

Statistical Significance in Error Tolerant Search Results

MASCOT : *Error tolerant search statistics* © 2021 Matrix Science



I'd like to talk about some improvements we've made in the reporting of Error Tolerant search results

Error Tolerant Search: Overview

- First pass: standard search of specified database(s)
- Second pass: database entries that contain one or more significant peptide matches are selected and searched with
 - Relaxed enzyme specificity
 - Comprehensive list of chemical and post-translational modifications
 - Single residue substitutions or single nucleotide substitutions, insertions and deletions.

MASCOT : *Error tolerant search statistics* © 2021 Matrix Science



An error tolerant search is the most efficient way to find unsuspected modifications, non-specific cleavage products, and sequence variants.

MASCOT MS/MS Ions Search

Your name: Email:

Search title:

Database(s):

Taxonomy:

Enzyme: Allow up to: missed cleavages

Quantitation:

Crosslinking:

Fixed modifications:

Variable modifications:

Peptide tol. ±: ppm # ¹³C:

MS/MS tol. ±: Da

Peptide charge: Monoisotopic Average

Data file: No file chosen

Data format:

Instrument:

Precursor: m/z

Error tolerant

Target PSM FDR:

Decoy

MASCOT : Error tolerant search statistics © 2021 Matrix Science 

The big difference in Mascot Server 2.8 is that we now use target-decoy to assign significance to the all matches, including those found in the second pass search.

As in previous releases, all you need to do to perform an error tolerant search is to check a box on the search form.

In Mascot Server 2.8, you can (and should) also check the box to use target-decoy. Without a decoy, expect values are derived from counting trials – that is, the number of candidate peptides that have been tested. This estimate is not always accurate; particularly when there is something wrong with the choice of database or search parameters, making a large fraction of potential matches unavailable. Ticking the checkbox to search a decoy database gives a solid, empirical basis for the statistics.

There is also a control to specify the required false discovery rate. The reason we ask for is up front is that the FDR determines the set of proteins selected for the second pass search. For example, the first pass search might identify significant peptide matches to 500 proteins at an FDR of 5%, and these are sent through to the second pass. If the FDR was reduced to 1%, the number of proteins selected for the second pass might drop to 400. Although the FDR can be tweaked at the report stage, this will not give perfectly identical results to setting the required FDR in the search form.

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

	Target	Decoy	FDR
Protein family members	59	0	0.00%
PSMs	4279	42	0.98%

Note: Protein FDR 0% means there are not enough decoy protein hits for a meaningful FDR calculation.

Significance threshold for first pass search is **0.02075**, and second pass search **0.05448**. Target PSM FDR from combined first and second pass searches is **1%**.
Decoy results are available in [the decoy report](#).

Proteins (59) [Report Builder](#) [Unassigned \(22268\)](#)

Protein families 1-10 (out of 44)

10 per page 1 2 3 4 5 [Next](#) [Expand all](#) [Collapse all](#)

Accession contains Find Clear

▶ 1

▶ 2

1 1::P04264	12581	SWISS-PROT:P04264 Tax_Id=9606 Gene_Symbol=KRT14
2 1::P35908	5826	SWISS-PROT:P35908 Tax_Id=9606 Gene_Symbol=KRT24
3 1::P02538	1977	SWISS-PROT:P02538 Tax_Id=9606 Gene_Symbol=KRT6A
4 1::P04259	1772	SWISS-PROT:P04259 Tax_Id=9606 Gene_Symbol=KRT5B
5 1::P13647	1532	SWISS-PROT:P13647 Tax_Id=9606 Gene_Symbol=KRT5A

1 2::HSP72_YEAST	10625	Heat shock protein SSA2 OS=Saccharomyces cerevisiae (
2 2::HSP71_YEAST	10283	Heat shock protein SSA1 OS=Saccharomyces cerevisiae (
3 2::HSP74_YEAST	5160	Heat shock protein SSA4 OS=Saccharomyces cerevisiae (
4 2::HSP73_YEAST	4255	Heat shock protein SSA3 OS=Saccharomyces cerevisiae (
5 2::BIP_YEAST	2728	Endoplasmic reticulum chaperone BIP OS=Saccharomyces
6 2::SSB1_YEAST	1991	Ribosome-associated molecular chaperone SSB1 OS=Sacc

MASCOT : Error tolerant search statistics © 2021 Matrix Science

When the results come back, you have a single report that combines the results from both passes.

The required FDR is applied independently to the results from the first and second pass searches. Since this is based on counts of PSMs, it also holds true for the combined results.

▼954 peptide matches (173 non-duplicate, 781 duplicate)

Auto-fit to window

Query Dupes	Observed	Mr(expt)	Mr(calcd)	ppm	M Score	Expect	Rank	U	Peptide
#12574 ▶2	804.4050	1606.7955	1606.8025	-4.31	0	81	3e-06	▶1	U N.FNGNTLNDNDIMLIK.L
#12741 ▶1	812.3828	1622.7511	1622.7536	-1.51	0	39	0.04	▶1	U R.LGEHNIDVLEGNQ.F + Carbamidomethyl (N-term)
#13143 ▶4	827.3561	1652.6976	1652.6923	3.23	0	84	8.1e-07	▶1	U R.SCAAAGTECLISGWGN.T
#13307 ▶1	830.9304	1659.8463	1659.8468	-0.25	0	68	6e-05	▶1	U N.IDVLEGNQFINAAK.I
#13830 ▶1	855.8650	1709.7153	1709.7137	0.94	0	65	5.8e-05	▶1	U R.SCAAAGTECLISGWGN.T + Carbamidomethyl (N-term)
#13877 ▶5	857.4082	1712.8018	1712.8006	0.70	0	58	0.0006	▶1	U R.LGEHNIDVLEGNQ.F.I
#14490 ▶4	883.8943	1765.7741	1765.7764	-1.29	0	48	0.005	▶1	U R.SCAAAGTECLISGWGNK.S + 2 [-1.0078 at C2,C9]
#14608 ▶8	887.9519	1773.8892	1773.8897	-0.25	0	113	2.2e-09	▶1	U H.NIDVLEGNQFINAAK.I
#14772	896.4172	1790.8199	1790.8258	-3.25	0	57	0.00082	▶1	U R.SCAAAGTECLISGWGNK.S + [-33.9877 at C9]
#14785 ▶2	897.4366	1792.8586	1792.8566	1.11	0	88	6e-09	▶1	U K.VCNVYVNIQQITAAK.-
#15193 ▶5	912.4043	1822.7940	1822.7978	-2.10	0	57	0.00066	▶1	U R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term); 2 [-1.0078 at C2,C9]
#15293 ▶3	916.4605	1830.9065	1830.9111	-2.56	0	93	2.6e-07	▶1	U H.NIDVLEGNQFINAAK.I + Carbamidomethyl (N-term)
#15889 ▶9	941.9230	1881.8313	1881.8349	-1.90	0	114	8.5e-12	▶1	U R.SCAAAGTECLISGWGNK.S
#15890	628.2845	1881.8317	1881.8349	-1.72	0	48	3.5e-05	▶1	U R.SCAAAGTECLISGWGNK.S
#15914	628.6180	1882.8323	1882.8189	7.09	0	44	0.014	▶1	U R.SCAAAGTECLISGWGNK.S + [+0.9840 at M16]
#16103 ▶3	948.9313	1895.8480	1895.8506	-1.38	0	112	2.7e-09	▶1	U R.SCAAAGTECLISGWGNK.S + [+14.0156 at C-term K]
#16220	955.9276	1909.8407	1909.8298	5.70	0	79	4.4e-06	▶1	U R.SCAAAGTECLISGWGNK.S + [+27.9949 at T17]
#16225	637.6266	1909.8581	1909.8662	-4.27	0	54	0.0018	▶1	U R.SCAAAGTECLISGWGNK.S + [+28.0313 at M16]
#16242 ▶4	637.6288	1909.8645	1909.8662	-0.90	0	60	0.00041	▶1	U R.SCAAAGTECLISGWGNK.S + [+28.0313 at C-term
#16244 ▶9	955.9400	1909.8654	1909.8662	-0.45	0	117	8.5e-10	▶1	U R.SCAAAGTECLISGWGNK.S + [+28.0313 at C-term K]
#16287	956.4820	1910.9494	1910.9486	0.43	0	47	0.01	▶1	U E.HNIDVLEGNQFINAAK.I
#16330 ▶1	957.9163	1913.8181	1913.8070	5.79	0	77	5.8e-06	▶1	U R.SCAAAGTECLISGWGNK.S + [+31.9721 at M14]
#16432 ▶1	962.9273	1923.8400	1923.8567	-8.68	0	85	1.2e-06	▶1	U R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term); [+42.0218 at C2]
#16434	642.2883	1923.8430	1923.8567	-7.13	0	50	0.0035	▶1	U R.SCAAAGTECLISGWGNK.S + [+42.0218 at C-term K]
#16469 ▶1	963.9348	1925.8551	1925.8611	-3.12	0	65	0.00011	▶1	U R.SCAAAGTECLISGWGNK.S + [+44.0262 at G15]
#16485	964.9608	1927.9069	1927.9098	-1.48	0	45	0.015	▶1	U K.IYHPVFNQNTLNDIM.L
#16568	969.9386	1937.8627	1937.8611	0.80	0	47	0.0085	▶1	U R.SCAAAGTECLISGWGNK.S + [+56.0262 at S12]
#16581 ▶4	970.4324	1938.8502	1938.8564	-3.21	0	85	4.7e-09	▶1	U R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term)
#16582	647.2907	1938.8503	1938.8564	-3.15	0	28	0.002	▶1	U R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term)

Possible assignments:
Methyl (C-term) [+14.0156]
Methyl (K) [+14.0156]

MASCOT : Error tolerant search statistics © 2021 Matrix Science



Expect values are reported for both first and second pass matches. In earlier versions of Mascot, an error tolerant search could not be combined with target-decoy, and expect values based on counting trials were only reported for first pass matches.

Error Tolerant Search: Target-Decoy

- **The target and decoy proteins are treated as pairs**
 - Target and decoy databases are of identical size
 - All significant peptide matches (PSMs) from the first pass are represented
- **Blindly discard second pass results for queries that get a significant match in the first pass search**

MASCOT : *Error tolerant search statistics* © 2021 Matrix Science



The way it works is that target and decoy proteins are treated as pairs. After the first pass search, when proteins are selected, each significant match, whether target or decoy, causes the relevant pair of target and decoy proteins to be selected for the second pass. This means that the target and decoy databases are of identical size and contain all significant peptide matches (PSMs) from the first pass.

If a query gets a significant match in the first pass search, this is what we report, and we blindly discard the second pass results for this query. Sometimes, this means a stronger match is missed, but to do otherwise would be statistically dishonest. For example, if the significance threshold for a particular query in the first pass search corresponds to a score of 40, and we get a match with a score of 52, this is what we report, even if the second pass search might give us an even better match. This is not ideal, but the alternative is to burden all matches with statistics based on both passes. To illustrate why this is a problem, imagine we were to look at the second pass results and find nothing better. Now, we have a larger search space and the score threshold has increased to 55, so we have to discard our first pass match with a score of 52 because it is no longer significant.

▼954 peptide matches (173 non-duplicate, 781 duplicate)

Auto-fit to window

Query Dupes	Observed	Mr (expt)	Mr (calc)	ppm	M	Score	Expect	Rank	U	Peptide
#12574 ▶2	804.4050	1606.7955	1606.8025	-4.31	0	81	3e-06	▶1	U	N.FNGNTLNDIMLIK.L
#12741 ▶1	812.3828	1622.7511	1622.7536	-1.51	0	39	0.04	▶1	U	R.LGEINIDVLEGNQ.F + Carbamidomethyl (N-term)
#13143 ▶4	827.3561	1652.6976	1652.6923	3.23	0	84	8.1e-07	▶1	U	R.SCAAAGTECLISGWGN.T
#13307 ▶1	830.9304	1659.8463	1659.8468	-0.25	0	68	6e-05	▶1	U	N.IDVLEGNQFVNAAK.I
#13830 ▶1	855.8650	1709.7153	1709.7137	0.94	0	65	5.8e-05	▶1	U	R.SCAAAGTECLISGWGN.T + Carbamidomethyl (N-term)
#13877 ▶5	857.4082	1712.8018	1712.8006	0.70	0	58	0.0006	▶1	U	R.LGEINIDVLEGNQF.I
#14490 ▶4	883.8943	1765.7741	1765.7764	-1.29	0	48	0.005	▶1	U	R.SCAAAGTECLISGWGNK.S + 2 [-1.0078 at C2,C9]
#14608 ▶8	887.9519	1773.8892	1773.8897	-0.25	0	113	2.2e-09	▶1	U	H.NIDVLEGNQFVNAAK.I
#14772	896.4172	1790.8199	1790.8258	-3.25	0	57	0.00082	▶1	U	R.SCAAAGTECLISGWGNK.S + [-33.9877 at C9]
#14785 ▶2	897.4366	1792.8586	1792.8566	1.11	0	88	6e-09	▶1	U	K.VCNYVNWVQQVIAAN.-
#15193 ▶5	912.4043	1822.7940	1822.7978	-2.10	0	57	0.00066	▶1	U	R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term); 2 [-1.0078 at C2,C9]
#15293 ▶3	916.4605	1830.9065	1830.9111	-2.56	0	93	2.6e-07	▶1	U	H.NIDVLEGNQFVNAAK.I + Carbamidomethyl (N-term)
#15880 ▶9	941.9930	1881.8313	1881.8349	-1.90	0	114	8.6e-19	▶1	H	R.SCAAAGTECLISGWGNK.S
#15890	928.2845	1881.8317	1881.8349	-1.72	0	48	3.5e-05	▶1	U	R.SCAAAGTECLISGWGNK.S
#15914	928.6180	1882.8323	1882.8189	7.09	0	44	0.014	▶1	U	R.SCAAAGTECLISGWGNK.S + [+0.9840 at N16]
#16103 ▶3	948.9313	1895.8480	1895.8506	-1.38	0	112	2.7e-09	▶1	U	R.SCAAAGTECLISGWGNK.S + [+14.0156 at C-term K]
#16220	955.9276	1909.8407	1909.8298	5.70	0	79	4.4e-06	▶1	U	R.SCAAAGTECLISGWGNK.S + [+27.9949 at T17]
#16229	637.6266	1909.8581	1909.8662	-4.27	0	54	0.0018	▶1	U	R.SCAAAGTECLISGWGNK.S + [+28.0313 at N16]
#16242 ▶4	637.6288	1909.8645	1909.8662	-0.90	0	60	0.00041	▶1	U	R.SCAAAGTECLISGWGNK.S + [+28.0313 at C-term K]
#16244 ▶9	955.9400	1909.8654	1909.8662	-0.45	0	117	8.5e-10	▶1	U	R.SCAAAGTECLISGWGNK.S + [+28.0313 at C-term K]
#16287	956.4820	1910.9494	1910.9486	0.43	0	47	0.01	▶1	U	E.HNIDVLEGNQFVNAAK.I
#16330 ▶1	957.9163	1913.8181	1913.8070	5.79	0	77	5.8e-06	▶1	U	R.SCAAAGTECLISGWGNK.S + [+31.9721 at W14]
#16432 ▶1	962.9273	1923.8400	1923.8567	-8.68	0	85	1.2e-06	▶1	U	R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term); [+42.0218 at C2] ←
#16434	642.2883	1923.8430	1923.8567	-7.13	0	50	0.0035	▶1	U	R.SCAAAGTECLISGWGNK.S + [+42.0218 at C-term K]
#16465 ▶1	963.9348	1925.8551	1925.8611	-3.12	0	65	0.00011	▶1	U	R.SCAAAGTECLISGWGNK.S + [+44.0262 at G15]
#16485	964.9608	1927.9069	1927.9098	-1.48	0	45	0.015	▶1	U	K.IITHFNFNGNTLNDIM.L
#16568	969.9386	1937.8627	1937.8611	0.80	0	47	0.0085	▶1	U	R.SCAAAGTECLISGWGNK.S + [+56.0262 at S12]
#16581 ▶4	970.4324	1938.8502	1938.8564	-3.21	0	85	4.7e-09	▶1	U	R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term) ←
#16582	647.2907	1938.8503	1938.8564	-3.15	0	28	0.002	▶1	U	R.SCAAAGTECLISGWGNK.S + Carbamidomethyl (N-term)

MASCOT : Error tolerant search statistics © 2021 Matrix Science



Usually, the search space for the second pass search is larger than for the first pass. This means that the significance threshold is more stringent for second pass matches. Here, for example, query 16581 gets a score of 85 in the first pass search which corresponds to an expect value of 4.7E-9. Query 16432 gets the same score in the second pass search, but the expect value is 1.2E-6, worse by a factor of 250

Error Tolerant Search: Tips

- **In most cases, best to exclude**
 - Isotopic labels
 - Modifications larger than 1 kDa
- **Very abundant modifications should be specified as variable**
 - Otherwise, miss doubly modified peptides

MASCOT : *Error tolerant search statistics* © 2021 Matrix Science



Finally, some tips.

The Unimod database contains a large number of entries that you do not expect to find in a general search. You can reduce the search time by excluding isotopic labels and very large modifications, which rarely give strong matches. 1 kDa is a good cut-off. Typically, this reduces the number of modifications by a third.

Remember that the second pass search works through the list of modifications serially. It doesn't look for combinations of modifications on a single peptide. So, if you have a very abundant modification, affecting 10% or more of the peptides, it is a good idea to specify this as a variable modification, so that you can find matches to peptides with both the abundant modification and an additional unsuspected one.

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

	<u>Target</u>	<u>Decoy</u>	<u>FDR</u>
Protein family members	59	0	0.00%
PSMs <input type="button" value="v"/> above <input type="button" value="homology v"/>	4279	42	0.98%

▼ **Modification statistics for all protein families**

Modification	Delta	Type	Site	Total matches
Carbamidomethyl	57.021464	variable	N-term	644
Oxidation	15.994915	variable	M	554
Non-specific cleavage		ET	-	543
Carbamidomethyl	57.021464	fixed	C	400
Deamidated	0.984016	ET	N	141
Ethyl	28.031299	ET	N-term	63
Ethyl	28.0313	ET	K	43
Methyl	14.01565	ET	E	30
Guanidinyl	42.021792	ET	N-term	27
Methyl	14.01565	ET	K	24
Dehydro	-1.007825	ET	C	22
Deamidated	0.984016	ET	Q	22
Dehydrated	-18.010565	ET	T	21
Gln->pyro-Glu	-17.026532	ET	N-term	21
Acetyl	42.010562	ET	N-term	17

MASCOT : Error tolerant search statistics © 2021 Matrix Science 

The example used for these screen shots was a search of LTQ-Orbitrap Velos data acquired by the Medical University of Graz and deposited in PRIDE project PXD002726. There is quite a bit of non-specific cleavage. N-term Carbamidomethyl was very common, so this was specified as a variable modification. The next most common modification is deamidation of asparagine, but it only affects 3% of the peptides, so it is not worth specifying as a variable modification.

Error Tolerant Search: Tips

- **In most cases, best to exclude**
 - Isotopic labels
 - Modifications larger than 1 kDa
- **Very abundant modifications should be specified as variable**
 - Otherwise, miss doubly modified peptides
- **Adjust *Min. number of sig. unique sequences* for good protein FDR**

Even when the FDR for PSMs is well controlled, the FDR for proteins will often be high for an error tolerant search because only a few entries are searched in the second pass.

Format
Significance threshold p< 0.02075
Max. number of families AUTO
[\[help\]](#)

Target FDR (overrides sig. threshold) 1%

Display non-sig. matches

Error tolerant matches: Reliable

Preferred taxonomy All entries

FDR type PSM

Min. number of sig. unique sequences 1

Dendrograms cut at 0

▼Sensitivity and FDR (reversed protein sequences)

	Target	Decoy	FDR
Protein family members	88	20	22.73%
PSMs	4279	42	0.98%

Format
Significance threshold p< 0.02075
Max. number of families AUTO
[\[help\]](#)

Target FDR (overrides sig. threshold) 1%

Display non-sig. matches

Error tolerant matches: Reliable

Preferred taxonomy All entries

FDR type PSM

Min. number of sig. unique sequences 2

Dendrograms cut at 0

▼Sensitivity and FDR (reversed protein sequences)

	Target	Decoy	FDR
Protein family members	59	0	0.00%
PSMs	4279	42	0.98%

MASCOT : Error tolerant search statistics © 2021 Matrix Science

In our example search, at 1% FDR for PSMs, the protein FDR is 23%, which sounds awful. This is simply because the 42 significant decoy matches are scattered randomly across 20 decoy proteins. If we increase the ‘Min. number of sig. unique sequences’ from 1 to 2 and choose ‘Format’, we eliminate one hit wonders, and the protein FDR drops to a more satisfactory 0%.