edit Delete
	C-Term	D				
Lys-C	C-Term	K	P	no	no	Edit Delete
Lys-C/P	C-Term	K		no	no	Edit Delete
PepsinA	C-Term	FL		no	no	Edit Delete
Tryp-CNBr	C-Term	KMR	P	no	no	Edit Delete
TrypChymo	C-Term	FKLRWY	P	no	no	Edit Delete
Trypsin/P	C-Term	KR		no	no	Edit Delete
V8-DE	C-Term	BDEZ	P	no	no	Edit Delete
V8-E	C-Term	EZ	P	no	no	Edit Delete
semiTrypsin	C-Term	KR	P	no	yes	Edit Delete
LysC+AspN	N-Term	BD		no	no	Edit Delete
	C-Term	K	P			
None						

MASCOT : Search Parameters

© 2007-2010 Matrix Science



The list of enzymes is user configurable. Standard entries are described in the help. If you wish, you can modify the definitions or create new ones using the configuration editor. The configuration editor was new in Mascot 2.2, so if you are using an earlier version of Mascot, you'll need to edit the configuration file called enzymes in a text editor. The format is described in the Mascot Installation and Setup Manual.

Mascot supports two categories of mixed enzyme definitions. An independent mixed enzyme is used where multiple sample aliquots have been digested separately, and the digests combined for analysis. This means that the sample could contain (say) tryptic peptides and Asp-N peptides, but no peptides that are tryptic at one end and Asp-N at the other. The second category simulates a single sample aliquot being digested simultaneously or serially by more than one cleavage agent. For example CNBr followed by trypsin.

Enzyme

PMF ✓ SQ ✓ MS/MS ✓

MASCOT Sequence Query

Your name: Expert User, Email: ama@se@matrixscience.com

Search title: Glu Fib

Databases: MSP1_mouse, NCBInr, UniProt

Enzyme: TrypsinP

Allow up to: 1 missed cleavages

Taxonomy: All entries

Fixed modifications: none selected

Variable modifications: none selected

Peptide tol.: 0.3, MS/MS tol.: 0.5

Peptide charge: 1+

Query: 1569.7 seqM-QGNGN

Start Search, Reset Form

MASCOT : Search Parameters

© 2007-2010 Matrix Science



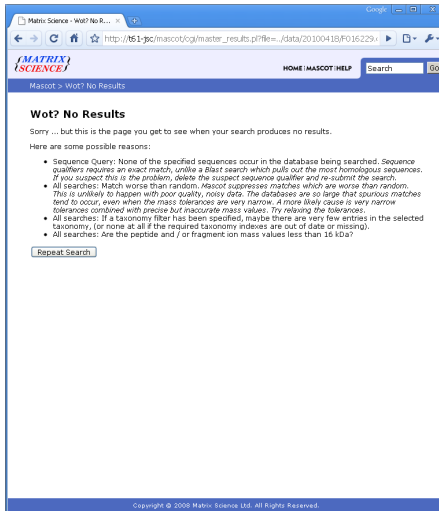
Remember that enzyme specificity also applies to Sequence Queries

One of the most common emails we receive is "Mascot is broken. I did a search for this peptide and I know its in the database but Mascot failed to find it"

For example, here's a search for glu-fib, a very common sequencing standard. The mass is correct and the sequence is correct. But, when we do a search of Swiss-Prot -

Enzyme

PMF ✓ SQ ✓ MS/MS ✓



MASCOT : Search Parameters

© 2007-2010 Matrix Science



No results!
Why?

Enzyme

PMF ✓ SQ ✓ MS/MS ✓

Mascot Search Results

Protein View

Match to: **FIBR_HUMAN** Score: 72 Expect: 0.017
Fibrinogen beta chain precursor [Contains: Fibrinopeptide B] - Homo sapiens (Human)

Nominal mass (M_0): 55892; Calculated pI value: 8.54
NCBI BLAST search of **FIBR_HUMAN** against nr
Unformatted [sequence listing](#) for pasting into other applications

Taxonomy: [Homo sapiens](#)

No enzyme cleavage specificity
Sequence Coverage: 2%

Matched peptides shown in **Bold Red**

```
1 MFRKVSVFF KLKTNHLLL LLLCVLVES QVNDREKFF FSARGRPLD
31 KFEELAPLR PAPPVSDGG YHAPARAA TQVYKRAF DAGCCLRAQ
101 IDGLDPTQC QGKALQGF FIFNSPEEL NNNYFASQT SSSSPVTVL
151 LFDLWQKQK QVQDNHNVN EYSELEKQ LTIETVNSH IFTNLVLR
201 TLNLRRIQ KLEDSVSAK EYQVTPCTP CHIPVYQK CEIIRKQK
251 TSEMTLQFD SSVKPYVVC DNTNNGOUT VIGRQGGV DFRKNDPK
301 QQFQVATIT DQNYVCLPG EYVLRNDIS QLTNRPTEL LIERDRGSD
351 KQKATQGGT VQKARVQK DNVVPTGIG HALLRQKQ KRNRTTTH
401 NQFFETTER DNDGLTSDP KQCKREKQK GVVNDFCAA NFNQRYGGG
451 QTTDRKRG TDQGVVNVN KQWTRRKM SKKIPFFPQ Q
```

Sort Peptides By: Residue Number Increasing Mass Decreasing Mass

Start	End	Observed	H(e)xp(t)	H(e)ca(e)	Delta	Miss	Sequence
31	44	1569.7010	1568.4927	1568.4955	0.0072	0	S-GQVNDREKFFSAR.G (No match)

MASCOT : Search Parameters

© 2007-2010 Matrix Science



Because glu-fib in Swiss-Prot is not a tryptic peptide. The N-terminus is created by a post-translational cleavage after serine. If you now go back to the search form and select enzyme type none, bingo ... you'll get a match

Modifications

PMF ✓ SQ ✓ MS/MS ✓

The screenshot shows the 'Modifications' section of the Mascot search interface. It features two columns of modification lists. The left column has a 'Fixed modifications' section with a dropdown menu currently showing 'Carbamidomethyl (C)' and a 'Variable modifications' section with a dropdown menu showing 'Oxidation (M)'. Between these columns is a checkbox labeled 'Display all modifications' which is currently unchecked. To the right of these sections are two sets of arrow buttons (> and <) for moving items between lists. On the far right is a large list box containing a scrollable list of modification names: Acetyl (K), Acetyl (N-term), Acetyl (Protein N-term), Amidated (C-term), Amidated (Protein C-term), Ammonia-loss (N-term C), Biotin (K), Biotin (N-term), Carbamyl (K), Carbamyl (N-term), and Carboxymethyl (C).

- Get details of current modifications, download updates, and define new entries at <http://www.unimod.org>
- User definable with an in-house Mascot installation

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

This screen shot shows how modifications are displayed in the search form in Mascot 2.3. If you are using an earlier version, there are just two list boxes, one for fixed modifications and one for variable. In 2.3, you move modifications from the single list on the right to and from the lists on the left. This makes it easier to see at a glance what has been selected for the search. If the checkbox labelled 'Display all modifications' is clear, as shown here, you get a relatively short list of the most common modifications. If you check the box, a much longer list is available. In Mascot 2.2 and earlier, this checkbox could be found on the search form defaults page, which we will mention later

You can keep your list of modifications up-to-date by downloading the latest information from Unimod. If you have a modification which you don't want to share with others, you can add it to the local configuration file. We'll describe how to go about doing this in detail in the Mascot Server Administration talk.

Modifications

PMF✓ SQ✓ MS/MS✓

Modifications

- Fixed / static modifications cost nothing
- Variable / differential modifications are very expensive
- Use minimum variable modifications, especially for PMF
 - Maybe oxidation of M
 - Maybe alkylation of C

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

Modifications in database searching are handled in two ways. First, there are the fixed or static or quantitative 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 or differential or non-quantitative modifications are expensive in the sense that they increase the time taken for a search and reduce its specificity. This is 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. The so-called combinatorial explosion.

Hence, it is very important to be as sparing as possible with variable modifications. Especially in a peptide mass fingerprint, where the increase in the number of calculated peptides quickly makes it impossible to find a statistically significant match.

Quantitation

PMF✗ SQ✓ MS/MS✓

Quantitation

•More later ...

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

Quantitation is the subject of a complete presentation.

Protein mass

PMF✓ SQ✗ MS/MS✗

Protein mass kDa

- Applied as sliding window because there is no guarantee that the database entry represents the processed protein
- Slows down the search
- Never useful for MS/MS search. Only useful for Peptide Mass Fingerprint when
 - Analyte is small fragment of very large entry
 - Low complexity entry.

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

The protein mass is the mass of the intact protein in kDa applied as a sliding window. That is, the mass of the contiguous stretch of sequence which contains all of the matched peptide mass values. This will generally be less than the mass of the entire sequence entry. Consequently, if you specify a value for the protein mass, this acts only as a ceiling. Not only will you see smaller proteins on the hit list, you will also see larger ones, but all of the reported matches will be within a stretch of sequence less than or equal to the specified mass.

If this field is left blank, there is no restriction on protein mass

Specifying a protein mass will slow down the search a little.

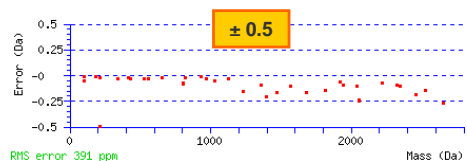
Its hard to find examples where this parameter is useful. We include it mainly because many people requested it. It could give a better score if the analyte was small fragment of very large entry, or a low complexity protein. But, you can't know this in advance, so our general recommendation is to leave the protein mass open

Peptide tolerance

PMF ✓ SQ ✓ MS/MS ✓

Peptide tol. ± 0.3 Da

Specifying too tight a mass tolerance is the most common reason for failing to get a match



MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

This is the error window on experimental peptide mass values, not the error window for MS/MS fragment ion mass values, which is set using the MS/MS tol. \pm parameter.

Units can be selected from: percentage, milli-mass units, parts per million, or Daltons.

Specifying too tight a tolerance is the most common reason for failing to get a match.

Making an estimate of the mass accuracy doesn't have to be a guessing game. Protein View includes a graph of the mass errors for intact peptides. Just search a strong standard and look at the error graph. You'll normally see some kind of trend. Add on a safety margin and this is your error estimate. If you see something that looks like this, a mass tolerance of ± 0.5 Da is about right. It gives some safety margin. Remember that there will always be the odd outlier, like the data point at the lower left. It is the general trend and distribution of the majority of the data points that is important.

For a peptide mass fingerprint, the score depends on the peptide tolerance. In an MS/MS search, this parameter has no effect on the ions score. However, it does affect the search time. The larger the tolerance, the longer the search will take.

Peptide tolerance ¹³C

PMF ✗ SQ ✗ MS/MS ✓

¹³C 0

Sometimes, peak detection chooses the ¹³C peak

The normal test for a precursor match is:

$$\text{TOL} > \text{absolute}(\text{exp} - \text{calc})$$

If this field is set to 1, the test will also succeed for

$$\text{TOL} > \text{absolute}(\text{exp} - \text{calc} - 1)$$

If this field is set to 2, the test will succeed for the above two conditions, plus:

$$\text{TOL} > \text{absolute}(\text{exp} - \text{calc} - 2)$$

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

Sometimes, peak detection chooses the ¹³C peak rather than the ¹²C. In extreme cases, it may pick the ¹³C₂ peak. The normal test for a precursor match is:

$$\text{TOL} > \text{absolute}(\text{exp} - \text{calc})$$

Assuming the mass values and tolerance are in Da, if this field is set to 1, the test will also succeed for

$$\text{TOL} > \text{absolute}(\text{exp} - \text{calc} - 1)$$

If this field is set to 2, the test will succeed for the above two conditions, plus:

$$\text{TOL} > \text{absolute}(\text{exp} - \text{calc} - 2)$$

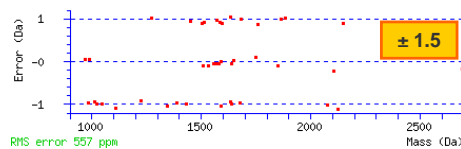
This means that you can use a tight mass tolerance and still get a match to a ¹³C peak. If you are using a very high accuracy instrument, note that the precise shifts are the carbon isotope spacings of 1.00335 and 2.00670, rather than 1 and 2.

MS/MS tolerance

PMF✗ SQ✓ MS/MS✓

MS/MS tol. ± 0.3 Da

Specifying too tight or too loose a mass tolerance will reduce the ions score



MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

This is the error window on MS/MS fragment mass values.

Units can be milli-mass units or Daltons.

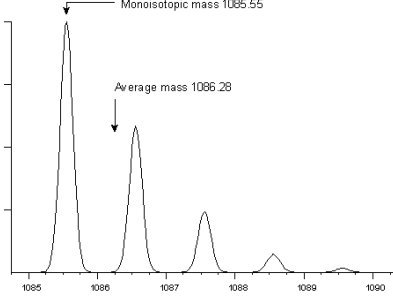
Specifying too tight or too loose a mass tolerance will reduce the ions score. Peptide View includes a graph of the mass errors for fragment ions.

Here, the mass tolerance is much too high. A more appropriate tolerance might be ± 0.5 . Having a tolerance which is much too high can sometimes lead to artefacts and false positives


Mass type

PMF ✓ SQ ✓ MS/MS ✓

Monoisotopic
Average



If you get this setting wrong, the mass errors will be very large and show a strong trend

MASCOT : Search Parameters © 2007-2010 Matrix Science 

Mass type specifies whether the experimental mass values are average or monoisotopic. Monoisotopic mass is the mass of the peptide where all atoms are the most abundant natural isotopes of their elements, e.g. Carbon 12, Nitrogen 14, Hydrogen 1, etc. In most cases, this is the first peak of the natural isotope distribution. Average mass is the chemical mass, which is the centre of gravity of the isotope distribution.

In Mascot, you cannot mix the two, and have (say) average precursors and monoisotopic fragments.

Most modern instruments produce monoisotopic mass values. You will only have an average mass 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%.

Charge

PMF ✓ SQ ✓ MS/MS ✓

Mass values
 MH⁺
 M_r
 M-H⁻

Peptide charge

- 1+ means MH⁺, 1- means M-H⁻, etc.
- For MS/MS, this setting is a default, which is rarely used.

MASCOT : Search Parameters © 2007-2010 Matrix Science

These fields are used to specify the peptide charge state. The radio buttons are from the peptide mass fingerprint form. The drop down list is used on the sequence query and MS/MS forms.

The notation "1+", "2+", etc. is used to save space and because some HTML form fields do not support the use of superscripts and subscripts. "1+" always means MH⁺, "1-" always means M-H⁻, etc.

For MALDI-PSD, the precursor peptides will generally be MH⁺, so the charge state should be set to "1+"

For an MS/MS search, the value specified here is a default. Most peak lists always specify a charge state, so default is never used.

Data (PMF)

PMF ✓ SQ ✓ MS/MS ✗

The screenshot shows a web form with two main sections. The top section is labeled 'Data file' and contains a text input field followed by a 'Browse...' button. The bottom section is labeled 'Query' and contains a large text area. To the left of the 'Query' text area, there is a note: 'NB Contents of this field are ignored if a data file is specified.'

- Mass [intensity] [additional text]
- Applied Biosystems Data Explorer (.pkm)
- Bruker Analysis AutoXecute Data Report
- Bruker XML
- mzData (1.05)
- .mzML

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

The contents of the query window on the peptide mass fingerprint form are only used when no data file has been specified.

The data format for a peptide mass fingerprint is auto detected. It can be a simple list of mass values, one per line. If a second values is present, it is assumed to be intensity. Any further values on the same line are ignored

Mascot also supports other peak list formats, as listed.

mzData is the standard interchange format sponsored by the HUPO Proteomics Standards Initiative working group

Data (MS/MS)

PMF ✗ SQ ✗ MS/MS ✓

Data file	<input type="text"/>	Browse...
Data format	Mascot generic	Precursor <input type="text"/> m/z

- Mascot Generic Format (.MGF)
- Finnigan (.ASC)
- Sequest (.DTA)
- PerSeptive (.PKS)
- Micromass (.PKL)
- Sciex API III
- Bruker (.XML)
- mzData (.XML)
- mzML (.mzML)

MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

Data for MS/MS ion searches must be supplied as an ASCII file in one of these supported formats. The format cannot be auto-detected, and must be specified using the drop down list.

Certain data file formats, SCIEX API III, PerSeptive (.PKS), and Bruker (.XML), do not include m/z information for the precursor peptide. For these formats only, the Precursor field is used to specify the m/z value of the parent peptide.

A data file may include embedded search parameters. Most embedded parameters can only appear once, at the head of the data file. In a Mascot generic format file, a few parameters can appear within an MS/MS dataset. See the Data File Format help page for further details

If there is a conflict between the values of the embedded parameters and values entered into search form fields, the embedded parameters always take precedence. The search form fields are essentially defaults for values missing from the data file.

Instrument

PMF ✗ SQ ✓ MS/MS ✓

Instrument ESI-QUAD-TOF

- Click on the help link to see which ions series are used



MASCOT : Search Parameters

© 2007-2010 Matrix Science

MATRIX
SCIENCE

For an MS/MS Ions Search, choose the description which best matches the type of instrument used to acquire the data. This setting determines which fragment ion series will be used for scoring, according to the following table.

Instrument

PMF ✗ SQ ✓ MS/MS ✓

The screenshot shows a web browser window titled "Mascot configuration - Microsoft Internet Explorer" displaying the "Mascot Configuration: Instruments" page. The page contains a table with columns for various instrument types and ion series. The table is as follows:

Ion series	Default	ESI QUAD TOF	MALDI TOF PICO	ESI TRAP	ESI QUAD	ESI FTICR	MALDI TOF TOF	ESI 4SECTOR	FIMS ECD	ETD TRAP	MALDI QUAD TOF	MALDI QIT TOF	ETD-W
1+	X	X	X	X	X	X	X	X	X	X	X	X	X
2+	X	X		X	X	X			X	X	X		X
(precursor-3+)			X				X	X			X	X	
mmonium	X	X					X	X					X
a	X	X					X	X					X
a*	X	X					X	X					X
ad		X					X	X					X
b	X	X	X	X	X	X	X	X					X
b*	X	X	X	X	X	X	X	X					X
bd	X	X	X	X	X	X	X	X					X
c									X	X			X
i													
y	X	X	X	X	X	X	X	X	X	X	X	X	X
y*	X	X	X	X	X	X	X	X					X
y0		X	X	X	X	X	X	X					X
z									X				X
zb							X	X					X
z0							X	X					X
y must be significant													
y must be highest score									X	X			X
d							X						
v							X						X
w							X						X
z+2									X	X			X
Minimum mass													
Max mass	700,000	700,000	700,000	700,000	700,000	700,000	700,000	700,000	700,000	700,000	700,000	700,000	700,000
	Delete	Delete	Delete	Delete	Delete	Delete	Delete	Delete	Delete	Delete	Delete	Delete	Delete
	Edit	Edit	Edit	Edit	Edit	Edit	Edit	Edit	Edit	Edit	Edit	Edit	Edit

MASCOT : Search Parameters

© 2007-2010 Matrix Science



"Default" corresponds to the configuration used in Mascot version 1.7 and earlier.

Many of the instruments are very similar.

You can modify instrument settings or create new ones using the configuration editor. In this screenshot, the right hand column is an experiment to see how the addition of w ions affects ETD matching

Error Tolerant

PMF ✗ SQ ✗ MS/MS ✓

Error tolerant

MASCOT : Search Parameters

© 2007-2010 Matrix Science



If you have MS/MS data, and are interested in finding post-translational modifications, you can perform an error tolerant search by checking this box on the search form. This is a much more efficient way to discover unusual modifications, as well as non-specific peptides and sequence variants. More about this in a later presentation.

Decoy

PMF ✓ SQ ✓ MS/MS ✓

Decoy

The screenshot shows a Microsoft Internet Explorer browser window displaying a search results page from the Molecular & Cellular Proteomics website. The page title is "PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATION". A search bar at the top right contains the text "Decoy". Below the search bar, a table is displayed with the following data:

	Sprot	Decoy	False discovery rate
Peptide matches above identity threshold	3290	8	0.24 %
Peptide matches above homology or identity threshold	6037	224	3.71 %

The table is highlighted with a red border. Below the table, there is a red-bordered box containing the following text: "For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, e.g., the results of randomized database searches or other computational approaches."

MASCOT : Search Parameters

© 2007-2010 Matrix Science



The decoy checkbox enables you to validate the false discovery rate according to the approach recommended in the Molecular & Cellular Proteomics Guidelines for Publication: “For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, e.g., the results of randomized database searches or other computational approaches”

Report

PMF✓ SQ✓ MS/MS✓

Report top | AUTO hits

Report top should normally be set to auto.

MASCOT : Search Parameters

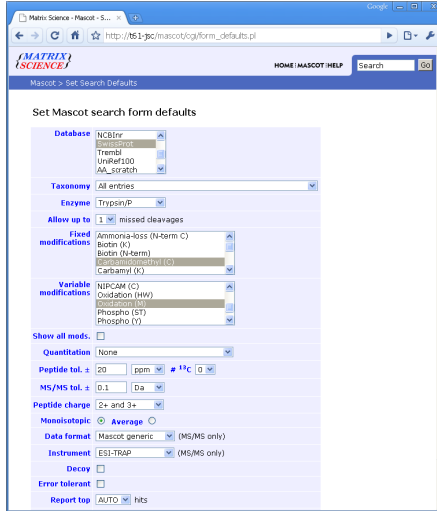
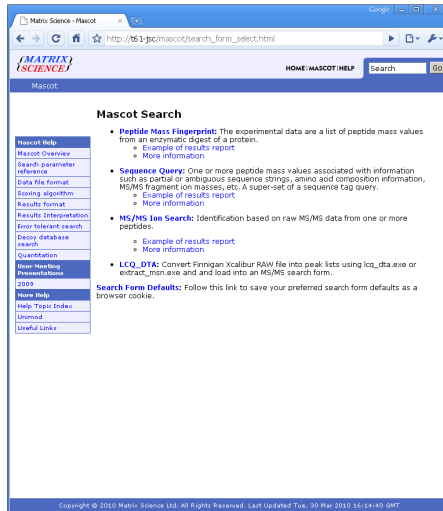
© 2007-2010 Matrix Science

MATRIX
SCIENCE

REPORT determines the *maximum* number of hits displayed in a search results report. Choose AUTO to display only protein hits with significant scores. In a protein summary report, one additional hit is reported after the cutoff at the significant score. This is to ensure that the report shows the highest scoring hit, even though it is not significant.

Setting defaults

PMF ✓ SQ ✓ MS/MS ✓



MASCOT : Search Parameters

© 2007-2010 Matrix Science



You can choose your own defaults for the search forms. Look for the link at the bottom of the search form selection page

In particular, if you are using Mascot 2.2 or earlier, this is where you choose whether to display the full modifications list or just a short list of the most common mods

When you save the defaults, they are saved as a browser cookie. If you go to a different PC, or switch to a different browser, you'll need to repeat this step

Final Tip

DANGER!

- Iteratively adjusting search parameters to get a better score can give misleading results
- Beware of
 - Narrowing the taxonomy
 - Reducing mass tolerances
 - Removing modifications
 - Selecting spectra or mass values

Set search parameters using standard samples

A final word of advice: It is easy to distort the search results without realising.

Basically, it is risky to adjust the search parameters interactively to get a better score for an unknown.

For example, you search the complete database and don't get a significant match. However, a very interesting looking protein is near the top of the list, surrounded by some others that are clearly wrong. You change the taxonomy filter so as to exclude the "wrong" proteins. Sorry, but this is cheating.

Search parameters should be set using standards. Broadening the search if you get a negative result is usually OK, but not narrowing the search.