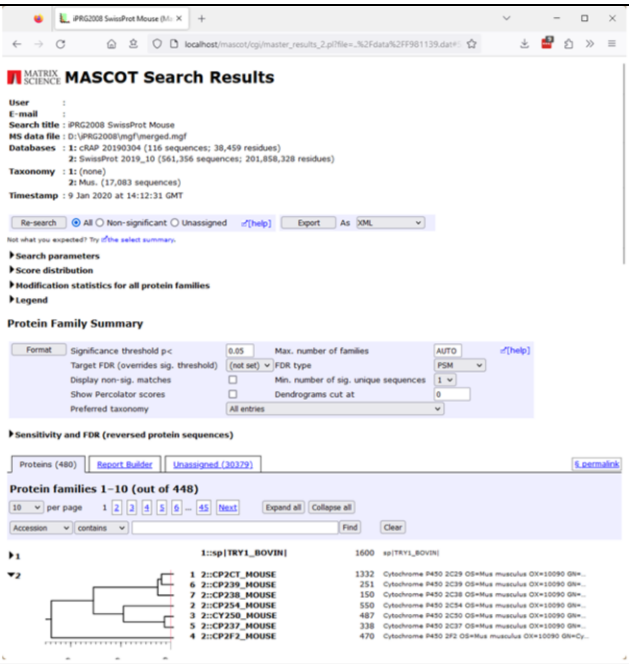
Scoring & Statistics



iPRG2008 SwissProt Mouse (M

localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat#

MATRIX SCIENCE **MASCOT Search Results**

```
User         :
E-mail       :
Search title : iPRG2008 SwissProt Mouse
MS data file : D:\iPRG2008\mgf\merged.mgf
Databases    : 1: cRAP 20190304 (116 sequences; 38,459 residues)
               2: SwissProt 2019_10 (561,356 sequences; 201,858,328 residues)
Taxonomy     : 1: (none)
               2: Mus. (17,083 sequences)
Timestamp    : 9 Jan 2020 at 14:12:31 GMT
```

Re-search ● All ○ Non-significant ○ Unassigned [help] Export As XML

Not what you expected? Try the select summary.

- Search parameters
- Score distribution
- Modification statistics for all protein families
- Legend

Protein Family Summary

Format

Significance threshold p< 0.05 | Max. number of families AUTO | [help]

Target FDR (overrides sig. threshold) (not set) | FDR type PSM

Display non-sig. matches □ | Min. number of sig. unique sequences 1

Show Percolator scores □ | Dendrograms cut at 0

Preferred taxonomy All entries

- Sensitivity and FDR (reversed protein sequences)

Proteins (480) | Report Builder | Unassigned (30379) | permalink

Protein families 1–10 (out of 448)

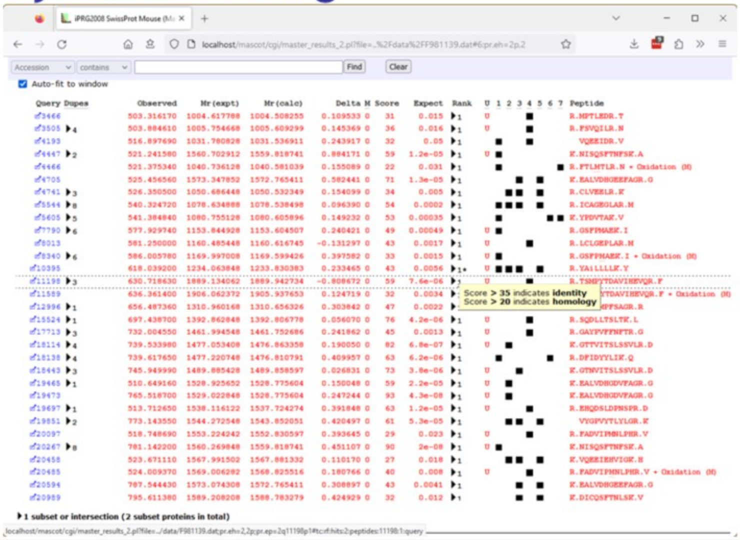
10 per page 1 2 3 4 5 6 ... 45 Next Expand all Collapse all

Accession contains Find Clear

```
▸1        1::sp|TRY1_BOVIN|     1600  sp|TRY1_BOVIN|
▾2      1 2::CP2CT_MOUSE        1332  Cytochrome P450 2C29 OS=Mus musculus OX=10090 GN=...
        6 2::CP239_MOUSE         251  Cytochrome P450 2C39 OS=Mus musculus OX=10090 GN=...
        7 2::CP238_MOUSE         150  Cytochrome P450 2C38 OS=Mus musculus OX=10090 GN=...
        2 2::CP254_MOUSE         550  Cytochrome P450 2C54 OS=Mus musculus OX=10090 GN=...
        3 2::CY250_MOUSE         487  Cytochrome P450 2C50 OS=Mus musculus OX=10090 GN=...
        5 2::CP237_MOUSE         338  Cytochrome P450 2C37 OS=Mus musculus OX=10090 GN=...
        4 2::CP2F2_MOUSE         470  Cytochrome P450 2F2 OS=Mus musculus OX=10090 GN=Cy...
```

This is the Mascot result report for a MS/MS ion search. A list of proteins is reported with a score calculated from individual peptide matches. Proteins with lots of strong peptide matches come at the top of the report. However, it is very important to understand that the protein score in an MS/MS search is not statistically rigorous. It is just a way of ranking the protein hits.

Probability based scoring


iPRG2008 SwissProt Mouse (M
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat#6:pr.eh=2p.2
Accession
contains
Find
Clear

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	3	4	5	6	7	Peptide
3466		503.316170	1004.617788	1004.508255	0.109533	0	31	0.015	▶1	U				■				R.MPTLEDR.T
3505	▶4	503.884610	1005.754668	1005.609299	0.145369	0	36	0.016	▶1	U				■				R.FSVQILR.N
4193		516.897690	1031.780828	1031.536911	0.243917	0	32	0.05	▶1		■			■				VQEEIDR.V
4447	▶2	521.241580	1560.702912	1559.818741	0.884171	0	59	1.2e-05	▶1	U	■							K.NISQSFTNFSK.A
4466		521.375340	1040.736128	1040.581039	0.155089	0	22	0.031	▶1		■						■	R.FTLMTLR.N + Oxidation (M)
4705		525.456560	1573.347852	1572.765411	0.582441	0	71	1.3e-05	▶1				■		■			K.EALVDHGEEFAGR.G
4741	▶3	526.350500	1050.686448	1050.532349	0.154099	0	34	0.005	▶1			■	■		■			R.CLVEELR.K
5544	▶8	540.324720	1078.634888	1078.538498	0.096390	0	54	0.0002	▶1		■	■	■		■			R.ICAGEGLAR.M
5605	▶5	541.384840	1080.755128	1080.605896	0.149232	0	53	0.00035	▶1		■					■	■	K.YPDVTAK.V
7790	▶6	577.929740	1153.844928	1153.604507	0.240421	0	49	0.00049	▶1	U	■							R.GSFPMAEK.I
8013		581.250000	1160.485448	1160.616745	-0.131297	0	43	0.0017	▶1	U				■				R.LCLGEPLAR.M
8340	▶6	586.005780	1169.997008	1169.599426	0.397582	0	33	0.0015	▶1	U	■							R.GSFPMAEK.I + Oxidation (M)
10395		618.039200	1234.063848	1233.830383	0.233465	0	43	0.0056	▶1•	U	■	■	■		■			R.YAiLLLLK.Y
11198	▶3	630.718630	1889.134062	1889.942734	-0.808672	0	59	7.6e-06	▶1	U				■				R.TSMPYTDAVIHEVQR.F
11589		636.361400	1906.062372	1905.937653	0.124719	0	32	0.0034	▶									YTDAVIHEVQR.F + Oxidation (M)
12996	▶1	656.487360	1310.960168	1310.656326	0.303842	0	47	0.0022	▶									PFSAGR.R
15524	▶1	697.438700	1392.862848	1392.806778	0.056070	0	76	4.2e-06	▶1	U				■				R.SQDLLTSLTK.L
17713	▶3	732.004550	1461.994548	1461.752686	0.241862	0	45	0.0013	▶1	U				■				R.GAYFVFFNFTR.G
18114	▶4	739.533980	1477.053408	1476.863358	0.190050	0	82	6.8e-07	▶1	U		■						K.GTTVITSLSSVLR.D
18138	▶4	739.617650	1477.220748	1476.810791	0.409957	0	63	6.2e-06	▶1		■					■		R.DFIDYYLIK.Q
18443	▶3	745.949990	1489.885428	1489.858597	0.026831	0	73	3.8e-06	▶1	U			■					K.GTNVITSLSSVLR.D
19465	▶1	510.649160	1528.925652	1528.775604	0.150048	0	59	2.2e-05	▶1	U		■						K.EALVDHGDVFAGR.G
19473		765.518700	1529.022848	1528.775604	0.247244	0	93	4.3e-08	▶1	U		■						K.EALVDHGDVFAGR.G
19697	▶1	513.712650	1538.116122	1537.724274	0.391848	0	63	1.2e-05	▶1	U				■				R.EHQDSLDPNSPR.D
19851	▶2	773.143550	1544.272548	1543.852051	0.420497	0	61	5.3e-05	▶1			■	■		■			VYGPVYTLYLGR.K
20097		518.748690	1553.224242	1552.830597	0.393645	0	29	0.023	▶1	U				■				R.FADVIPMNLPHR.V
20267	▶8	781.142200	1560.269848	1559.818741	0.451107	0	90	2e-08	▶1	U	■							K.NISQSFTNFSK.A
20458		523.671110	1567.991502	1567.881332	0.110170	0	27	0.018	▶1			■	■		■			K.VQEEIEHVIGK.H
20485		524.009370	1569.006282	1568.825516	0.180766	0	40	0.008	▶1	U				■				R.FADVIPMNLPHR.V + Oxidation (M)
20594		787.544430	1573.074308	1572.765411	0.308897	0	43	0.0041	▶1				■		■			K.EALVDHGEEFAGR.G
20989		795.611380	1589.208208	1588.783279	0.424929	0	32	0.012	▶1				■		■			K.DICQSFTNLSK.V

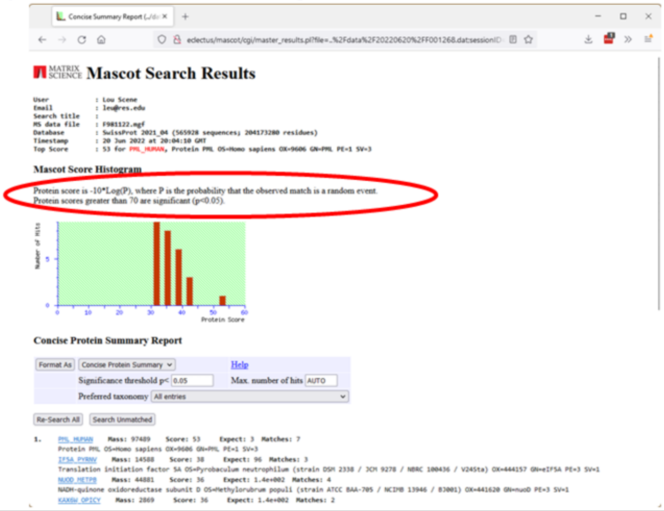
Score > 35 indicates **identity**
Score > 20 indicates **homology**

▶1 subset or intersection (2 subset proteins in total)

localhost/mascot/cgi/master_results_2.pl?file=../data/F981139.dat;pr.eh=2,2p;pr.ep=2q11198p1#tc:rf:hits:2:peptides:11198:1:query

If we expand one of the protein matches, we can see a list of peptides assigned to the protein. Each peptide has a score and expect value, these are probabilistic scores. If you mouse over the rank column you can see the score cutoffs for each query. The identity and homology threshold will be described in more detail in this talk (slides 17 and 18).

Probability based scoring



Mascot Search Results

```
User           : Lou Scene
Email          : leu@res.edu
Search title   :
MS data file   : F981122.mgf
Database       : SwissProt 2021_04 (565928 sequences; 204173280 residues)
Timestamp      : 20 Jun 2022 at 20:04:10 GMT
Top Score      : 53 for PML_HUMAN, Protein PML OS=Homo sapiens OX=9606 GN=PML PE=1 SV=3
```

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 70 are significant (p<0.05).

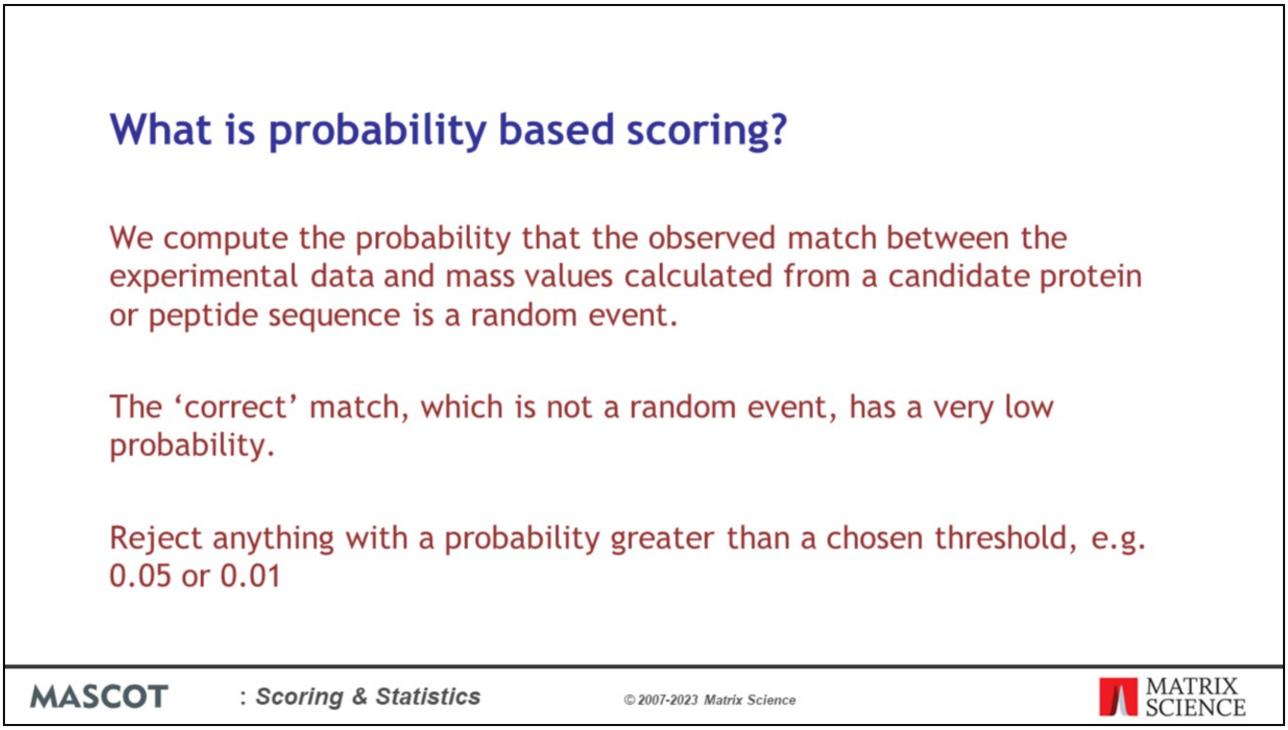
Concise Protein Summary Report

Significance threshold p< 0.05   Max. number of hits AUTO

Preferred taxonomy All entries

```
1.   PML_HUMAN     Mass: 97489    Score: 53    Expect: 3  Matches: 7
     Protein PML OS=Homo sapiens OX=9606 GN=PML PE=1 SV=3
     IFSA_PYRNV    Mass: 14588    Score: 38    Expect: 96  Matches: 3
     Translation initiation factor 5A OS=Pyrobaculum neutrophilum (strain DSM 2338 / JCM 9278 / NBRC 100436 / V24Sta) OX=444157 GN=eIF5A PE=3 SV=1
     NUOD_METPB    Mass: 44881    Score: 36    Expect: 1.4e+002  Matches: 4
     NADH-quinone oxidoreductase subunit D OS=Methylorubrum populi (strain ATCC BAA-705 / NCIMB 13946 / BJ001) OX=441620 GN=nuoD PE=3 SV=1
     KAX6W_OPICY   Mass: 2869     Score: 36    Expect: 1.4e+002  Matches: 2
```

This is the Mascot result report for a peptide mass fingerprint search. There is also a list of proteins each of which matches some of the experimental peptide masses. In peptide fingerprint, we use probability-based scoring, and the report tells us that these matches are not statistically significant. The score threshold for this search is 70, and the top scoring match is 53. The graph is a histogram of the scores of the top ten matches and, as you see, all of them are in the area shaded green to indicate random, meaningless matches.



What is probability based scoring?

We compute the probability that the observed match between the experimental data and mass values calculated from a candidate protein or peptide sequence is a random event.

The ‘correct’ match, which is not a random event, has a very low probability.

Reject anything with a probability greater than a chosen threshold, e.g. 0.05 or 0.01

MASCOT : Scoring & Statistics

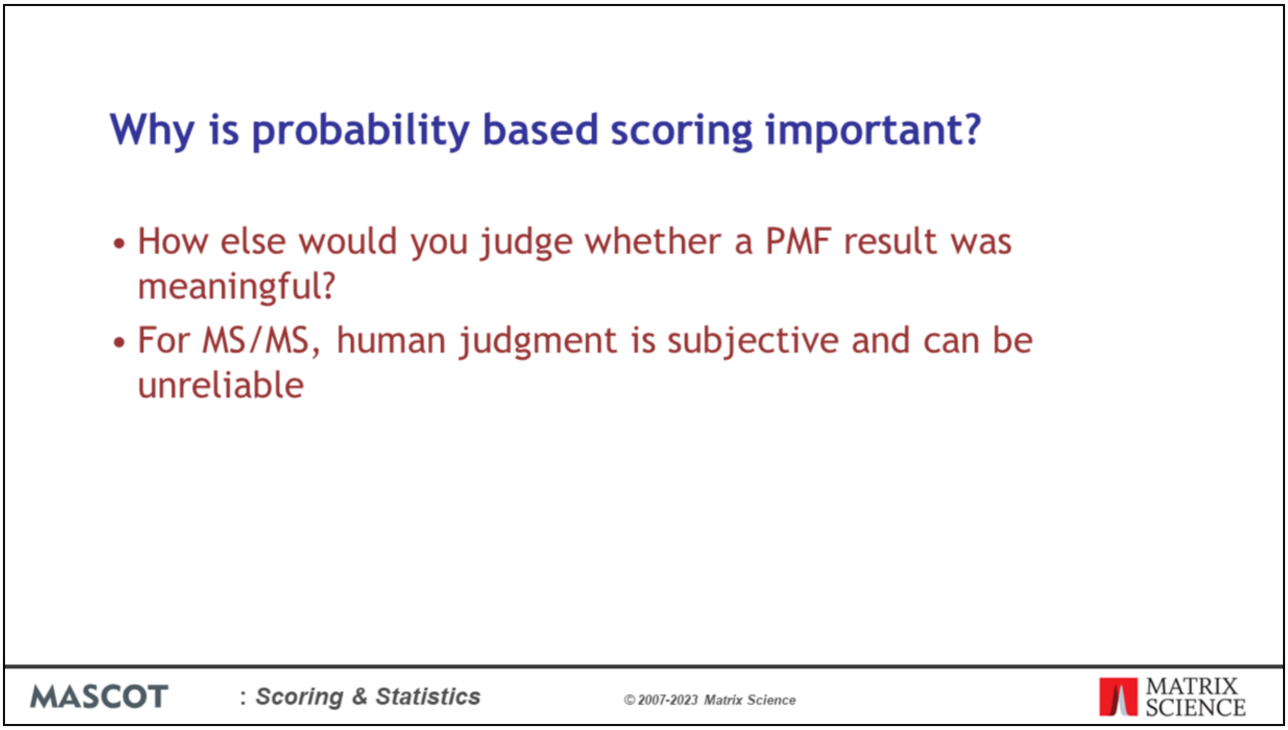
MATRIX SCIENCE

What exactly do I mean by probability based scoring?

We calculate, as accurately as possible, the probability that the observed match between the experimental data, and mass values calculated from a candidate peptide or protein sequence, is a random event.

The real match, which is not a random event, then has a very low probability.

We can then reject anything with a probability greater than a chosen threshold, e.g. 1%



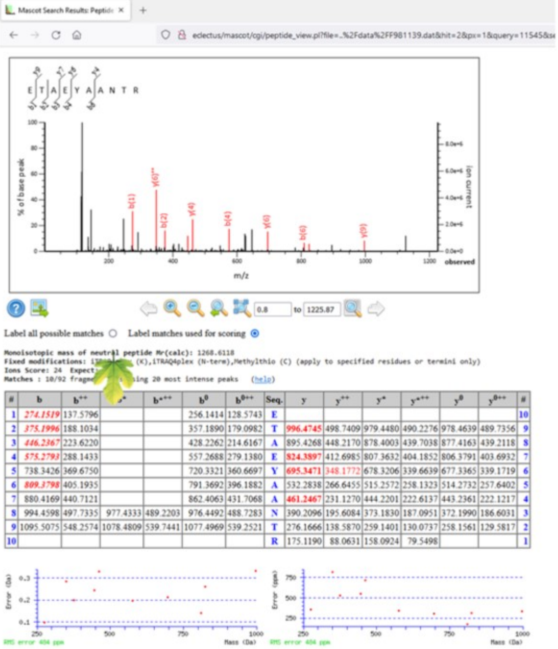
Why is probability based scoring important?

- How else would you judge whether a PMF result was meaningful?
- For MS/MS, human judgment is subjective and can be unreliable

Why is probability based scoring important?

Well, how else would you judge whether a protein hit in a peptide mass fingerprint search was meaningful?

In the case of MS/MS data, it is very difficult to judge whether a match is significant or not by looking at the spectrum. Let me illustrate this with an example.



Label all possible matches ○ Label matches used for scoring ●

```
Monoisotopic mass of neutral peptide Mr(calc): 1268.6118
Fixed modifications: iT        (K),iTRAQ4plex (N-term),Methylthio (C) (apply to specified residues or termini only)
Ions Score: 24  Expect:
Matches : 10/92 fragme       sing 20 most intense peaks   (help)
```

#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	274.1519	137.5796			256.1414	128.5743	E							10
2	375.1996	188.1034			357.1890	179.0982	T	996.4745	498.7409	979.4480	490.2276	978.4639	489.7356	9
3	446.2367	223.6220			428.2262	214.6167	A	895.4268	448.2170	878.4003	439.7038	877.4163	439.2118	8
4	575.2793	288.1433			557.2688	279.1380	E	824.3897	412.6985	807.3632	404.1852	806.3791	403.6932	7
5	738.3426	369.6750			720.3321	360.6697	Y	695.3471	348.1772	678.3206	339.6639	677.3365	339.1719	6
6	809.3798	405.1935			791.3692	396.1882	A	532.2838	266.6455	515.2572	258.1323	514.2732	257.6402	5
7	880.4169	440.7121			862.4063	431.7068	A	461.2467	231.1270	444.2201	222.6137	443.2361	222.1217	4
8	994.4598	497.7335	977.4333	489.2203	976.4492	488.7283	N	390.2096	195.6084	373.1830	187.0951	372.1990	186.6031	3
9	1095.5075	548.2574	1078.4809	539.7441	1077.4969	539.2521	T	276.1666	138.5870	259.1401	130.0737	258.1561	129.5817	2
10							R	175.1190	88.0631	158.0924	79.5498			1

MASCOT : *Scoring & Statistics* © 2007-2023 *Matrix Science* MATRIX SCIENCE

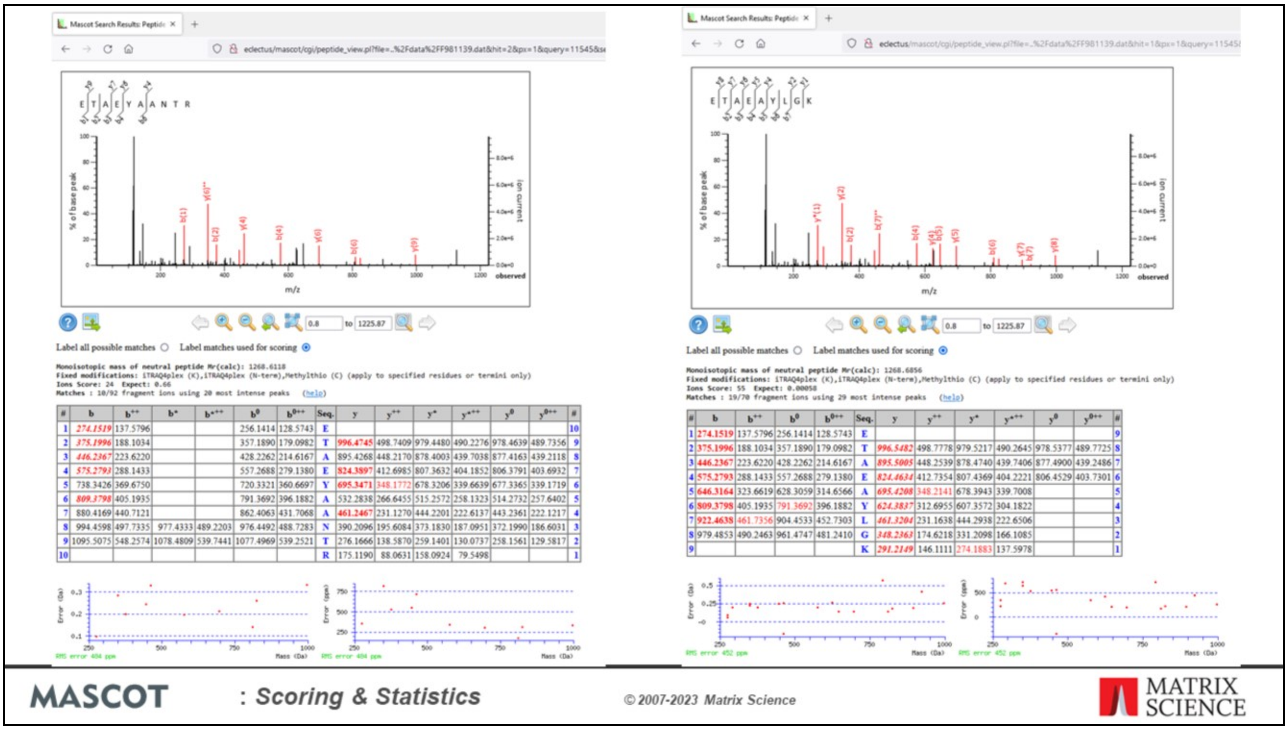
This match has a good number of matches to y and b ions, highlighted in red. Almost all the major peaks above 200 Da seem to be labelled. Could such a good match have occurred by chance?

You cannot tell, because you can match anything to anything if you try hard enough.

If this sounds strange, here's a simple analogy. If I say that I was tossing a coin and got ten heads in a row, does that mean there was something strange about the coin, like it had two heads? You cannot tell, because you need to know how many times I tossed the coin in total. If I picked it up off the table, tossed it ten times, then put it down, yes, that would suggest this was not a fair coin. However, if I tossed it ten thousand times, I would expect to get ten heads in a row more than once.

So, it isn't just a matter of how good the match is, i.e. how many y or b ions you found, it's a case of how hard you tried to find the match. In the case of a database search, this means how large is the database, what is the mass tolerance, how many variable modifications, etc., etc. These are very difficult calculations to do in your head, but they are easy calculations for the search engine.

If we look at the expectation value for this match, it is 0.66. That is, we could expect to get this match purely by chance. It looks good, but it's a random match.


Mascot Search Results: Peptide
eclectus/mascot/cgi/peptide_view.pl?file=../%2Fdata%2FF981139.dat&hit=2&px=1&query=11545&se
E T A E Y A A N T R
% of base peak
ion current
m/z
observed
b(1)
y(6)++
b(2)
y(4)
b(4)
y(6)
b(6)
y(9)
0.8
to
1225.87
Error (Da)
Error (ppm)
Mass (Da)
RMS error 484 ppm
Mascot Search Results: Peptide
eclectus/mascot/cgi/peptide_view.pl?file=../%2Fdata%2FF981139.dat&hit=1&px=1&query=11545
E T A E A Y L G K
y*(1)
y(2)
b(2)
b(7)++
b(4)
y(4)
b(5)
y(5)
b(6)
y(7)
b(7)
y(8)
RMS error 452 ppm

Label all possible matches ○ Label matches used for scoring ●

```
Monoisotopic mass of neutral peptide Mr(calc): 1268.6118
Fixed modifications: iTRAQ4plex (K),iTRAQ4plex (N-term),Methylthio (C) (apply to specified residues or termini only)
Ions Score: 24  Expect: 0.66
Matches : 10/92 fragment ions using 20 most intense peaks   (help)
```

#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	*274.1519*	137.5796			256.1414	128.5743	E							10
2	*375.1996*	188.1034			357.1890	179.0982	T	996.4745	498.7409	979.4480	490.2276	978.4639	489.7356	9
3	*446.2367*	223.6220			428.2262	214.6167	A	895.4268	448.2170	878.4003	439.7038	877.4163	439.2118	8
4	*575.2793*	288.1433			557.2688	279.1380	E	824.3897	412.6985	807.3632	404.1852	806.3791	403.6932	7
5	738.3426	369.6750			720.3321	360.6697	Y	695.3471	348.1772	678.3206	339.6639	677.3365	339.1719	6
6	*809.3798*	405.1935			791.3692	396.1882	A	532.2838	266.6455	515.2572	258.1323	514.2732	257.6402	5
7	880.4169	440.7121			862.4063	431.7068	A	461.2467	231.1270	444.2201	222.6137	443.2361	222.1217	4
8	994.4598	497.7335	977.4333	489.2203	976.4492	488.7283	N	390.2096	195.6084	373.1830	187.0951	372.1990	186.6031	3
9	1095.5075	548.2574	1078.4809	539.7441	1077.4969	539.2521	T	276.1666	138.5870	259.1401	130.0737	258.1561	129.5817	2
10							R	175.1190	88.0631	158.0924	79.5498			1

Label all possible matches ○ Label matches used for scoring ●

```
Monoisotopic mass of neutral peptide Mr(calc): 1268.6856
Fixed modifications: iTRAQ4plex (K),iTRAQ4plex (N-term),Methylthio (C) (apply to specified residues or termini only)
Ions Score: 55  Expect: 0.00058
Matches : 19/70 fragment ions using 29 most intense peaks   (help)
```

#	b	b++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	274.1519	137.5796	256.1414	128.5743	E							9
2	375.1996	188.1034	357.1890	179.0982	T	*996.5482*	498.7778	979.5217	490.2645	978.5377	489.7725	8
3	446.2367	223.6220	428.2262	214.6167	A	*895.5005*	448.2539	878.4740	439.7406	877.4900	439.2486	7
4	575.2793	288.1433	557.2688	279.1380	E	*824.4634*	412.7354	807.4369	404.2221	806.4529	403.7301	6
5	646.3164	323.6619	628.3059	314.6566	A	*695.4208*	348.2141	678.3943	339.7008			5
6	809.3798	405.1935	791.3692	396.1882	Y	*624.3837*	312.6955	607.3572	304.1822			4
7	922.4638	461.7356	904.4533	452.7303	L	*461.3204*	231.1638	444.2938	222.6506			3
8	979.4853	490.2463	961.4747	481.2410	G	*348.2363*	174.6218	331.2098	166.1085			2
9					K	*291.2149*	146.1111	274.1883	137.5978			1

MASCOT : *Scoring & Statistics*

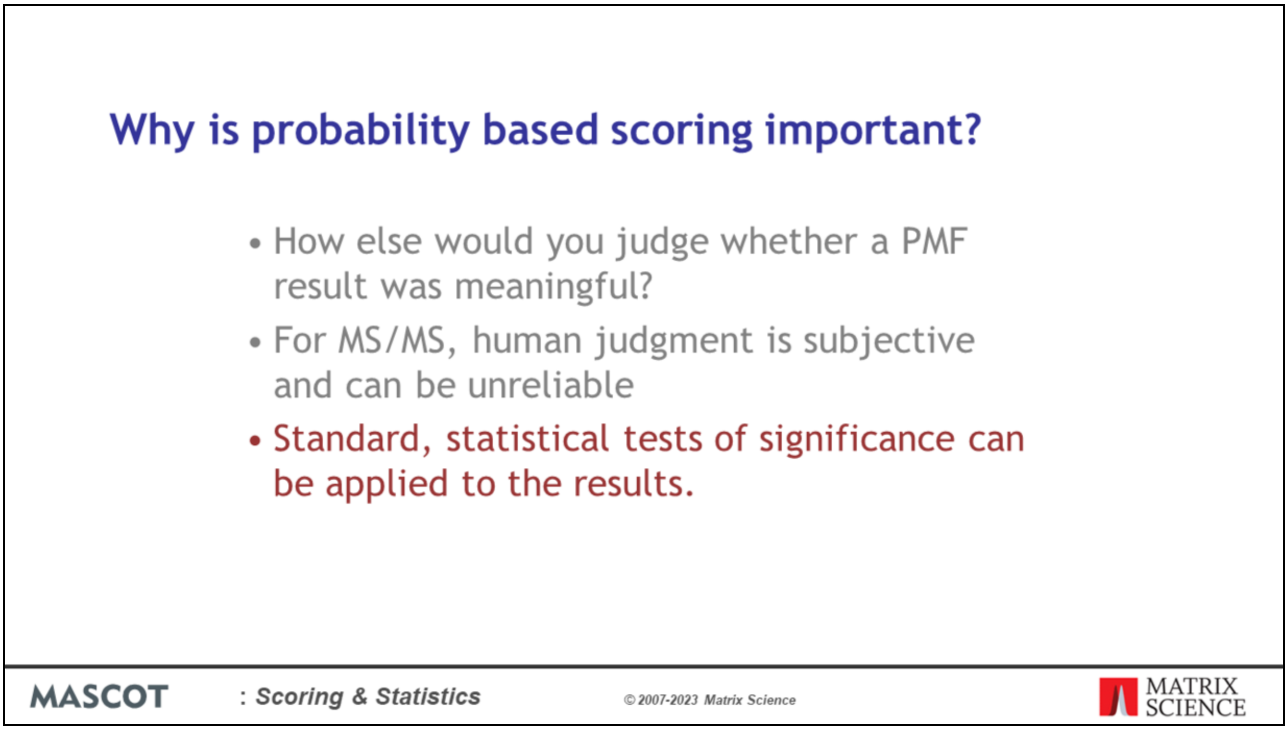
© 2007-2023 *Matrix Science*

MATRIX SCIENCE

If I show you a better match, then it is easy to dismiss the previous one as inferior. We can all make that judgement very easily. This match has an expectation value of less than 1 in 5,000. It is definitely not random.

The challenge is, what if you don't have the better match to compare against? Maybe this sequence wasn't in the database. If you only had the inferior match, how would you decide by looking at it whether it was significant or not?

The other interesting question is whether this is the "correct" match. Who can say that a better match isn't possible, where we get the extra y ion or the first and last b ions?

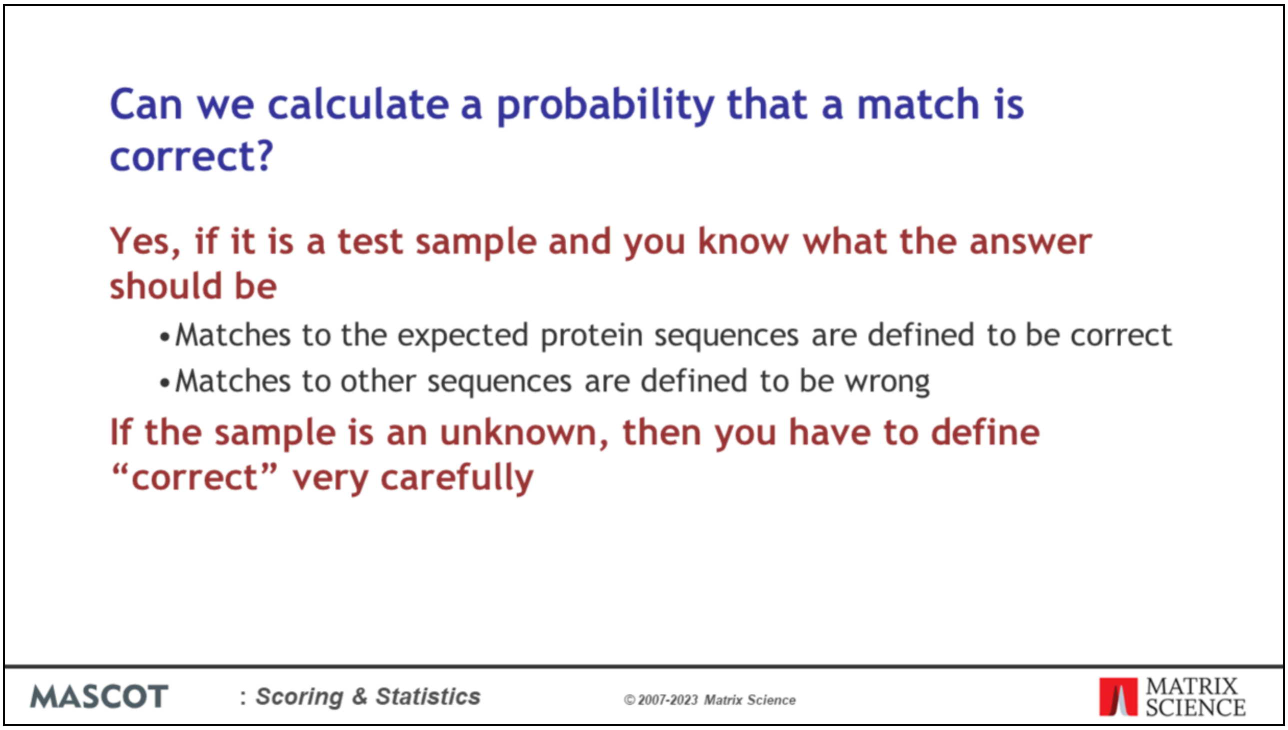


Why is probability based scoring important?

- How else would you judge whether a PMF result was meaningful?
- For MS/MS, human judgment is subjective and can be unreliable
- Standard, statistical tests of significance can be applied to the results.

If we use probability based scoring, we can apply standard, statistical tests of significance to the results.

If we don't do this, then the only way to know the level of false positives is a target decoy search, and this isn't always possible, e.g. when searching a small number of spectra or doing a targeted search of only a few protein sequences



Can we calculate a probability that a match is correct?

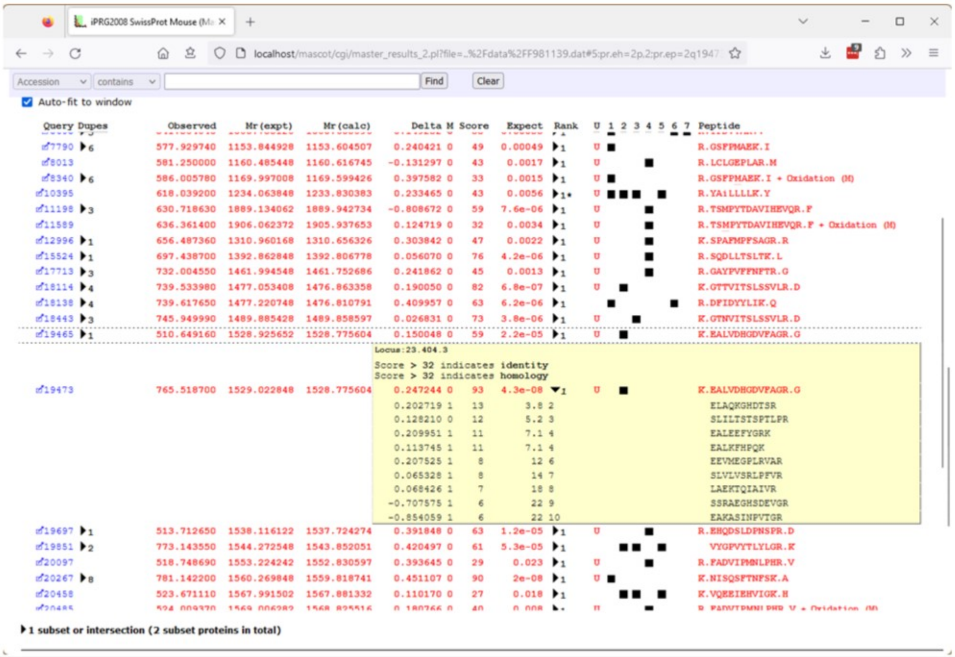
Yes, if it is a test sample and you know what the answer should be

- Matches to the expected protein sequences are defined to be correct
- Matches to other sequences are defined to be wrong

If the sample is an unknown, then you have to define “correct” very carefully

Probability based scoring calculates the probability that the match is random. This is, the probability that the match is meaningless. Many people ask whether we can report the probability that the match is correct. Is this possible?

It is certainly possible if you are analysing a known protein or standard mixture of proteins. If you know what the sequences are, or think you know, then the matches to the known sequences are defined to be correct and those to any other sequence are defined to be wrong. If the sample is an unknown, then it is difficult even to define what is meant by a correct match.



iPRG2008 SwissProt Mouse (M

localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981139.dat#5:pr.eh=2p,2:pr.ep=2q1947

Accession ⌄ contains ⌄ Find Clear

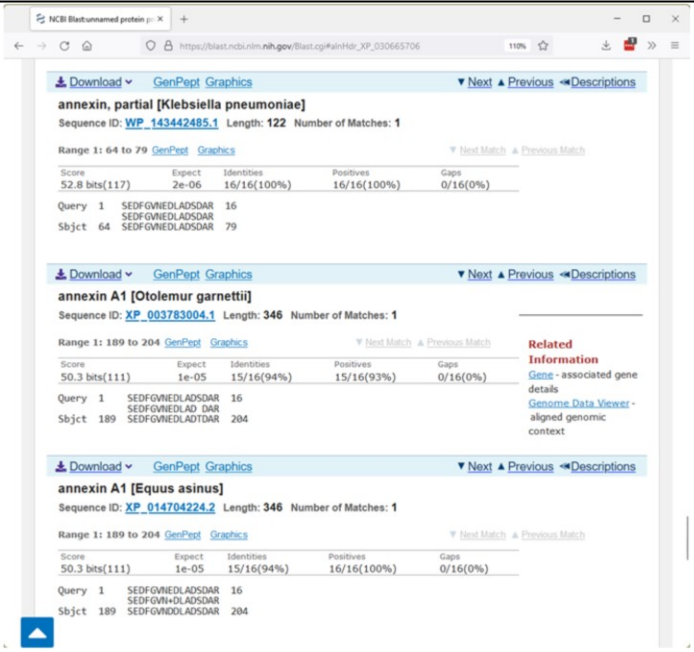
☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	3	4	5	6	7	Peptide
7790	▶6	577.929740	1153.844928	1153.604507	0.240421	0	49	0.00049	▶1	U	■							R.GSFPMAEK.I
8013		581.250000	1160.485448	1160.616745	-0.131297	0	43	0.0017	▶1	U				■				R.LCLGEPLAR.M
8340	▶6	586.005780	1169.997008	1169.599426	0.397582	0	33	0.0015	▶1	U	■							R.GSFPMAEK.I + Oxidation (M)
10395		618.039200	1234.063848	1233.830383	0.233465	0	43	0.0056	▶1•	U	■	■	■		■			R.YAiLLLLK.Y
11198	▶3	630.718630	1889.134062	1889.942734	-0.808672	0	59	7.6e-06	▶1	U				■				R.TSMPYTDAVIHEVQR.F
11589		636.361400	1906.062372	1905.937653	0.124719	0	32	0.0034	▶1	U				■				R.TSMPYTDAVIHEVQR.F + Oxidation (M)
12996	▶1	656.487360	1310.960168	1310.656326	0.303842	0	47	0.0022	▶1	U				■				K.SPAFMPFSAGR.R
15524	▶1	697.438700	1392.862848	1392.806778	0.056070	0	76	4.2e-06	▶1	U				■				R.SQDLLTSLTK.L
17713	▶3	732.004550	1461.994548	1461.752686	0.241862	0	45	0.0013	▶1	U				■				R.GAYPVFFNFTR.G
18114	▶4	739.533980	1477.053408	1476.863358	0.190050	0	82	6.8e-07	▶1	U		■						K.GTTVITSLSSVLR.D
18138	▶4	739.617650	1477.220748	1476.810791	0.409957	0	63	6.2e-06	▶1		■					■		R.DFIDYYLIK.Q
18443	▶3	745.949990	1489.885428	1489.858597	0.026831	0	73	3.8e-06	▶1	U			■					K.GTNVITSLSSVLR.D
19465	▶1	510.649160	1528.925652	1528.775604	0.150048	0	59	2.2e-05	▶1	U		■						K.EALVDHGDVFAGR.G
					Locus:23.404.3 Score > 32 indicates identity Score > 32 indicates homology													
19473		765.518700	1529.022848	1528.775604	0.247244	0	93	4.3e-08	▼1	U		■						K.EALVDHGDVFAGR.G
					0.202719	1	13	3.8	2									ELAQKGHDTSR
					0.128210	0	12	5.2	3									SLILTSTSPTLPR
					0.209951	1	11	7.1	4									EALEEFYGRK
					0.113745	1	11	7.1	4									EALKFHPQK
					0.207525	1	8	12	6									EEVMEGPLRVAR
					0.065328	1	8	14	7									SLVLVSRLPFVR
					0.068426	1	7	18	8									LAEKTQIAIVR
					-0.707575	1	6	22	9									SSRAEGHSDEVGR
					-0.854059	1	6	22	10									EAKASINPVTGR
19697	▶1	513.712650	1538.116122	1537.724274	0.391848	0	63	1.2e-05	▶1	U				■				R.EHQDSLDPNSPR.D
19851	▶2	773.143550	1544.272548	1543.852051	0.420497	0	61	5.3e-05	▶1			■	■		■			VYGPVYTLYLGR.K
20097		518.748690	1553.224242	1552.830597	0.393645	0	29	0.023	▶1	U				■				R.FADVIPMNLPHR.V
20267	▶8	781.142200	1560.269848	1559.818741	0.451107	0	90	2e-08	▶1	U	■							K.NISQSFTNFSK.A
20458		523.671110	1567.991502	1567.881332	0.110170	0	27	0.018	▶1			■	■		■			K.VQEEIEHVIGK.H
20485		524.009370	1569.006282	1568.825516	0.180766	0	40	0.008	▶1	U				■				R.FADVIPMNLPHR.V + Oxidation (M)

▶1 subset or intersection (2 subset proteins in total)

This is a typical MS/MS search result, where we see a series of high scoring homologous peptides. The sequences of the top three matches are very similar, and their expectation values vary from random through to very unlikely to be random. The best match has an expectation value of 1.2E-7. However, we cannot be sure that this is an identity match to the analyte peptide. It is simply the best match we could find in the database. There is always the possibility that a better match exists, that is not in the database, so to call it the correct match would be misleading.

The important thing is that we have a mechanism to discard matches that are nothing more than random matches.


NCBI Blast:unnamed protein p
https://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_XP_030665706
110%
Download
GenPept Graphics
▼ Next ▲ Previous ◀◀Descriptions
▼ Next Match ▲ Previous Match

annexin, partial [Klebsiella pneumoniae]

Sequence ID: WP_143442485.1 Length: **122** Number of Matches: **1**

Range 1: 64 to 79 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
52.8 bits(117)	2e-06	16/16(100%)	16/16(100%)	0/16(0%)

```
Query  1    SEDFGVNEDLADSDAR  16
            SEDFGVNEDLADSDAR
Sbjct  64   SEDFGVNEDLADSDAR  79
```

annexin A1 [Otolemur garnettii]

Sequence ID: XP_003783004.1 Length: **346** Number of Matches: **1**

Range 1: 189 to 204 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
50.3 bits(111)	1e-05	15/16(94%)	15/16(93%)	0/16(0%)

```
Query  1    SEDFGVNEDLADSDAR  16
            SEDFGVNEDLAD DAR
Sbjct  189  SEDFGVNEDLADTDAR  204
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

annexin A1 [Equus asinus]

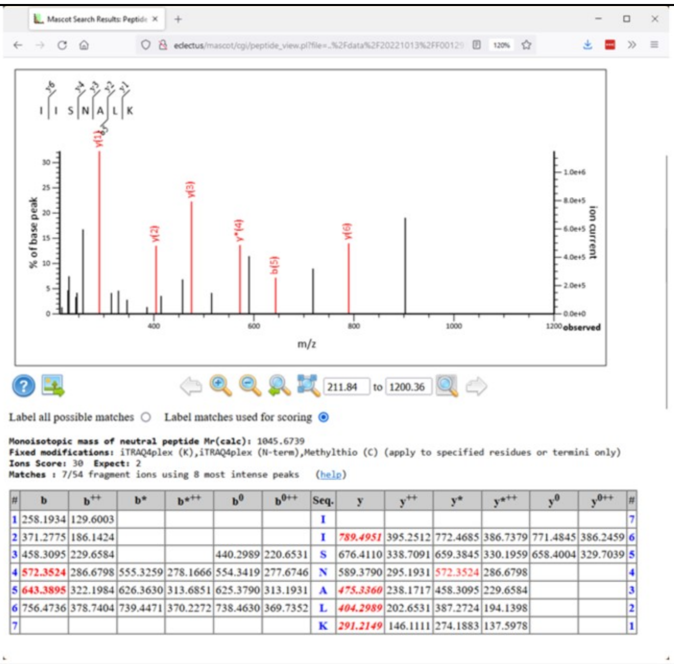
Sequence ID: XP_014704224.2 Length: **346** Number of Matches: **1**

Range 1: 189 to 204 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
50.3 bits(111)	1e-05	15/16(94%)	16/16(100%)	0/16(0%)

```
Query  1    SEDFGVNEDLADSDAR  16
            SEDFGVN+DLADSDAR
Sbjct  189  SEDFGVNDDLADSDAR  204
```

It is a similar situation in Blast, except that you have the luxury of seeing when you have a perfect identity match. Here, the identity match has an expectation value of 2E-6, which reminds us that it would be a random match if the database was 2 million times larger. The match with one different residue is not worthless, it has an expectation value of 1E-5 and is a very good match. It just isn't as good a match as the one above.



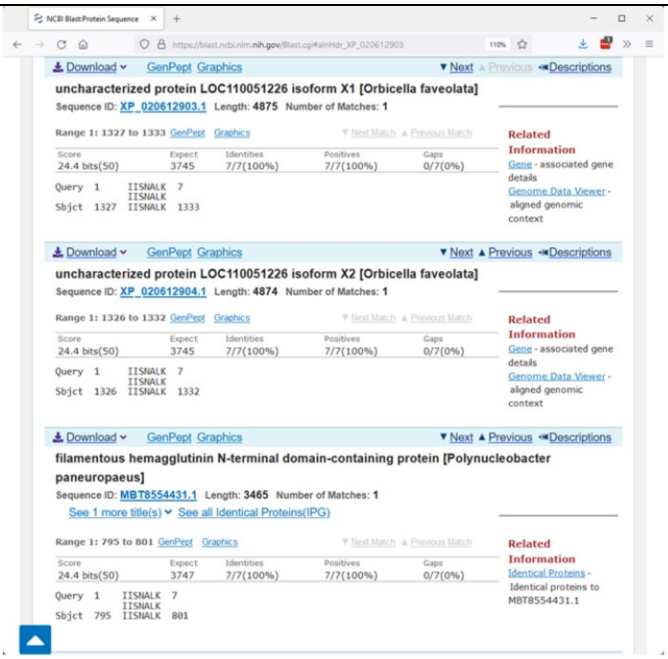
Label all possible matches ○ Label matches used for scoring ●

```
Monoisotopic mass of neutral peptide Mr(calc): 1045.6739
Fixed modifications: iTRAQ4plex (K),iTRAQ4plex (N-term),Methylthio (C) (apply to specified residues or termini only)
Ions Score: 30  Expect: 2
Matches : 7/54 fragment ions using 8 most intense peaks   (help)
```

#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	258.1934	129.6003					I							7
2	371.2775	186.1424					I	789.4951	395.2512	772.4685	386.7379	771.4845	386.2459	6
3	458.3095	229.6584			440.2989	220.6531	S	676.4110	338.7091	659.3845	330.1959	658.4004	329.7039	5
4	572.3524	286.6798	555.3259	278.1666	554.3419	277.6746	N	589.3790	295.1931	572.3524	286.6798			4
5	643.3895	322.1984	626.3630	313.6851	625.3790	313.1931	A	475.3360	238.1717	458.3095	229.6584			3
6	756.4736	378.7404	739.4471	370.2272	738.4630	369.7352	L	404.2989	202.6531	387.2724	194.1398			2
7							K	291.2149	146.1111	274.1883	137.5978			1

If we are doing probability based matching, we are not scoring the quality of the spectrum, we are scoring whether the match is random or not.

Even when the mass spectrum is of very high quality, if the peptide is so short that it could occur in the database by chance, then you will not get a very good score.


NCBI Blast:Protein Sequence
https://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_XP_020612903
110%

Download GenPept Graphics ▼ Next ▲ Previous ◂◂Descriptions

uncharacterized protein LOC110051226 isoform X1 [Orbicella faveolata]

Sequence ID: XP_020612903.1 Length: **4875** Number of Matches: **1**

Range 1: 1327 to 1333 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
24.4 bits(50)	3745	7/7(100%)	7/7(100%)	0/7(0%)

```
Query  1     IISNALK  7
             IISNALK
Sbjct  1327  IISNALK  1333
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

Download GenPept Graphics ▼ Next ▲ Previous ◂◂Descriptions

uncharacterized protein LOC110051226 isoform X2 [Orbicella faveolata]

Sequence ID: XP_020612904.1 Length: **4874** Number of Matches: **1**

Range 1: 1326 to 1332 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
24.4 bits(50)	3745	7/7(100%)	7/7(100%)	0/7(0%)

```
Query  1     IISNALK  7
             IISNALK
Sbjct  1326  IISNALK  1332
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

Download GenPept Graphics ▼ Next ▲ Previous ◂◂Descriptions

filamentous hemagglutinin N-terminal domain-containing protein [Polynucleobacter paneuropaeus]

Sequence ID: MBT8554431.1 Length: **3465** Number of Matches: **1**

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 795 to 801 GenPept Graphics ▼ Next Match ▲ Previous Match

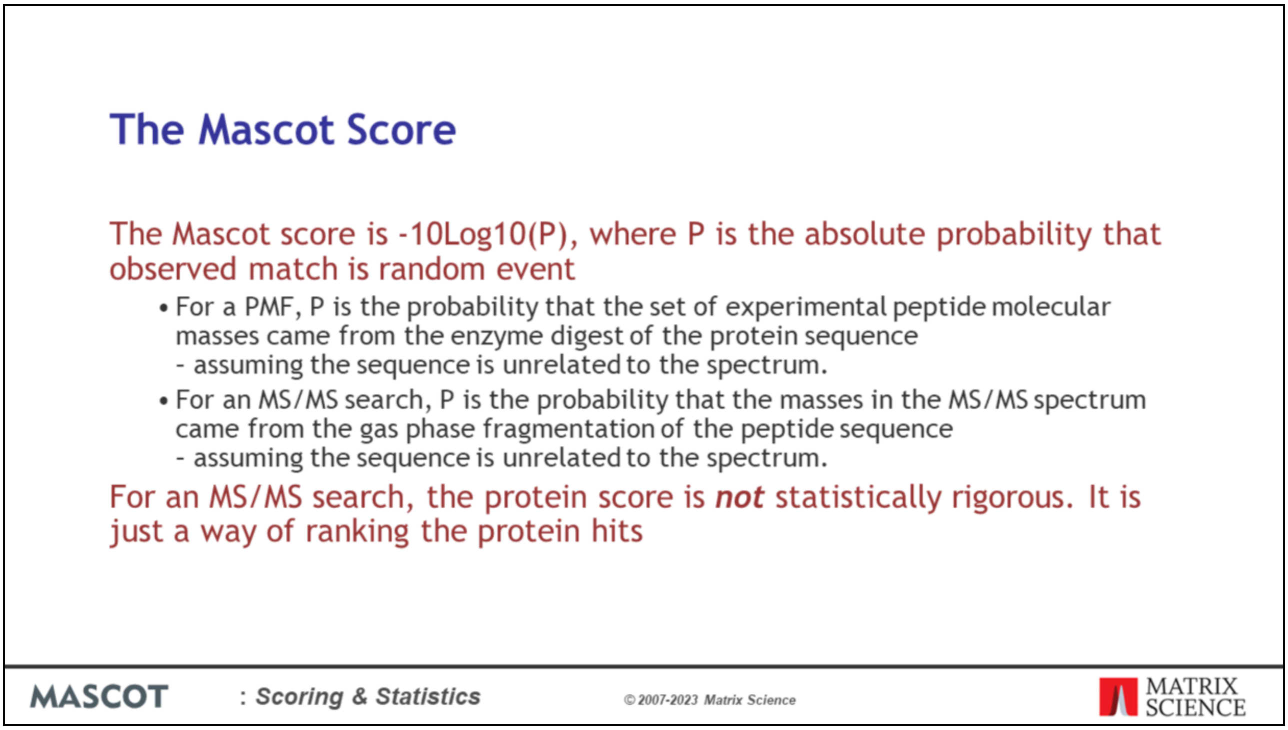
Score	Expect	Identities	Positives	Gaps
24.4 bits(50)	3747	7/7(100%)	7/7(100%)	0/7(0%)

```
Query  1    IISNALK  7
            IISNALK
Sbjct  795  IISNALK  801
```

Related Information

Identical Proteins - Identical proteins to MBT8554431.1

The situation in a Blast search is identical. Even though this is a perfect identity match, the expectation value is 3745. This is just a random match. Hence, the earlier tip to discard spectra from low mass precursors.



The Mascot Score

The Mascot score is -10Log10(P), where P is the absolute probability that observed match is random event

- For a PMF, P is the probability that the set of experimental peptide molecular masses came from the enzyme digest of the protein sequence
  - assuming the sequence is unrelated to the spectrum.
- For an MS/MS search, P is the probability that the masses in the MS/MS spectrum came from the gas phase fragmentation of the peptide sequence
  - assuming the sequence is unrelated to the spectrum.

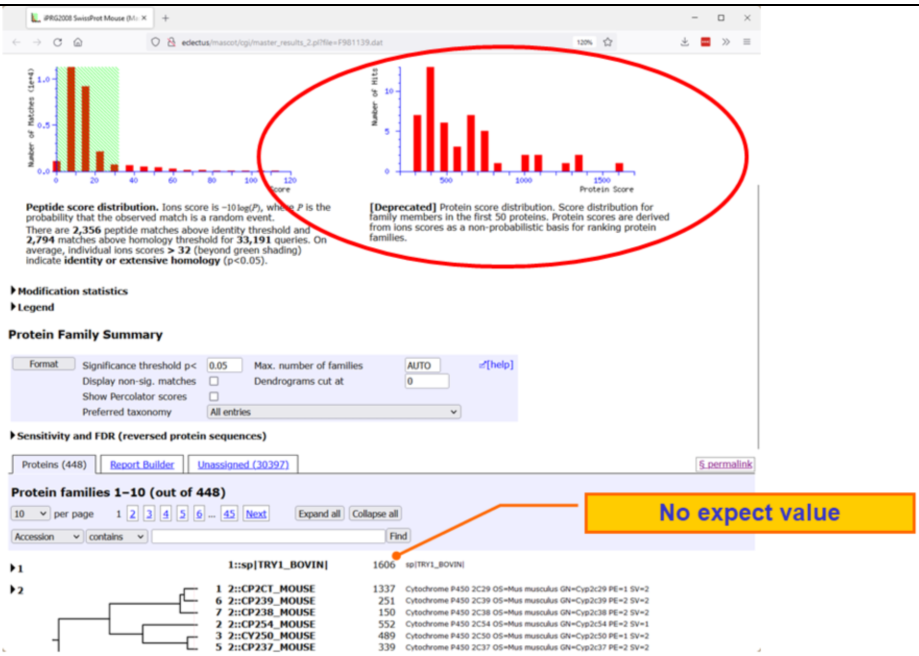
For an MS/MS search, the protein score is ***not*** statistically rigorous. It is just a way of ranking the protein hits

For a peptide mass fingerprint, there is just one score that matters: the protein score. This tells us whether the match is significant or not, and is determined by calculating the probability of getting the observed number of peptide mass matches if the protein sequence was random.

For an MS/MS search, we have two scores. The important one is the peptide match score or ions score. This is the probability of getting the observed number of fragment ion mass matches if the peptide sequence was an unrelated sequence.

However, most people are interested in which proteins are present, rather than which peptides have been found. So, we assign peptide matches to protein hits and provide protein scores for MS/MS searches, so that the proteins with lots of strong peptide matches come at the top of the report.

However, it is very important to understand that the protein score in an MS/MS search is not statistically rigorous. It is just a way of ranking the protein hits.



Peptide score distribution. Ions score is −10 log(P), where P is the probability that the observed match is a random event.
There are **2,356** peptide matches above identity threshold and **2,794** matches above homology threshold for **33,191** queries. On average, individual ions scores **> 32** (beyond green shading) indicate **identity or extensive homology** (p<0.05).

[Deprecated] Protein score distribution. Score distribution for family members in the first 50 proteins. Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein families.

▸ **Modification statistics**

▸ **Legend**

Protein Family Summary

Significance threshold p< 0.05 Max. number of families AUTO

Display non-sig. matches □ Dendrograms cut at 0
Show Percolator scores □
Preferred taxonomy All entries

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

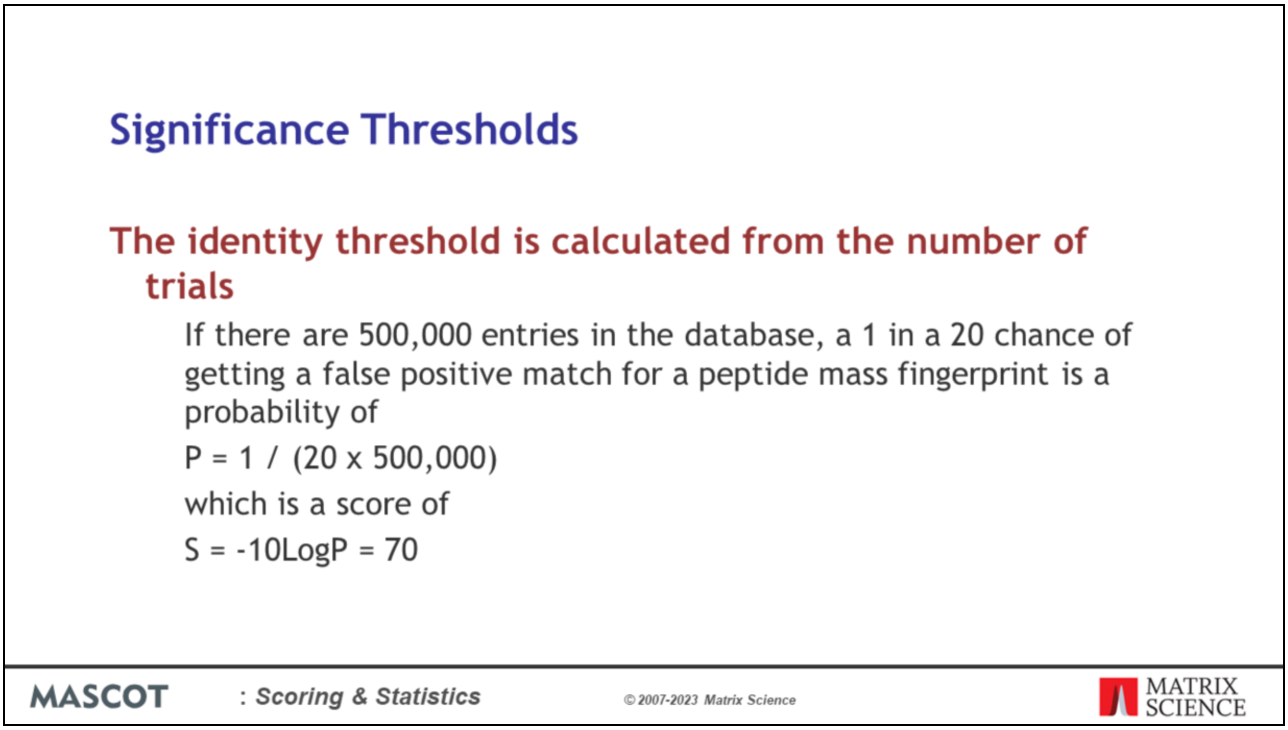
Protein families 1–10 (out of 448)

10 per page 1 2 3 4 5 6 ... 45 Next

```
▸1          1::sp|TRY1_BOVIN|     1606  sp|TRY1_BOVIN|
▸2        1 2::CP2CT_MOUSE        1337  Cytochrome P450 2C29 OS=Mus musculus GN=Cyp2c29 PE=1 SV=2
          6 2::CP239_MOUSE         251  Cytochrome P450 2C39 OS=Mus musculus GN=Cyp2c39 PE=2 SV=2
          7 2::CP238_MOUSE         150  Cytochrome P450 2C38 OS=Mus musculus GN=Cyp2c38 PE=2 SV=2
          2 2::CP254_MOUSE         552  Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
          3 2::CY250_MOUSE         489  Cytochrome P450 2C50 OS=Mus musculus GN=Cyp2c50 PE=1 SV=2
          5 2::CP237_MOUSE         339  Cytochrome P450 2C37 OS=Mus musculus GN=Cyp2c37 PE=2 SV=2
```

No expect value

This is why there is no expect value for the protein score in an MS/MS search, and why there is a short explanation at the top of every report.



Significance Thresholds

The identity threshold is calculated from the number of trials

If there are 500,000 entries in the database, a 1 in a 20 chance of getting a false positive match for a peptide mass fingerprint is a probability of

P = 1 / (20 x 500,000)

which is a score of

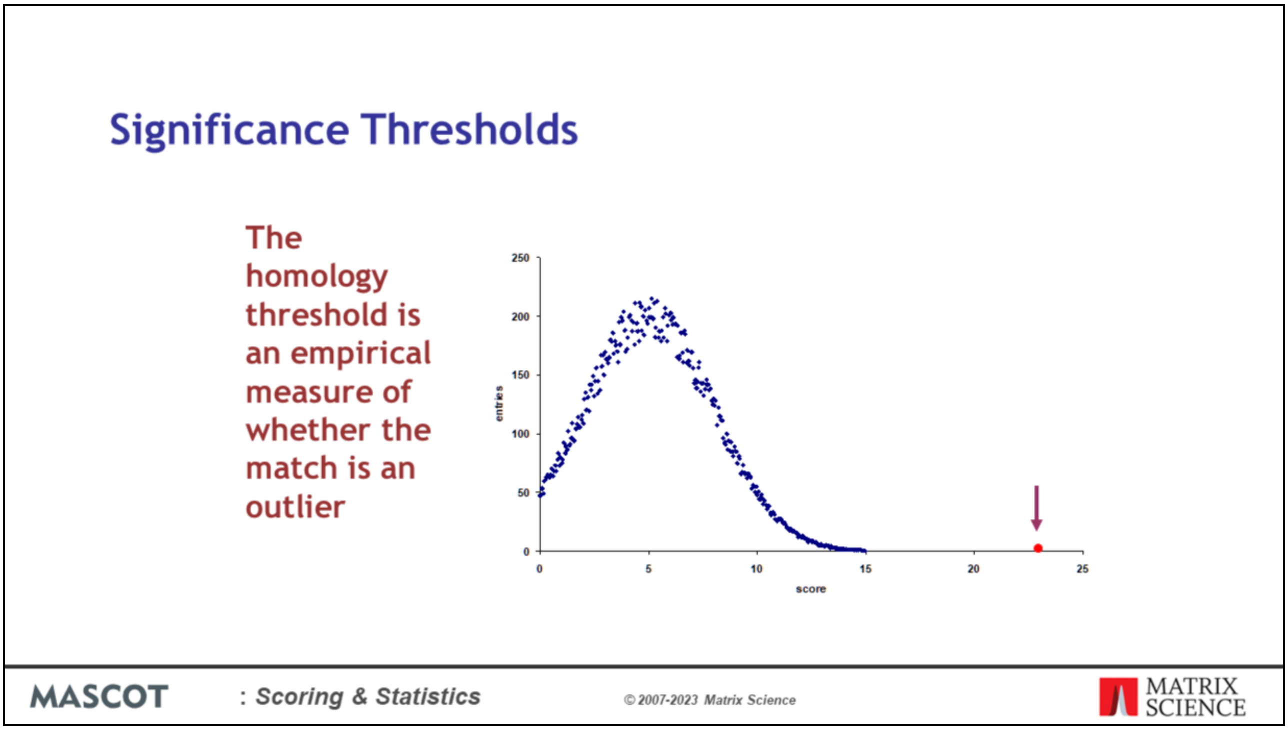
S = -10LogP = 70

Because a Mascot score is a log probability, assigning a significance threshold is very simple. It is just a function of the number of trials - the number of times we test for a match. For a peptide mass fingerprint, this is the number of entries in the database. For an MS/MS search, it is the number of peptides in the database that fit to the precursor mass tolerance. For an enzyme like trypsin, and a reasonable mass tolerance, this number will be less than the number of entries in the database. For a no-enzyme search, the number of trials will often be more than the number of entries in the database.

So, for example, if we are comfortable with a 1 in a 20 chance of getting a false positive match, and we are doing a PMF search of a database that contains 500,000 entries, we are looking for a probability of less than 1 / (20 x 500,000) which is a Mascot score of 70

If we could only tolerate a false positive rate of 1 in 200 then the threshold would be 80, 1 in 2000 90, etc.

For MS/MS searches with trypsin, and a reasonable mass tolerance, the numbers tend to be lower. The default identity threshold is typically a score of around 40.



Significance Thresholds

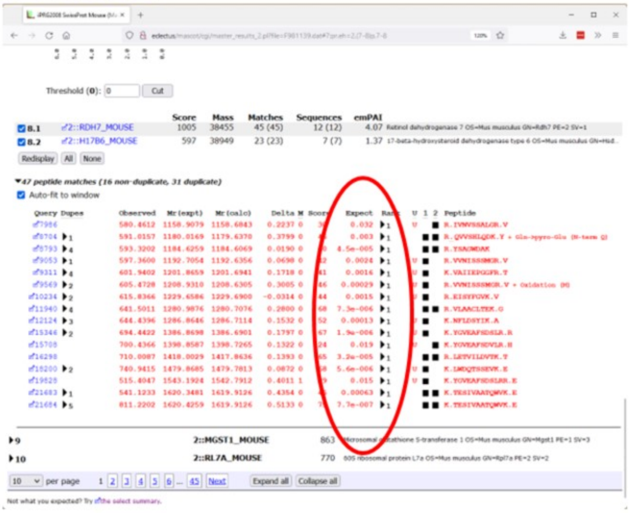
The homology threshold is an empirical measure of whether the match is an outlier

MASCOT : *Scoring & Statistics* © 2007-2023 *Matrix Science* MATRIX SCIENCE

Unfortunately, MS/MS spectra are often far from ideal, with poor signal to noise or gaps in the fragmentation. In such cases, it may not be possible to reach the identity threshold score, even though the best match in the database is a clear outlier from the distribution of random scores. To assist in identifying these outliers, we also report a second, lower threshold for MS/MS searches; the 'homology' threshold. This simply says the match is an outlier.

In practice, from measuring the actual false positive rate by searching large data sets against reversed or randomised databases, we find that the identity threshold is usually conservative, and the homology threshold can provide a useful number of additional true positive matches without exceeding the specified false positive rate.

Expectation values



Threshold (0): 0 Cut

		Score	Mass	Matches	Sequences	emPAI	
☑ 8.1	2::RDH7_MOUSE	1005	38455	45 (45)	12 (12)	4.07	Retinol dehydrogenase 7 OS=Mus musculus GN=Rdh7 PE=2 SV=1
☑ 8.2	2::H17B6_MOUSE	597	38949	23 (23)	7 (7)	1.37	17-beta-hydroxysteroid dehydrogenase type 6 OS=Mus musculus GN=Hsd...

Redisplay All None

▼47 peptide matches (16 non-duplicate, 31 duplicate)

☑ Auto-fit to window

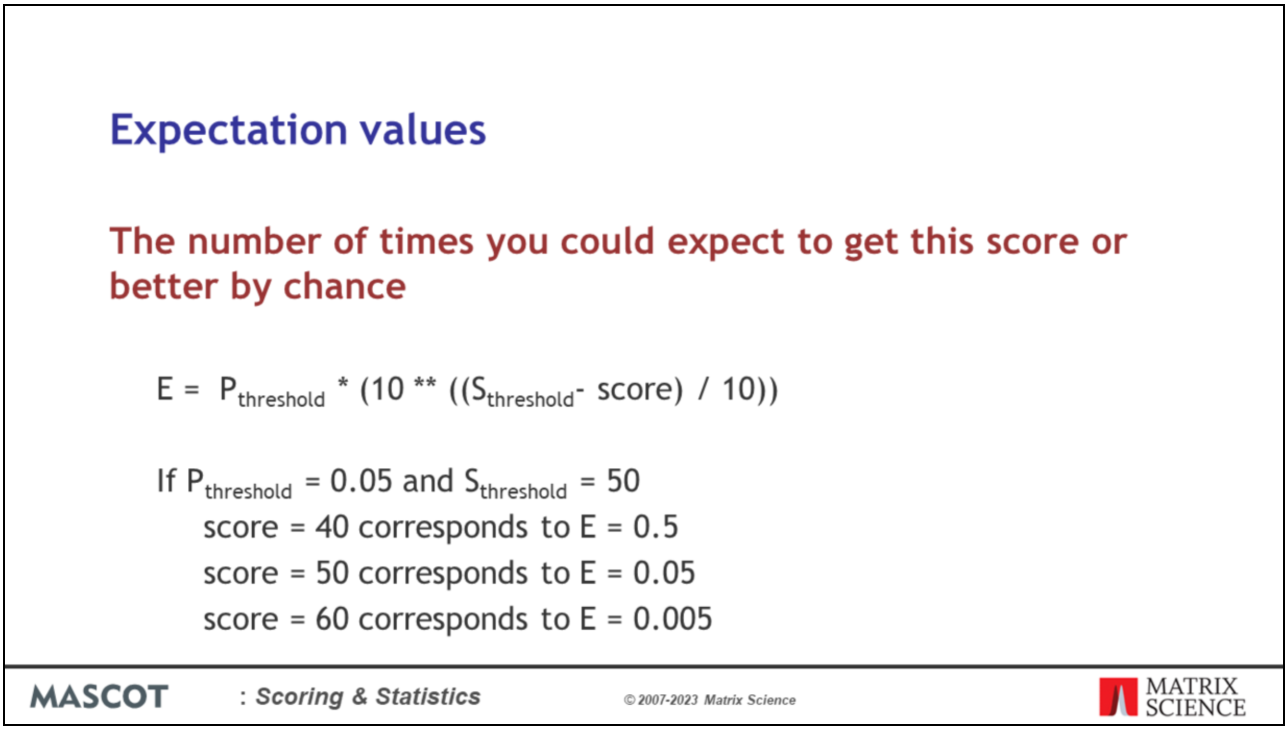
Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	Peptide
7986		580.4612	1158.9079	1158.6843	0.2237	0	[illegible]	0.032	▶1	U		■	R.IVMVSSALGR.V
8704	▶1	591.0157	1180.0169	1179.6370	0.3799	0	[illegible]	0.003	▶1		■	■	R.QVVSHLQDK.Y + Gln->pyro-Glu (N-term Q)
8793	▶4	593.3202	1184.6259	1184.6069	0.0190	0	[illegible]	4.5e-005	▶1		■	■	R.YSAGWDAK
9053	▶1	597.3600	1192.7054	1192.6356	0.0698	0	42	0.0024	▶1	U	■		R.VVNISSSMGR.V
9311	▶4	601.9402	1201.8659	1201.6941	0.1718	0	41	0.0016	▶1	U	■		K.VAIIEPGGFR.T
9569	▶2	605.4728	1208.9310	1208.6305	0.3005	0	46	0.00029	▶1	U	■		R.VVNISSSMGR.V + Oxidation (M)
10234	▶2	615.8366	1229.6586	1229.6900	-0.0314	0	44	0.0015	▶1	U	■		R.EISYFGVK.V
11940	▶4	641.5011	1280.9876	1280.7076	0.2800	0	68	7.3e-006	▶1		■	■	R.VLAACLTEK.G
12124	▶3	644.4396	1286.8646	1286.7114	0.1532	0	52	0.00013	▶1	U	■		K.NFLDSYIK.A
15346	▶2	694.4422	1386.8698	1386.6901	0.1797	0	67	1.9e-006	▶1	U	■		K.YGVEAFSDSLR.R
15708		700.4366	1398.8587	1398.7265	0.1322	0	24	0.019	▶1	U		■	K.YGVEAFSDVLR.H
16298		710.0087	1418.0029	1417.8636	0.1393	0	65	3.2e-005	▶1		■	■	R.LETVILDVTK.T
18200	▶2	740.9415	1479.8685	1479.7813	0.0872	0	[illegible]	5.6e-006	▶1	U	■		K.LMDQTSSEVK.E
19828		515.4047	1543.1924	1542.7912	0.4011	1	[illegible]	0.015	▶1	U	■		K.YGVEAFSDSLRR.E
21683	▶1	541.1233	1620.3481	1619.9126	0.4354	0	[illegible]	0.00063	▶1		■	■	K.TESIVAATQWVK.E
21684	▶5	811.2202	1620.4259	1619.9126	0.5133	0	[illegible]	7.7e-007	▶1		■	■	K.TESIVAATQWVK.E

▶9	2::MGST1_MOUSE	863	Microsomal glutathione S-transferase 1 OS=Mus musculus GN=Mgst1 PE=1 SV=3
▶10	2::RL7A_MOUSE	770	60S ribosomal protein L7a OS=Mus musculus GN=Rpl7a PE=2 SV=2

10 per page 1 2 3 4 5 6 ... 45 Next Expand all Collapse all

Not what you expected? Try the select summary.

We display an expect or expectation value in addition to the score.



Expectation values

The number of times you could expect to get this score or better by chance

$E = P_{threshold} * (10 ** ((S_{threshold} - score) / 10))$

If $P_{threshold} = 0.05$ and $S_{threshold} = 50$

score = 40 corresponds to E = 0.5

score = 50 corresponds to E = 0.05

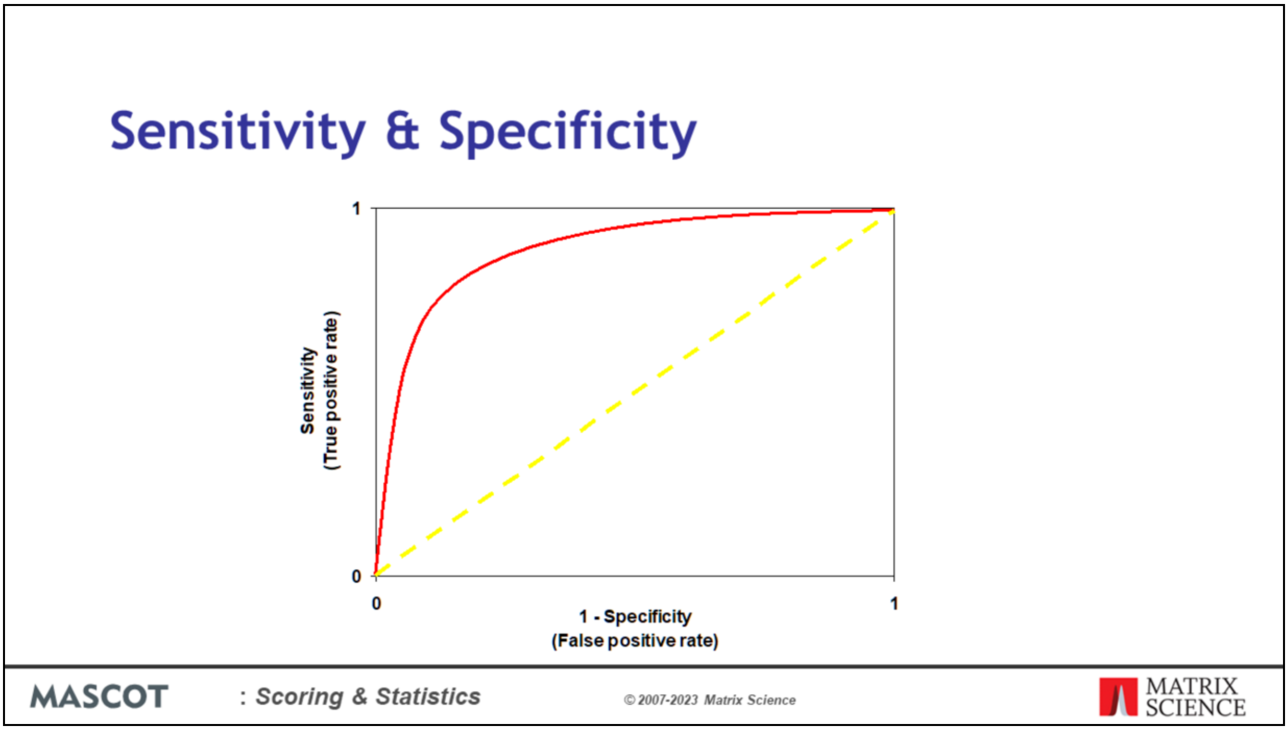
score = 60 corresponds to E = 0.005

The expectation value does not contain new information. It can be derived directly from the score and the threshold. The advantage is that it tells you everything you need to know in a single number.

It is the number of times you could expect to get this score or better by chance.

A completely random match has an expectation value of 1 or more.

The better the match, the smaller the expectation value.



Sensitivity & Specificity

MASCOT : *Scoring & Statistics*

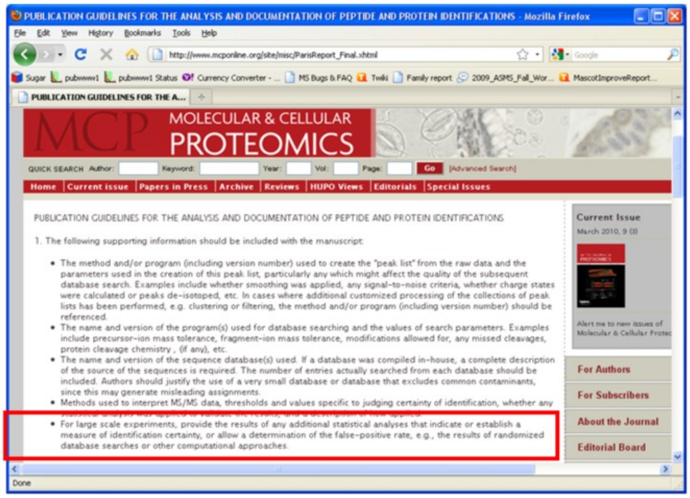
MATRIX SCIENCE

The most important attributes of a scoring scheme are sensitivity and specificity. That is, you want as many correct matches as possible, and as few incorrect matches as possible.

This is often illustrated in the form of a Receiver Operating Characteristic or ROC plot. This plots the relationship between the true positive and false positive rates as the threshold is varied. The origin is a very high threshold, which lets nothing through. At the top right, we have a very low threshold, that allows everything through. Neither extreme is a useful place to be. The diagonal represents a useless scoring algorithm, that is equally likely to let through a false match as a true one. The red curve shows a useful scoring algorithm, and the more it pushes the curve up towards the top left corner, the better. Setting a threshold towards this top left corner gives a high ratio of correct matches to false matches.

A few years ago, there was a little too much focus on sensitivity and not enough consideration given to specificity, so that some of the published lists of proteins were not as accurate as the authors might have hoped.

Validation


PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATIONS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.mcponline.org/site/misc/ParisReport_Final.xhtml
Sugar
pubwww1
pubwww1 Status
Currency Converter - ...
MS Bugs & FAQ
Twiki
Family report
2009_ASMS_Fall_Wor...
MascotImproveReport...
PUBLICATION GUIDELINES FOR THE A...
MCP
MOLECULAR & CELLULAR PROTEOMICS
QUICK SEARCH Author: Keyword: Year: Vol: Page: Go [Advanced Search]
Home
Current issue
Papers in Press
Archive
Reviews
HUPO Views
Editorials
Special Issues
Current Issue
March 2010, 9 (3)
Alert me to new issues of Molecular & Cellular Protec
For Authors
For Subscribers
About the Journal
Editorial Board
Done

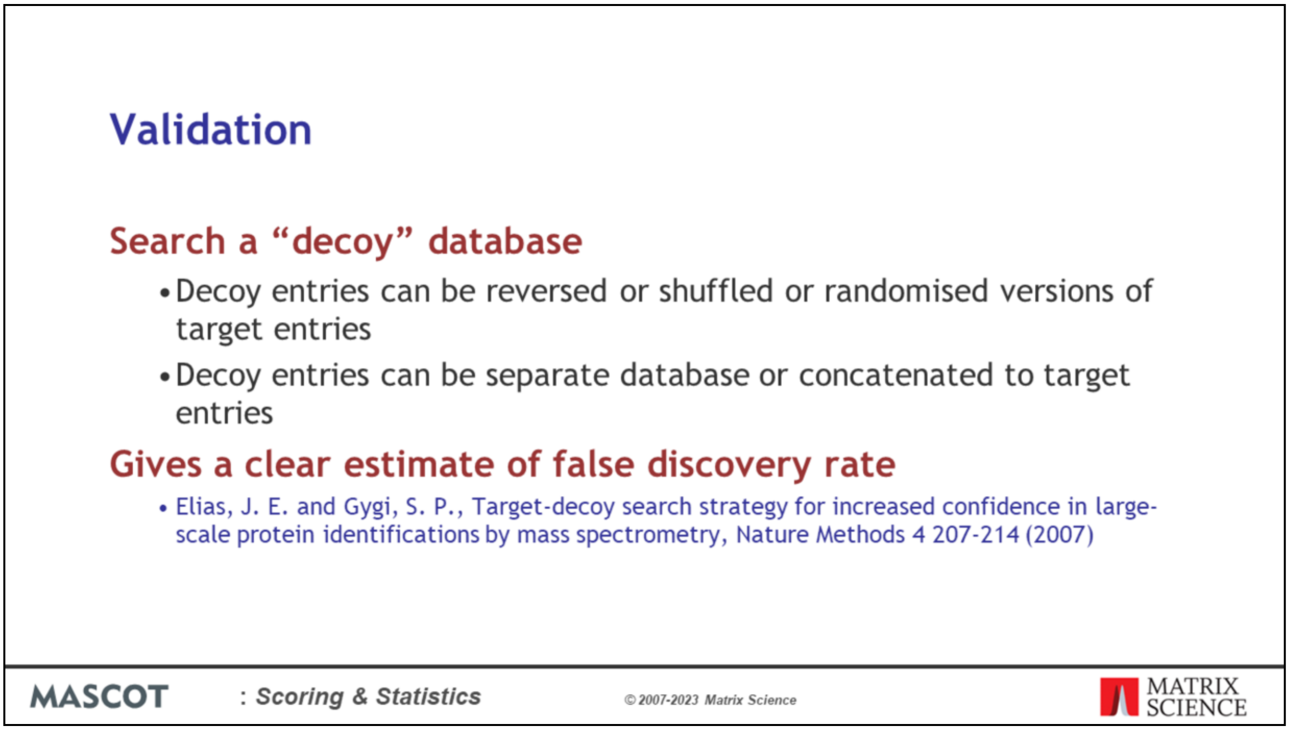
PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATIONS

1. The following supporting information should be included with the manuscript:

- The method and/or program (including version number) used to create the "peak list" from the raw data and the parameters used in the creation of this peak list, particularly any which might affect the quality of the subsequent database search. Examples include whether smoothing was applied, any signal-to-noise criteria, whether charge states were calculated or peaks de-isotoped, etc. In cases where additional customized processing of the collections of peak lists has been performed, e.g. clustering or filtering, the method and/or program (including version number) should be referenced.
- The name and version of the program(s) used for database searching and the values of search parameters. Examples include precursor-ion mass tolerance, fragment-ion mass tolerance, modifications allowed for, any missed cleavages, protein cleavage chemistry , (if any), etc.
- The name and version of the sequence database(s) used. If a database was compiled in-house, a complete description of the source of the sequences is required. The number of entries actually searched from each database should be included. Authors should justify the use of a very small database or database that excludes common contaminants, since this may generate misleading assignments.
- Methods used to interpret MS/MS data, thresholds and values specific to judging certainty of identification, whether any statistical analysis was applied to validate the results, and a description of how applied.
- 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.

A growing awareness of this problem led to initiatives from various quarters. Most notably, the Editors of Molecular and Cellular Proteomics, who held a workshop in 2005 to define a set of guidelines, which have been revised multiple times, last time in 2017.

For large scale studies, there is a requirement to estimate your false discovery rate. One of the most reliable ways to do this is with a so-called decoy database.



Validation

Search a “decoy” database

- Decoy entries can be reversed or shuffled or randomised versions of target entries
- Decoy entries can be separate database or concatenated to target entries

Gives a clear estimate of false discovery rate

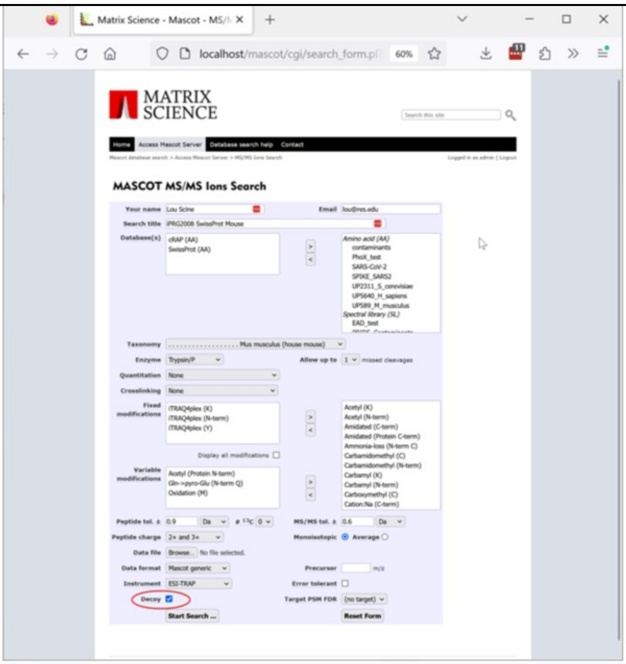
- Elias, J. E. and Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nature Methods 4 207-214 (2007)

This is very simple but very powerful. You repeat the search, using identical search parameters, against a database in which the sequences have been reversed or randomised. You do not expect to get any real matches from the decoy database. So, the number of matches that are found in the decoy database is an excellent estimate of the number of false positives in the results from the target database.

You’ll read a lot of discussion in the literature about whether the decoy sequences should be reversed or randomised; whether to search a single database containing both target and decoy sequences or separate databases. I suggest the most important thing is to do a decoy search; any decoy search. What you need to know is whether your level of false positives is 1% or 10% or 100%. Its less of a concern whether its 1% or 1.1%.

Although this is an excellent validation method for large data sets. It isn’t useful when you only have a small number of spectra, or a small database, because the numbers are too small to give an accurate estimate. Hence, this is not a substitute for a stable scoring scheme, but it is an excellent way of validating important results.

Validation


Matrix Science - Mascot - MS/I
localhost/mascot/cgi/search_form.pl
60%
MATRIX SCIENCE
Search this site
Home
Access Mascot Server
Database search help
Contact
Mascot database search > Access Mascot Server > MS/MS Ions Search
Logged in as admin | Logout

MASCOT MS/MS Ions Search

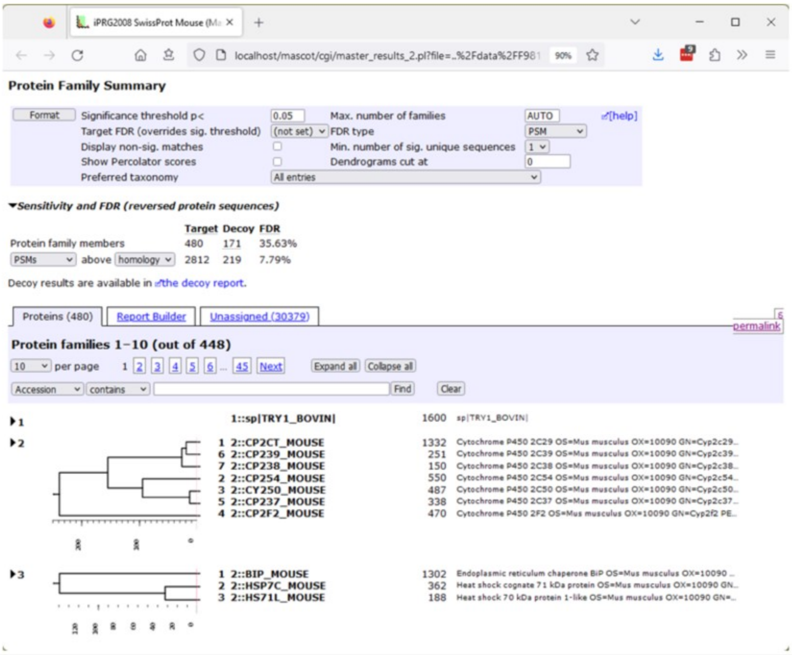
Your name: Lou Scine
Email: lou@res.edu
Search title: iPRG2008 SwissProt Mouse
Database(s): cRAP (AA), SwissProt (AA)

Taxonomy: .... Mus musculus (house mouse)
Enzyme: Trypsin/P
Allow up to 1 missed cleavages
Quantitation: None
Crosslinking: None
Fixed modifications: iTRAQ4plex (K), iTRAQ4plex (N-term), iTRAQ4plex (Y)
Display all modifications ☐
Variable modifications: Acetyl (Protein N-term), Gln->pyro-Glu (N-term Q), Oxidation (M)

Peptide tol. ± 0.9 Da # ^{13}C 0
MS/MS tol. ± 0.6 Da
Peptide charge: 2+ and 3+
Monoisotopic ● Average ○
Data file: Browse... No file selected.
Data format: Mascot generic
Precursor: m/z
Instrument: ESI-TRAP
Error tolerant ☐
Decoy ☑
Target PSM FDR: (no target)

The option to run a decoy search is built-in to Mascot Server. If you choose the Decoy checkbox on the search form, then every time a protein or peptide sequence from the target database is tested, a reversed or randomised sequence of the same length is automatically generated and tested. The average amino acid composition of the random sequences is the same as the average composition of the target database. The matches and scores for the decoy sequences are recorded separately in the result file. The result is identical to searching a separate database rather than a concatenated database.

On our public web site there is a help page devoted to decoy database searches. It includes a download link to a utility program that allows you to create a randomised or reversed database in the case that you don't want to use the built-in feature. If you have an early version of Mascot, or if you want to verify the results from another search engine, you can use this utility to create a decoy database for searching.


iPRG2008 SwissProt Mouse (M
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981
90%

Protein Family Summary

Format

Significance threshold p<	0.05	Max. number of families	AUTO
Target FDR (overrides sig. threshold)	(not set)	FDR type	PSM
Display non-sig. matches	□	Min. number of sig. unique sequences	1
Show Percolator scores	□	Dendrograms cut at	0
Preferred taxonomy	All entries		

[help]

Sensitivity and FDR (reversed protein sequences)

	Target	Decoy	FDR
Protein family members	480	171	35.63%
PSMs above homology	2812	219	7.79%

Decoy results are available in the decoy report.

Proteins (480) | Report Builder | Unassigned (30379) | permalink

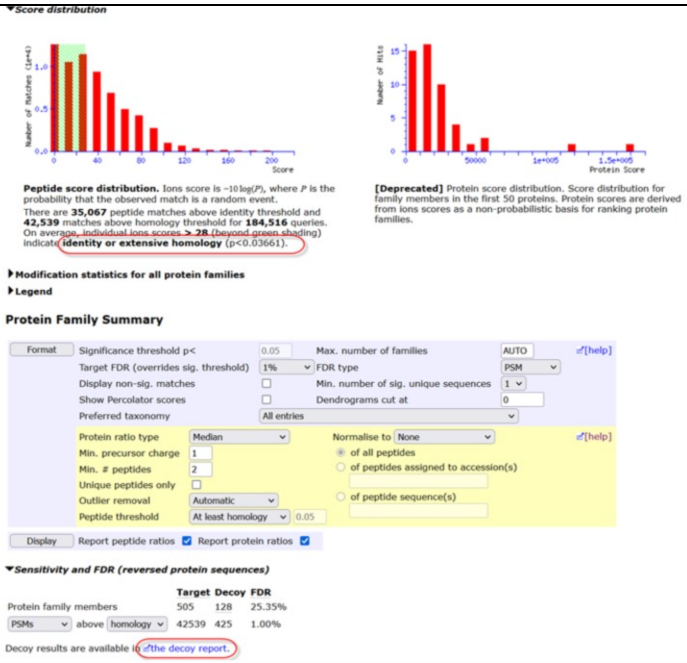
Protein families 1–10 (out of 448)

10 per page 1 2 3 4 5 6 ... 45 Next Expand all Collapse all

Accession contains Find Clear

1	1::sp|TRY1_BOVIN|	1600	sp|TRY1_BOVIN|
2	1 2::CP2CT_MOUSE	1332	Cytochrome P450 2C29 OS=Mus musculus OX=10090 GN=Cyp2c29...
	6 2::CP239_MOUSE	251	Cytochrome P450 2C39 OS=Mus musculus OX=10090 GN=Cyp2c39...
	7 2::CP238_MOUSE	150	Cytochrome P450 2C38 OS=Mus musculus OX=10090 GN=Cyp2c38...
	2 2::CP254_MOUSE	550	Cytochrome P450 2C54 OS=Mus musculus OX=10090 GN=Cyp2c54...
	3 2::CY250_MOUSE	487	Cytochrome P450 2C50 OS=Mus musculus OX=10090 GN=Cyp2c50...
	5 2::CP237_MOUSE	338	Cytochrome P450 2C37 OS=Mus musculus OX=10090 GN=Cyp2c37...
	4 2::CP2F2_MOUSE	470	Cytochrome P450 2F2 OS=Mus musculus OX=10090 GN=Cyp2f2 PE...
3	1 2::BIP_MOUSE	1302	Endoplasmic reticulum chaperone BiP OS=Mus musculus OX=10090 ...
	2 2::HSP7C_MOUSE	362	Heat shock cognate 71 kDa protein OS=Mus musculus OX=10090 GN...
	3 2::HS71L_MOUSE	188	Heat shock 70 kDa protein 1-like OS=Mus musculus OX=10090 GN=...

When the search is complete, the statistics for matches to the decoy sequences are reported in the result header. If you change the significance threshold, the numbers are recalculated. There is a option to adjust the significance threshold so as to achieve a chosen FDR value. For example, if we choose 1% FDR using the homology threshold.



▾*Score distribution*

Peptide score distribution. Ions score is $-10\log(P)$, where P is the probability that the observed match is a random event.

There are **35,067** peptide matches above identity threshold and **42,539** matches above homology threshold for **184,516** queries. On average, individual ions scores **> 28** (beyond green shading) indicate **identity or extensive homology** (p<0.03661).

[Deprecated] Protein score distribution. Score distribution for family members in the first 50 proteins. Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein families.

▸**Modification statistics for all protein families**

▸**Legend**

Protein Family Summary

Format | Significance threshold p< 0.05 | Max. number of families AUTO | [help]
Target FDR (overrides sig. threshold) 1% | FDR type PSM
Display non-sig. matches □ | Min. number of sig. unique sequences 1
Show Percolator scores □ | Dendrograms cut at 0
Preferred taxonomy All entries

Protein ratio type Median | Normalise to None | [help]
Min. precursor charge 1 | ● of all peptides
Min. # peptides 2 | ○ of peptides assigned to accession(s)
Unique peptides only □
Outlier removal Automatic | ○ of peptide sequence(s)
Peptide threshold At least homology 0.05

Display | Report peptide ratios ☑ Report protein ratios ☑

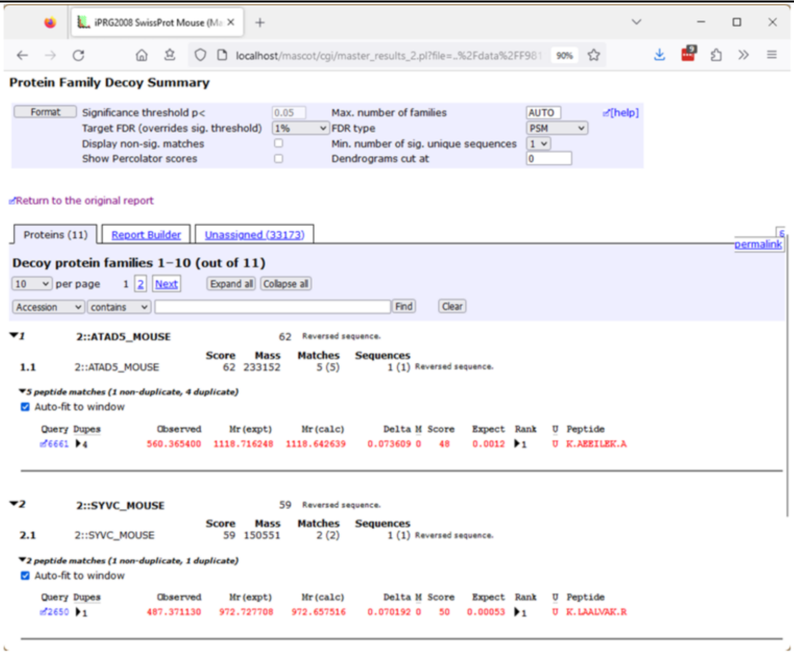
▾***Sensitivity and FDR (reversed protein sequences)***

	Target	Decoy	FDR
Protein family members	505	128	25.35%
PSMs above homology	42539	425	1.00%

Decoy results are available in the decoy report.

The significance threshold has been automatically adjusted. You can see the adjustment by expanding the Score distribution section report where it displays the change from 0.05 to 0.03661. In the report the expect values are recalculated and matches with an expect below 0.05 are significant.

Why do we get these false positives? Do they reflect some defect in the search engine? Let’s have a closer look. If you click the decoy report link as shown at the bottom of the screen shot in the expanded Sensitivity and FDR section, then you will see the results from searching the randomised database.


iPRG2008 SwissProt Mouse (Ma
localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2FF981
90%
Format
[help]
Return to the original report
Proteins (11)
Report Builder
Unassigned (33173)
permalink
Expand all
Collapse all
Find
Clear

Protein Family Decoy Summary

Significance threshold p< 0.05
Target FDR (overrides sig. threshold) 1%
Display non-sig. matches □
Show Percolator scores □
Max. number of families AUTO
FDR type PSM
Min. number of sig. unique sequences 1
Dendrograms cut at 0

Decoy protein families 1–10 (out of 11)

10 per page 1 2 Next

Accession contains

1 **2::ATAD5_MOUSE** 62 Reversed sequence.

		Score	Mass	Matches	Sequences	
1.1	2::ATAD5_MOUSE	62	233152	5 (5)	1 (1)	Reversed sequence.

5 peptide matches (1 non-duplicate, 4 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	Peptide
6661	▶4	560.365400	1118.716248	1118.642639	0.073609	0	48	0.0012	▶1	U	K.AEEILEK.A

2 **2::SYVC_MOUSE** 59 Reversed sequence.

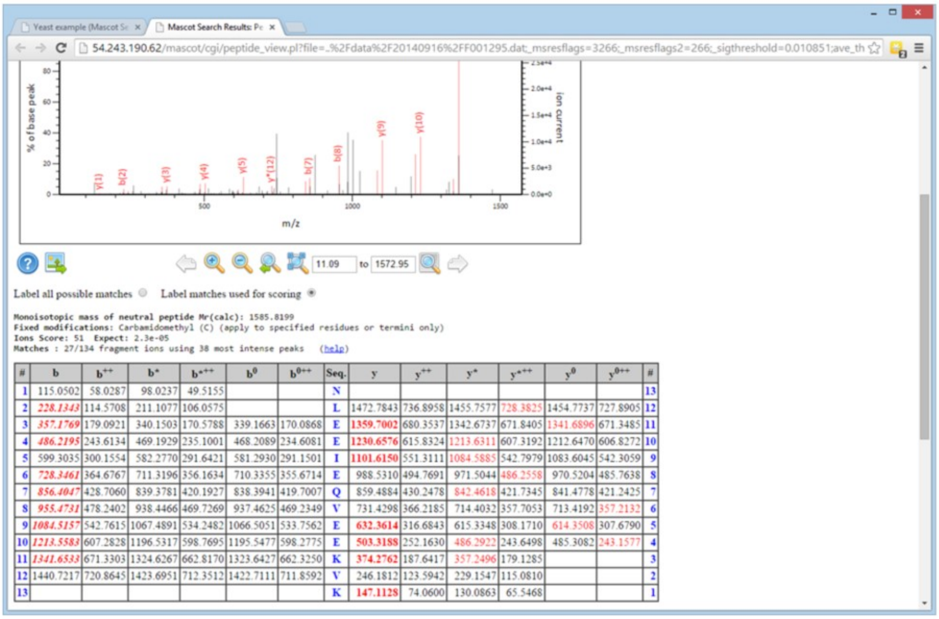
		Score	Mass	Matches	Sequences	
2.1	2::SYVC_MOUSE	59	150551	2 (2)	1 (1)	Reversed sequence.

2 peptide matches (1 non-duplicate, 1 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	Peptide
2650	▶1	487.371130	972.727708	972.657516	0.070192	0	50	0.00053	▶1	U	K.LAALVAK.R

The results from the matches to the randomised sequences are saved in a different section of the results file on the Mascot server. This means that we can view these results in exactly the same way as if we had performed a separate search against a randomised database that we had created manually. We can see matches here with scores of 37 and 51, with expect values well below 1%/0.01. If we click on the query number link to display the Peptide View of one of these matches …



Yeast example (Mascot Se ×
Mascot Search Results: Pe ×
54.243.190.62/mascot/cgi/peptide_view.pl?file=..%2Fdata%2F20140916%2FF001295.dat;_msresflags=3266;_msresflags2=266;_sigthreshold=0.010851;ave_th
% of base peak
80
60
40
20
0
y(1)
b(2)
y(3)
y(4)
y(5)
y*(12)
b(7)
b(8)
y(9)
y(10)
500
1000
1500
m/z
2.5e+4
2.0e+4
1.5e+4
1.0e+4
5.0e+3
0.0e+0
ion current
11.09
to
1572.95

Label all possible matches ○ Label matches used for scoring ●

```
Monoisotopic mass of neutral peptide Mr(calc): 1585.8199
Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini only)
Ions Score: 51  Expect: 2.3e-05
Matches : 27/134 fragment ions using 38 most intense peaks   (help)
```

#	b	b^{++}	b^{*}	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	115.0502	58.0287	98.0237	49.5155			N							13
2	***228.1343***	114.5708	211.1077	106.0575			L	1472.7843	736.8958	1455.7577	728.3825	1454.7737	727.8905	12
3	***357.1769***	179.0921	340.1503	170.5788	339.1663	170.0868	E	**1359.7002**	680.3537	1342.6737	671.8405	1341.6896	671.3485	11
4	***486.2195***	243.6134	469.1929	235.1001	468.2089	234.6081	E	**1230.6576**	615.8324	1213.6311	607.3192	1212.6470	606.8272	10
5	599.3035	300.1554	582.2770	291.6421	581.2930	291.1501	I	**1101.6150**	551.3111	1084.5885	542.7979	1083.6045	542.3059	9
6	***728.3461***	364.6767	711.3196	356.1634	710.3355	355.6714	E	988.5310	494.7691	971.5044	486.2558	970.5204	485.7638	8
7	***856.4047***	428.7060	839.3781	420.1927	838.3941	419.7007	Q	859.4884	430.2478	842.4618	421.7345	841.4778	421.2425	7
8	***955.4731***	478.2402	938.4466	469.7269	937.4625	469.2349	V	731.4298	366.2185	714.4032	357.7053	713.4192	357.2132	6
9	***1084.5157***	542.7615	1067.4891	534.2482	1066.5051	533.7562	E	**632.3614**	316.6843	615.3348	308.1710	614.3508	307.6790	5
10	***1213.5583***	607.2828	1196.5317	598.7695	1195.5477	598.2775	E	**503.3188**	252.1630	486.2922	243.6498	485.3082	243.1577	4
11	***1341.6533***	671.3303	1324.6267	662.8170	1323.6427	662.3250	K	**374.2762**	187.6417	357.2496	179.1285			3
12	1440.7217	720.8645	1423.6951	712.3512	1422.7111	711.8592	V	246.1812	123.5942	229.1547	115.0810			2
13							K	**147.1128**	74.0600	130.0863	65.5468			1

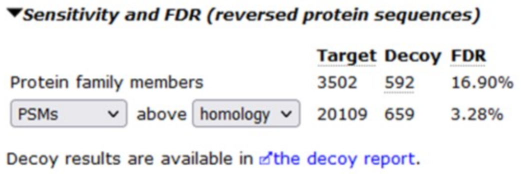
This is what it looks like. A pretty decent match from a decoy sequence. Tryptic peptide, no variable modifications, good run of b and y ions, most of the larger peaks matched.

Asking whether it is correct or wrong becomes almost a philosophical question.

The fact is, when we search large numbers of spectra against large sequence databases, we can get such matches by chance. No amount of expert manual inspection will prevent this. Database matching is a statistical process and, for this search, the number and magnitude of the false positives are well within the predicted range, which is all we can ask for.

Protein FDR

Automatically calculated when decoy option selected
Displayed in the decoy section:



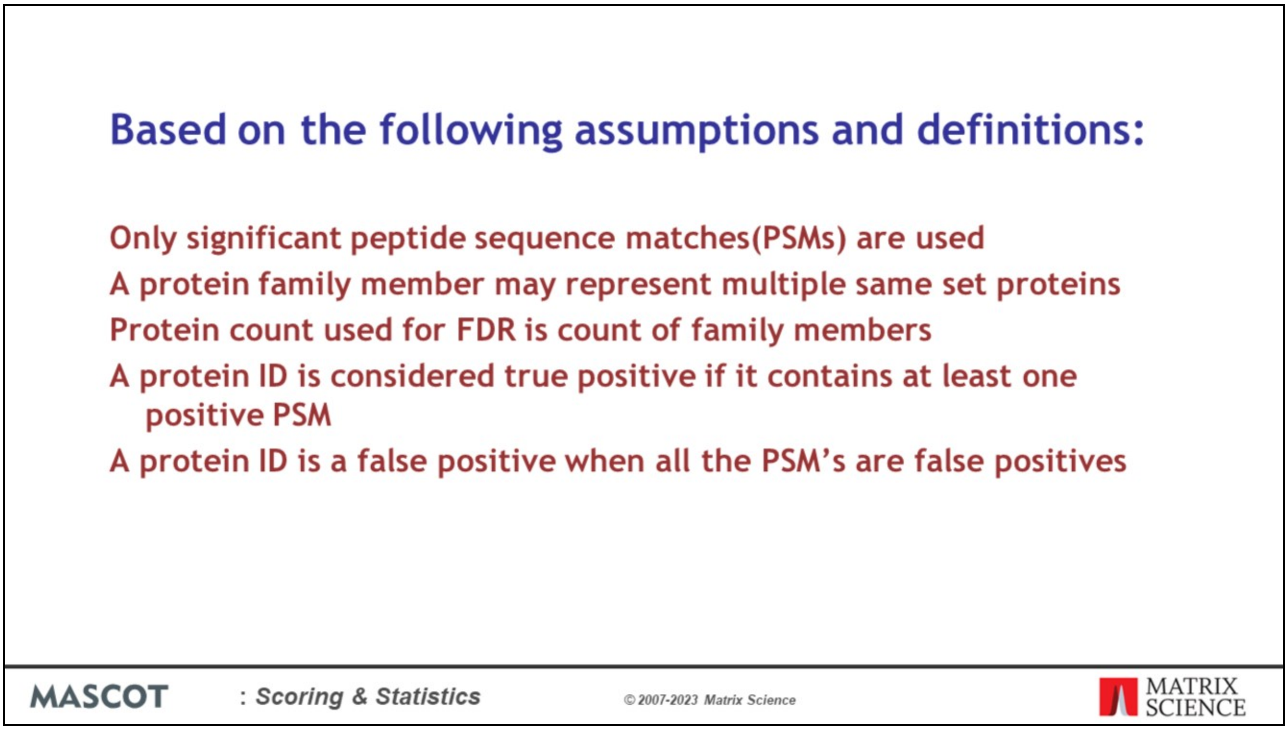
▼*Sensitivity and FDR (reversed protein sequences)*

	Target	Decoy	FDR
Protein family members	3502	592	16.90%
PSMs ⌄ above homology ⌄	20109	659	3.28%

Decoy results are available in the decoy report.

In Mascot Server 2.7, we introduced the Protein FDR value which is automatically calculated when the decoy search option is selected.

It is displayed above the peptide/PSM decoy results.



Based on the following assumptions and definitions:

Only significant peptide sequence matches(PSMs) are used

A protein family member may represent multiple same set proteins

Protein count used for FDR is count of family members

A protein ID is considered true positive if it contains at least one positive PSM

A protein ID is a false positive when all the PSM's are false positives

MASCOT : *Scoring & Statistics* © 2007-2023 Matrix Science MATRIX SCIENCE

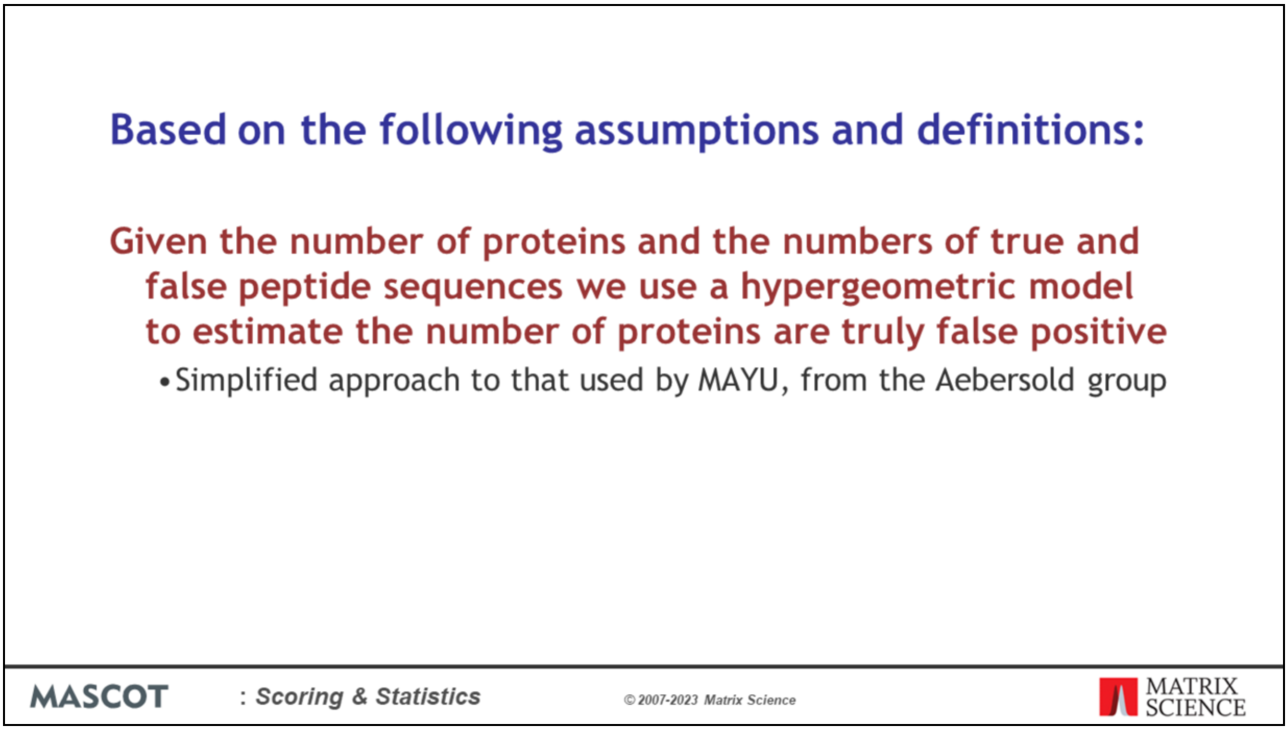
Protein FDR is based on the following assumptions and definitions:

By default, the protein family report only shows peptide sequence matches (PSMs) with significant scores and are used for the protein family assignment.

A protein family member may represent multiple same set proteins. Only members of all the protein families in the report, those that contain a unique peptide, are counted.

Protein count used for FDR is count of family members. That is, if the report contains 2 families, one with 4 members and the other with a single member, this counts as a total of 5 proteins. Same-set, sub-set and intersection proteins are not counted, because they have no independent peptide evidence.

While a protein ID is a false positive when all the PSM's are false positives. Just one true PMS would make the protein identification a true positive. This is very important.



Based on the following assumptions and definitions:

Given the number of proteins and the numbers of true and false peptide sequences we use a hypergeometric model to estimate the number of proteins are truly false positive

- Simplified approach to that used by MAYU, from the Aebersold group

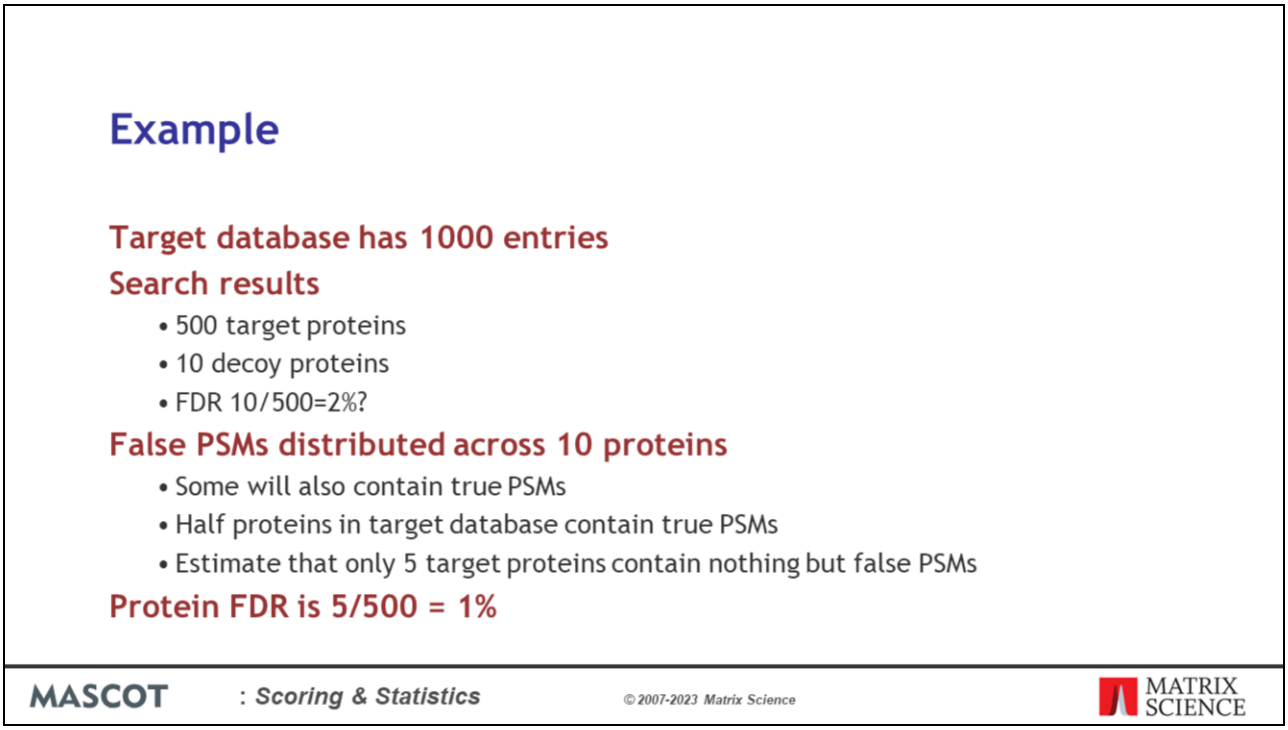
MASCOT : *Scoring & Statistics* © 2007-2023 Matrix Science MATRIX SCIENCE

Given the number of proteins and the numbers of true and false peptide sequences we use a hypergeometric model to estimate the number of proteins that are truly false positive.

The algorithm is a simplified approach to that used by MAYU, from the Aebersold group:

https://www.mcponline.org/content/8/11/2405https://www.mcponline.org/content/8/11/2405

The main differences are that we do not make a separate estimate of the FDR for one-hit wonders and we do not partition the database by protein size. We use a simpler estimate for the number of false proteins in the target database, based on the assumption that the number of decoy proteins never reaches a significant proportion of the database size.



Example

Target database has 1000 entries

Search results

- 500 target proteins
- 10 decoy proteins
- FDR 10/500=2%?

False PSMs distributed across 10 proteins

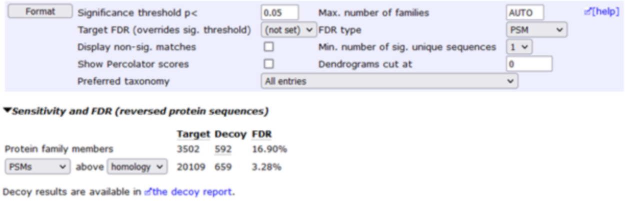
- Some will also contain true PSMs
- Half proteins in target database contain true PSMs
- Estimate that only 5 target proteins contain nothing but false PSMs

Protein FDR is 5/500 = 1%

Imagine the and the search results show 500 target proteins and 10 decoy proteins. Does this mean protein FDR is 10/500 = 2%? No, it does not. We can assume the false PSMs in the target are distributed across 10 proteins, but some of these will also contain true PSMs, so should not be counted as false. Since half the proteins in the target database contain true PSMs, a reasonable estimate would be that only 5 target proteins containing nothing but false PSMs, so that the protein FDR is 5/500 = 1%.

Adjusting the Protein FDR

Default protein FDR



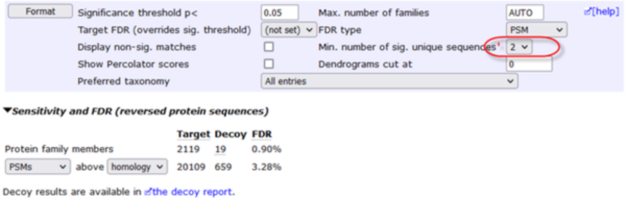
Format	Significance threshold p<	0.05	Max. number of families AUTO [help]
	Target FDR (overrides sig. threshold)	(not set)	FDR type PSM
	Display non-sig. matches	□	Min. number of sig. unique sequences 1
	Show Percolator scores	□	Dendrograms cut at 0
	Preferred taxonomy	All entries	

▼*Sensitivity and FDR (reversed protein sequences)*

	Target	Decoy	FDR
Protein family members	3502	592	16.90%
PSMs above homology	20109	659	3.28%

Decoy results are available in the decoy report.

Set Min. number sig. unique peptide sequences to 2



Format	Significance threshold p<	0.05	Max. number of families AUTO [help]
	Target FDR (overrides sig. threshold)	(not set)	FDR type PSM
	Display non-sig. matches	□	Min. number of sig. unique sequences 2
	Show Percolator scores	□	Dendrograms cut at 0
	Preferred taxonomy	All entries	

▼*Sensitivity and FDR (reversed protein sequences)*

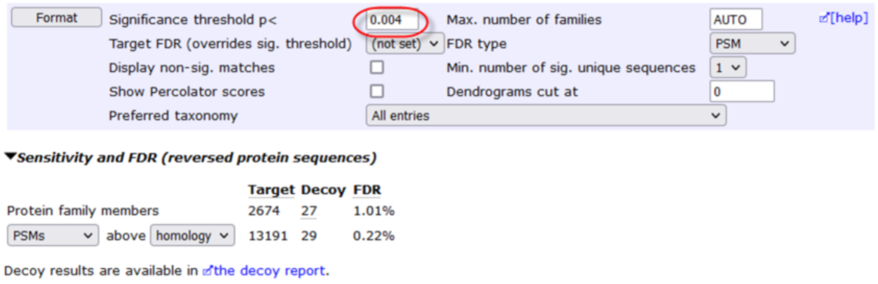
	Target	Decoy	FDR
Protein family members	2119	19	0.90%
PSMs above homology	20109	659	3.28%

Decoy results are available in the decoy report.

The default significance threshold for a Mascot search is usually 0.05 and this will often give a peptide FDR in the region of 5%. In this dataset the protein FDR is ~17%. If we want to lower the Protein FDR, we can try adjusting the peptide FDR to 1%. We can also try adjusting the report other ways.

We can adjust the Minimum number of significant unique sequences. This has quite a strong affect on the Protein FDR. If we change it to 2 and eliminate the “one hit wonders” the protein FDR drops to 0.9%. Note that “one hit wonders” are not necessarily false positives, just that they only have one significant peptide sequence match.

Adjusting the Protein FDR



Format	Significance threshold p<	0.004	Max. number of families	AUTO
	Target FDR (overrides sig. threshold)	(not set)	FDR type	PSM
	Display non-sig. matches	□	Min. number of sig. unique sequences	1
	Show Percolator scores	□	Dendrograms cut at	0
	Preferred taxonomy	All entries		

[help]

▼*Sensitivity and FDR (reversed protein sequences)*

	Target	Decoy	FDR
Protein family members	2674	27	1.01%
PSMs above homology	13191	29	0.22%

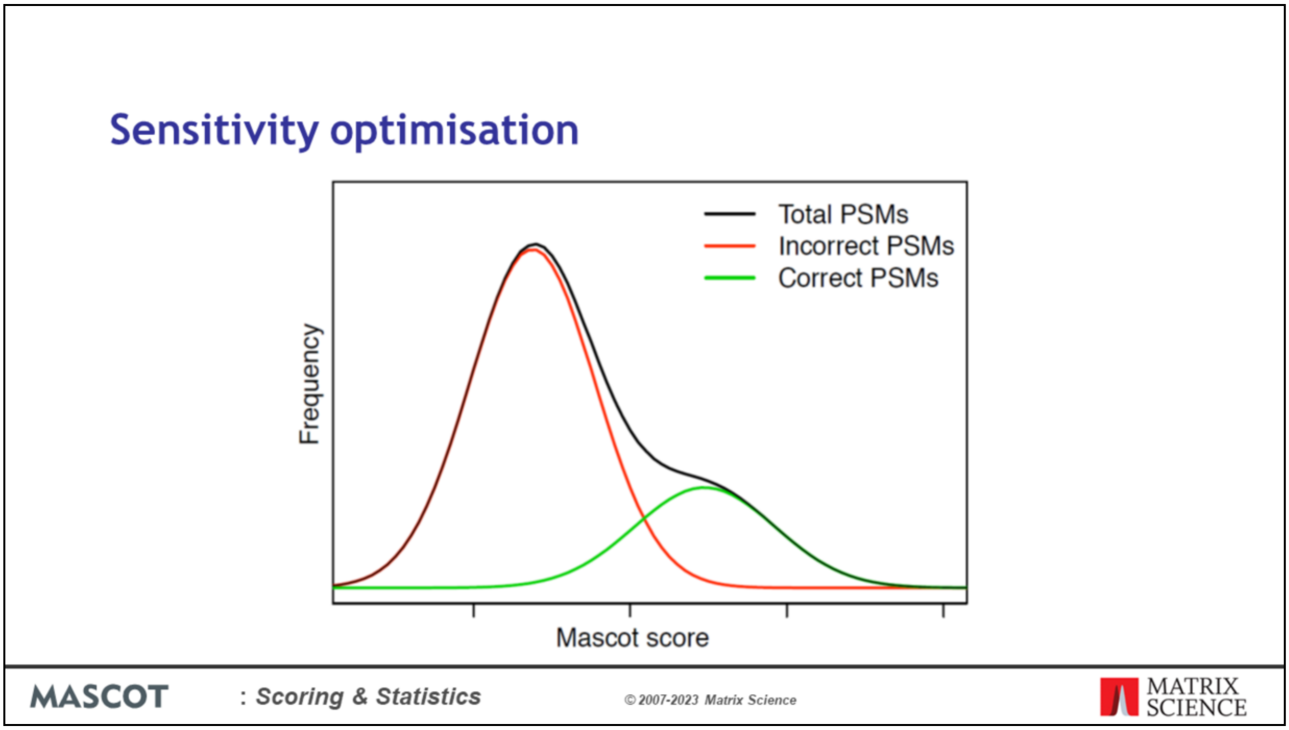
Decoy results are available in the decoy report.

Alternatively, we can adjust the Significance threshold for the results. I took a guess and reduced it by a factor of 10 from the default values of 0.05 to 0.004 and clicked the format button. The resulting Protein FDR is approximately 1%.

The current HUPO guidelines Interpretation Guidelines for large-scale results recommend adjusting the settings to lower than 1% protein-level global FDR so after formatting this search result would meet those guidelines.

When interpreting the results, a protein FDR of 1% only tells us that 1% of the proteins listed are wholly false. This doesn't mean the other 99% are "correct". In particular, where there are same-set proteins, we cannot say which one is "correct".

This is because database redundancy causes protein inference ambiguity and we can account for the PSM evidence using several sets of proteins. It is important to remember that a protein accession number in the summary report does not mean "this is the correct protein", it means "the correct protein is likely to be very similar to one of the set of proteins represented by this family member".

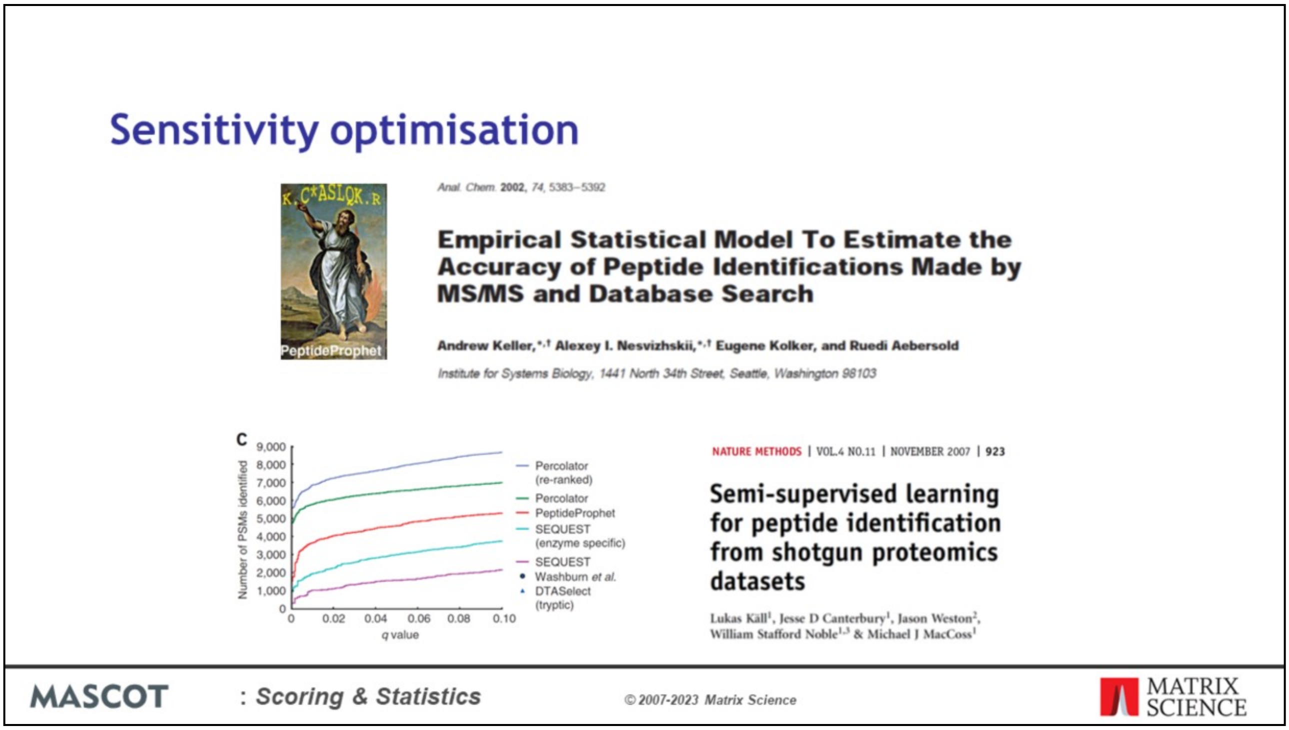


Sensitivity optimisation

MASCOT : *Scoring & Statistics* © 2007-2023 Matrix Science MATRIX SCIENCE

Sensitivity improvement is always a hot topic. A limitation of database matching is that even the best scoring scheme cannot fully separate the correct and incorrect matches, as shown here in a schematic way. The score distribution for the correct matches overlaps that of the incorrect matches. When we use a decoy search we are deciding where to place a threshold of some sort.

But, what if we could find ways to pull these two distributions further apart? In other words, improve the specificity of the scoring.



Sensitivity optimisation

Anal. Chem. 2002, 74, 5383–5392

Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search

Andrew Keller,*,† Alexey I. Nesvizhskii,*,† Eugene Kolker, and Ruedi Aebersold

Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103

NATURE METHODS | VOL.4 NO.11 | NOVEMBER 2007 | 923

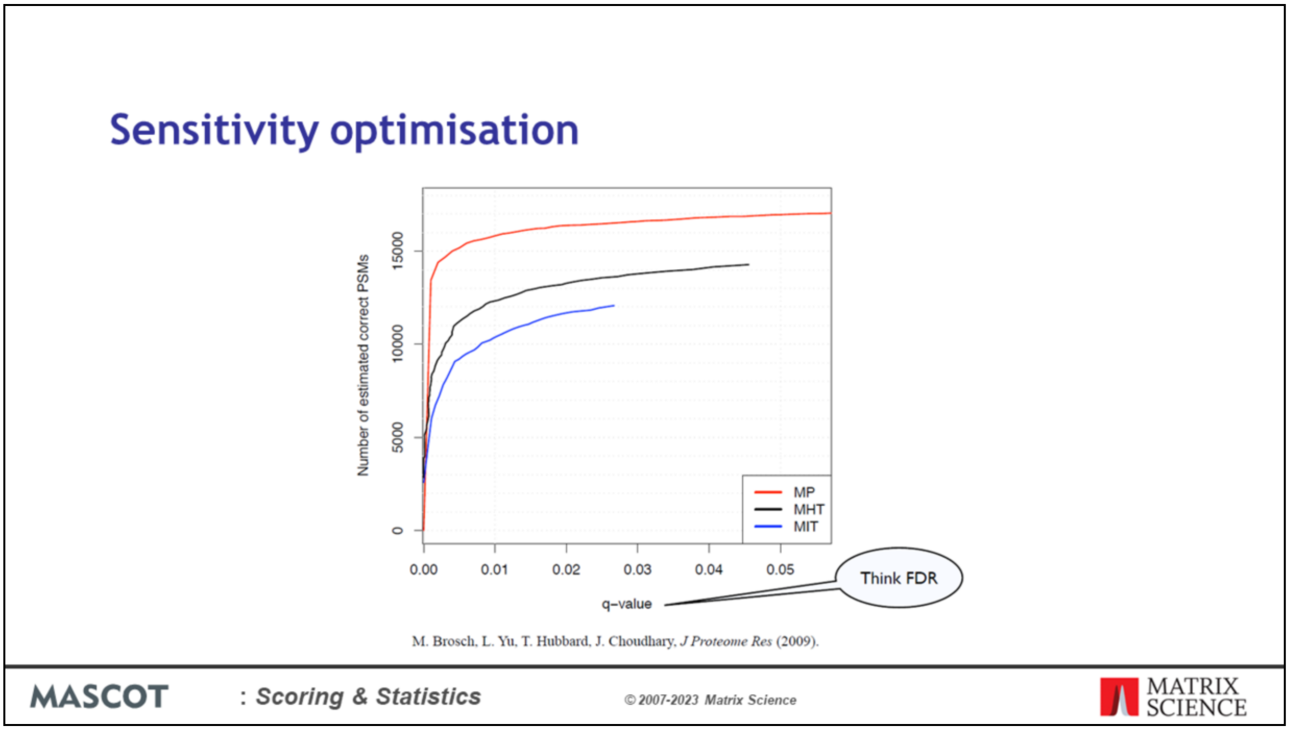
Semi-supervised learning for peptide identification from shotgun proteomics datasets

Lukas Käll[1], Jesse D Canterbury[1], Jason Weston[2], William Stafford Noble[1,3] & Michael J MacCoss[1]

One of the first attempts to do this was Peptide Prophet from the ISB. This was and is popular for transforming Sequest scores into probabilities.

It takes information about the matches in addition to the score, and uses an algorithm called expectation maximization to learn what distinguishes correct from incorrect matches. Examples of additional information would be precursor mass error, number of missed cleavages, or the number of tryptic terminii.

We can use the matches from a decoy database as negative examples for the classifier. Another popular tool, Percolator, trains a machine learning algorithm called a support vector machine to discriminate between a sub-set of the high-scoring matches from the target database, assumed correct, and the matches from the decoy database, assumed incorrect.

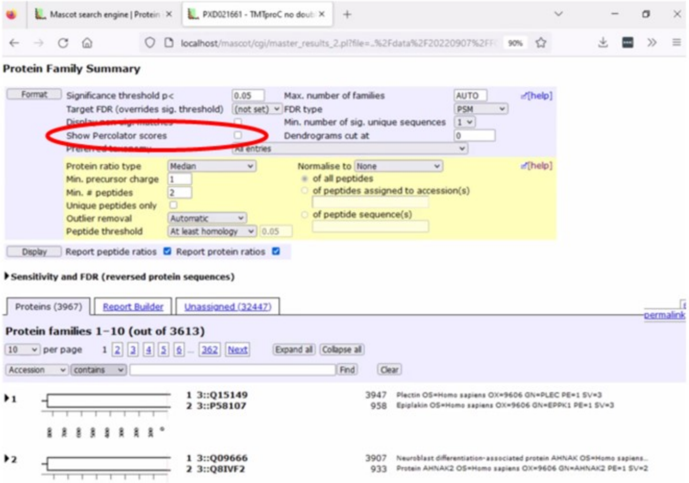


Sensitivity optimisation

M. Brosch, L. Yu, T. Hubbard, J. Choudhary, *J Proteome Res* (2009).

Percolator can give very substantial improvements in sensitivity. The original Percolator was implemented mainly with Sequest in mind, but Markus Brosch at the Sanger Centre wrote a wrapper that allowed it to be used with Mascot results and published results such as this. The black trace is the sensitivity using the Mascot homology threshold and the red trace is the sensitivity after processing through Percolator. It doesn't work for every single data set. But, when it does work, the improvements can be most impressive.

Sensitivity optimisation



Mascot search engine | Protein × PXD021661 - TMTproC no doub ×

localhost/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20220907%2FF

Protein Family Summary

Format Significance threshold p< 0.05 Max. number of families AUTO [help]
Target FDR (overrides sig. threshold) (not set) FDR type PSM
Display non-sig. matches □ Min. number of sig. unique sequences 1
Show Percolator scores □ Dendrograms cut at 0
Preferred taxonomy All entries

Protein ratio type Median Normalise to None [help]
Min. precursor charge 1 ● of all peptides
Min. # peptides 2 ○ of peptides assigned to accession(s)
Unique peptides only □
Outlier removal Automatic ○ of peptide sequence(s)
Peptide threshold At least homology 0.05

Display Report peptide ratios ☑ Report protein ratios ☑

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

Proteins (3967) | Report Builder | Unassigned (32447) permalink

Protein families 1–10 (out of 3613)

10 per page 1 2 3 4 5 6 … 362 Next Expand all Collapse all

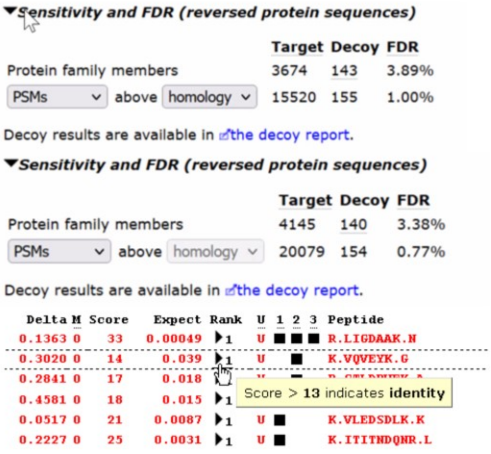
Accession contains Find Clear

▸1
1 3::Q15149 3947 Plectin OS=Homo sapiens OX=9606 GN=PLEC PE=1 SV=3
2 3::P58107 958 Epiplakin OS=Homo sapiens OX=9606 GN=EPPK1 PE=1 SV=3

▸2
1 3::Q09666 3907 Neuroblast differentiation-associated protein AHNAK OS=Homo sapiens…
2 3::Q8IVF2 933 Protein AHNAK2 OS=Homo sapiens OX=9606 GN=AHNAK2 PE=1 SV=2

The developers of Percolator have kindly agreed to allow us to distribute and install Percolator as part of Mascot Server. This option will be available for any search that has at least 100 MS/MS spectra and auto-decoy results, but it works best if there are several thousand spectra. To switch to Percolator scores, just check the box and then choose Filter. In this example we take a medium sized search result.

Sensitivity optimisation



▼Sensitivity and FDR (reversed protein sequences)

	Target	Decoy	FDR
Protein family members	3674	143	3.89%
PSMs ⌄ above homology ⌄	15520	155	1.00%

Decoy results are available in the decoy report.

▼Sensitivity and FDR (reversed protein sequences)

	Target	Decoy	FDR
Protein family members	4145	140	3.38%
PSMs ⌄ above homology ⌄	20079	154	0.77%

Decoy results are available in the decoy report.

Delta	M	Score	Expect	Rank	U	1	2	3	Peptide
0.1363	0	33	0.00049	▶1	U	■	■	■	R.LIGDAAK.N
0.3020	0	14	0.039	▶1	U		■		K.VQVEYK.G
0.2841	0	17	0.018	[illegible]	[illegible]		■		[illegible]
0.4581	0	18	0.015	▶1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.0517	0	21	0.0087	▶1	U	■			K.VLEDSDLK.K
0.2227	0	25	0.0031	▶1	U	■			K.ITITNDQNR.L

Score > 13 indicates **identity**

Using the Mascot homology threshold for a 1% false discovery rate, there are 15520 peptide matches. Re-scoring with Percolator gives a useful increase to 20079 matches.

Note that, in general, the scores are lower after switching to Percolator. The Posterior error probability is tabulated in the expect column. A Mascot score is calculated from the expect value and the single score threshold, which we describe as the identity threshold, has a fixed value of 13 (-10 log 0.05). By keeping the score, threshold, and expect value consistent, we hope to avoid breaking any third party software that expects to find these values.