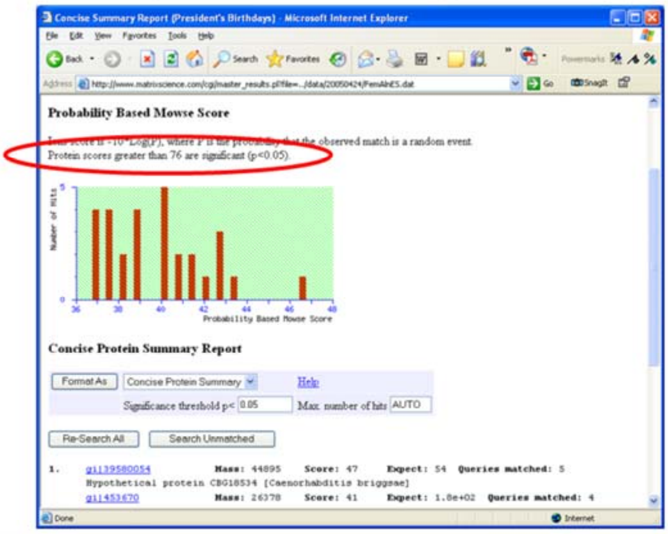
Scoring & Statistics

MASCOT

MATRIX SCIENCE

Probability based scoring



Concise Summary Report (President's Birthdays) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://www.matrixscience.com/cgi/master_results.pl?file=../data/20050424/FemAhES.dat
Go
SnagIt

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 76 are significant (p<0.05).

Number of Hits
5
0
36
38
40
42
44
46
48
Probability Based Mowse Score

Concise Protein Summary Report

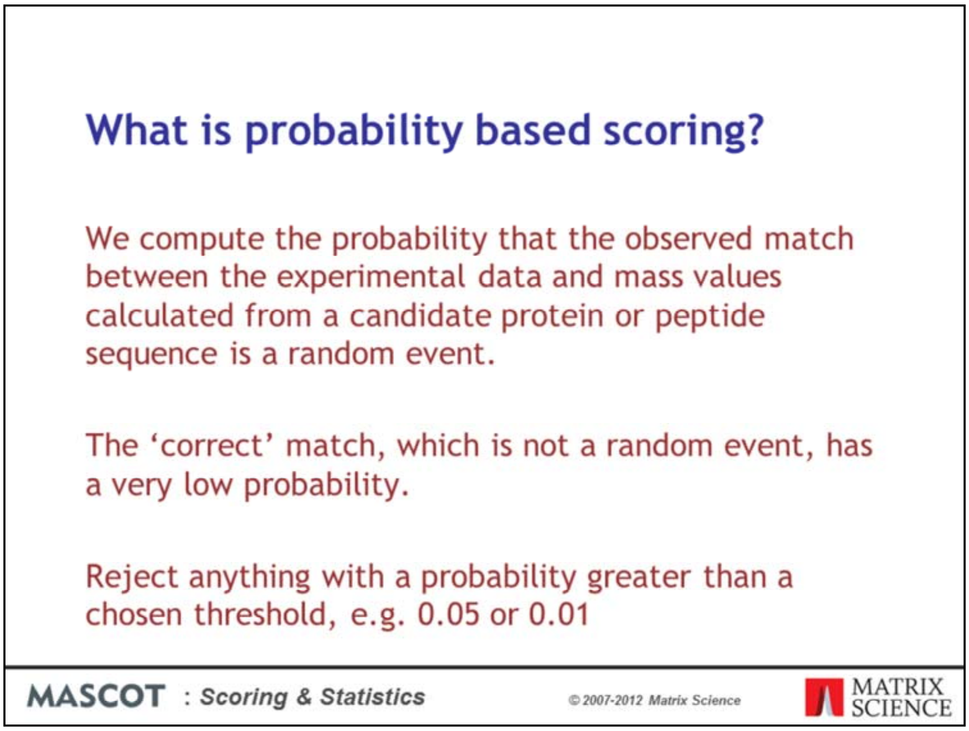
Format As
Concise Protein Summary
Help

Significance threshold p< 0.05   Max. number of hits AUTO

Re-Search All
Search Unmatched

```
1.   gi|39580054      Mass: 44895   Score: 47   Expect: 54   Queries matched: 5
     Hypothetical protein CBG18534 [Caenorhabditis briggsae]
     gi|453670        Mass: 26378   Score: 41   Expect: 1.8e+02   Queries matched: 4
```

This is the Mascot result report for a peptide mass fingerprint search. There is a list of proteins, each of which matches some of the experimental peptide masses, but the report tells us that these matches are not statistically significant. The score threshold for this search is 76, and the top scoring match is 47. 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

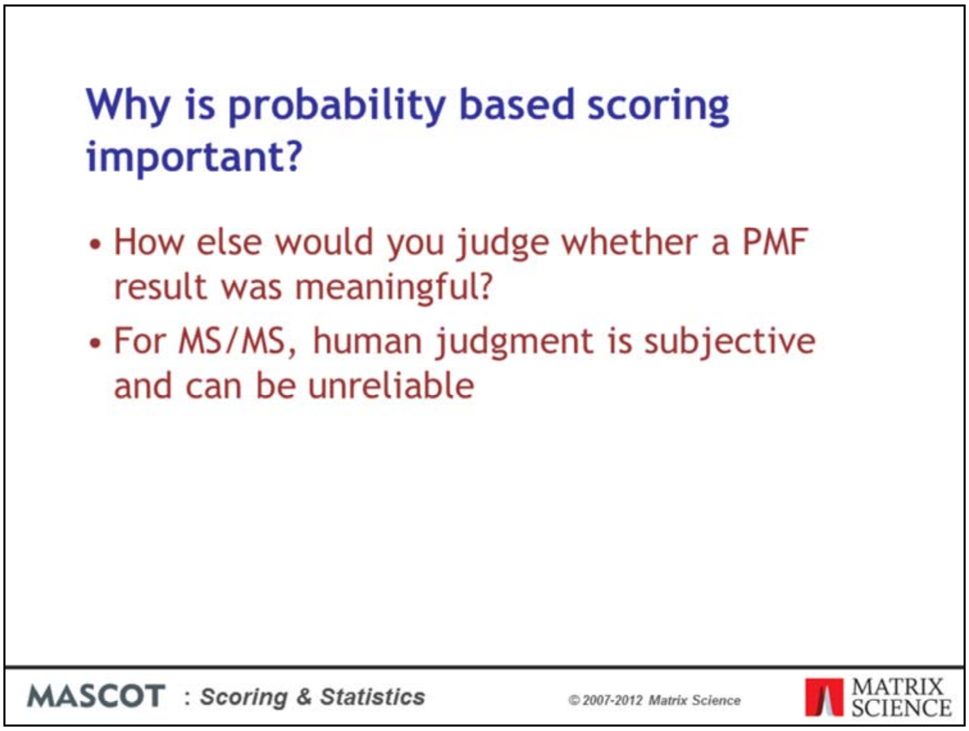
MASCOT : *Scoring & Statistics* © 2007-2012 *Matrix Science* MATRIX SCIENCE

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

MASCOT : *Scoring & Statistics*

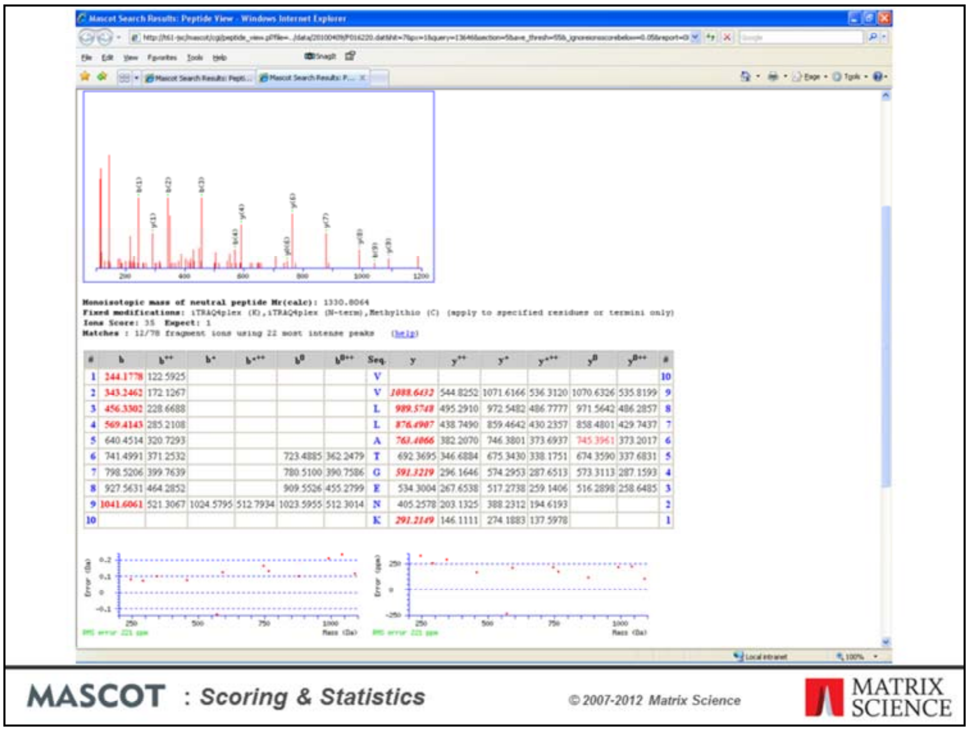
© 2007-2012 *Matrix* Science

MATRIX SCIENCE

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



Monoisotopic mass of neutral peptide Mr(calc): 1330.8064
Fixed modifications: iTRAQ4plex (K),iTRAQ4plex (N-term),Methylthio (C) (apply to specified residues or termini only)
Ions Score: 35 **Expect:** 1
Matches : 12/78 fragment ions using 22 most intense peaks (help)

#	b	b^{++}	b^{*}	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	244.1778	122.5925					V							10
2	343.2462	172.1267					V	1088.6432	544.8252	1071.6166	536.3120	1070.6326	535.8199	9
3	456.3302	228.6688					L	989.5748	495.2910	972.5482	486.7777	971.5642	486.2857	8
4	569.4143	285.2108					L	876.4907	438.7490	859.4642	430.2357	858.4801	429.7437	7
5	640.4514	320.7293					A	763.4066	382.2070	746.3801	373.6937	745.3961	373.2017	6
6	741.4991	371.2532			723.4885	362.2479	T	692.3695	346.6884	675.3430	338.1751	674.3590	337.6831	5
7	798.5206	399.7639			780.5100	390.7586	G	591.3219	296.1646	574.2953	287.6513	573.3113	287.1593	4
8	927.5631	464.2852			909.5526	455.2799	E	534.3004	267.6538	517.2738	259.1406	516.2898	258.6485	3
9	1041.6061	521.3067	1024.5795	512.7934	1023.5955	512.3014	N	405.2578	203.1325	388.2312	194.6193			2
10							K	291.2149	146.1111	274.1883	137.5978			1

MASCOT : *Scoring & Statistics*

© 2007-2012 Matrix Science

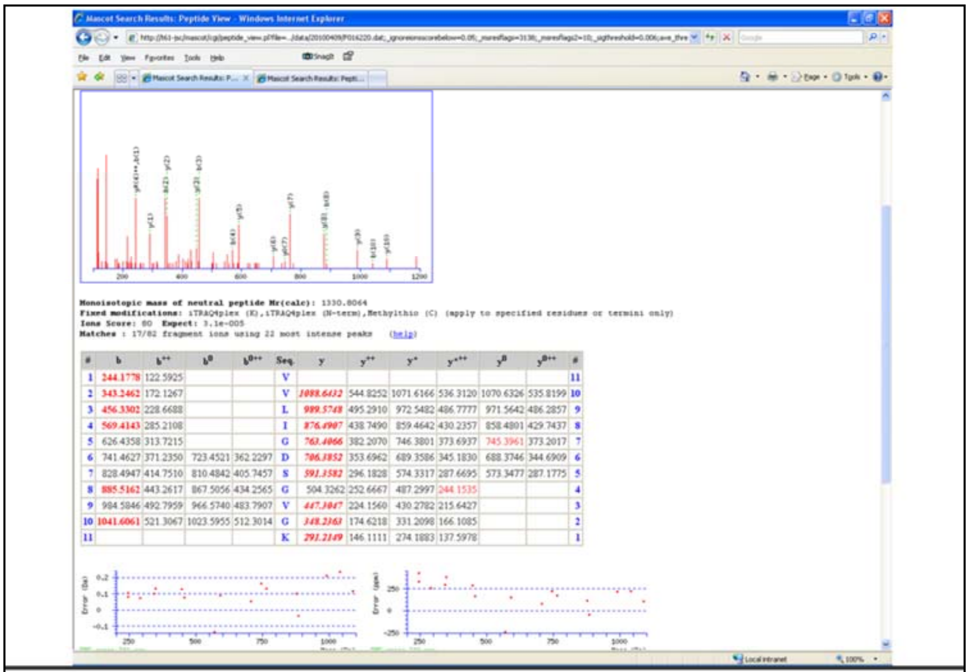
This match has a good number of matches to y and b ions, highlighted in red. All the major peaks above 200 Da seem to be labelled. Could such a good match have occurred by chance?

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 1. 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 View - Windows Internet Explorer
File Edit View Favorites Tools Help
Mascot Search Results: P...
Mascot Search Results: Pepti...
200
400
600
800
1000
1200
Error (Da)
Error (ppm)
250
500
750
1000

Monoisotopic mass of neutral peptide Mr(calc): 1330.8064
Fixed modifications: iTRAQ4plex (K),iTRAQ4plex (N-term),Methylthio (C) (apply to specified residues or termini only)
Ions Score: 80 **Expect:** 3.1e-005
Matches : 17/82 fragment ions using 22 most intense peaks (help)

#	b	b⁺⁺	b⁰	b⁰⁺⁺	Seq.	y	y⁺⁺	y*	y*⁺⁺	y⁰	y⁰⁺⁺	#
1	244.1778	122.5925			V							11
2	343.2462	172.1267			V	1088.6432	544.8252	1071.6166	536.3120	1070.6326	535.8199	10
3	456.3302	228.6688			L	989.5748	495.2910	972.5482	486.7777	971.5642	486.2857	9
4	569.4143	285.2108			I	876.4907	438.7490	859.4642	430.2357	858.4801	429.7437	8
5	626.4358	313.7215			G	763.4066	382.2070	746.3801	373.6937	745.3961	373.2017	7
6	741.4627	371.2350	723.4521	362.2297	D	706.3852	353.6962	689.3586	345.1830	688.3746	344.6909	6
7	828.4947	414.7510	810.4842	405.7457	S	591.3582	296.1828	574.3317	287.6695	573.3477	287.1775	5
8	885.5162	443.2617	867.5056	434.2565	G	504.3262	252.6667	487.2997	244.1535			4
9	984.5846	492.7959	966.5740	483.7907	V	447.3047	224.1560	430.2782	215.6427			3
10	1041.6061	521.3067	1023.5955	512.3014	G	348.2363	174.6218	331.2098	166.1085			2
11					K	291.2149	146.1111	274.1883	137.5978			1

MASCOT : *Scoring & Statistics*

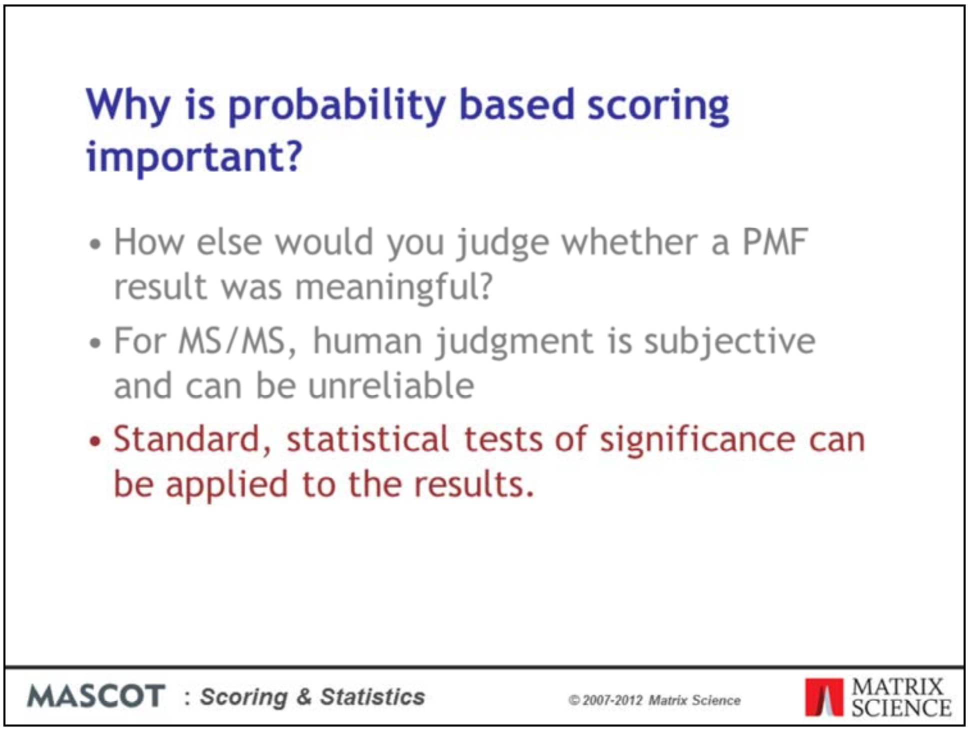
© 2007-2012 *Matrix Science*

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 10,000. 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 last y ion or some more of the b ions fall into line?



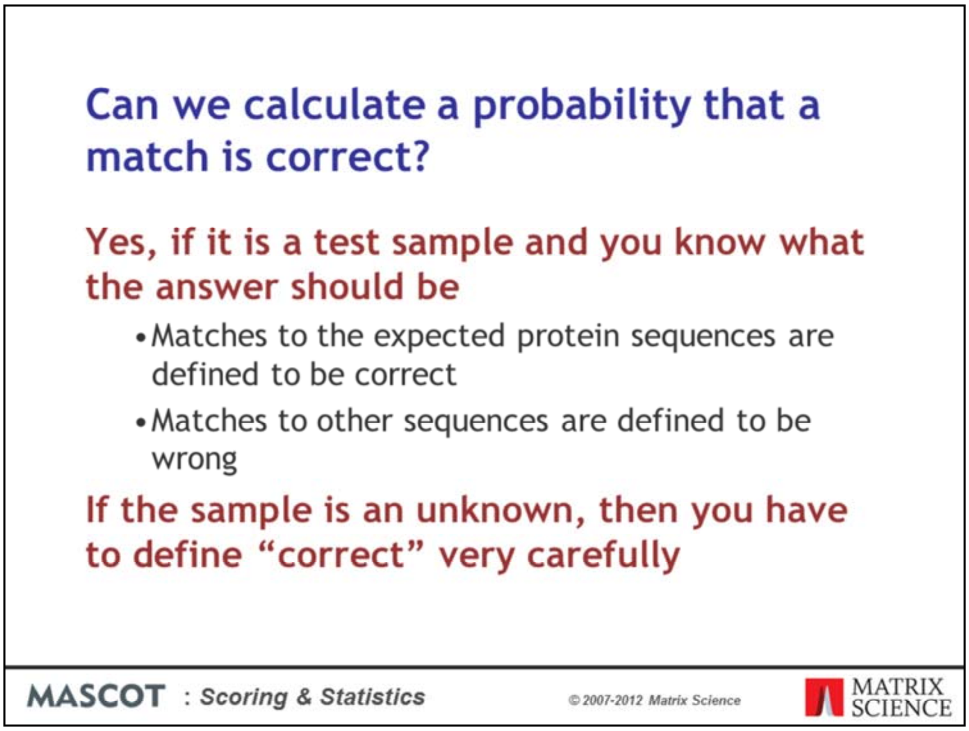
Why is probability based scoring important?

- 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.

MASCOT : *Scoring & Statistics* © 2007-2012 *Matrix Science* MATRIX SCIENCE

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



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

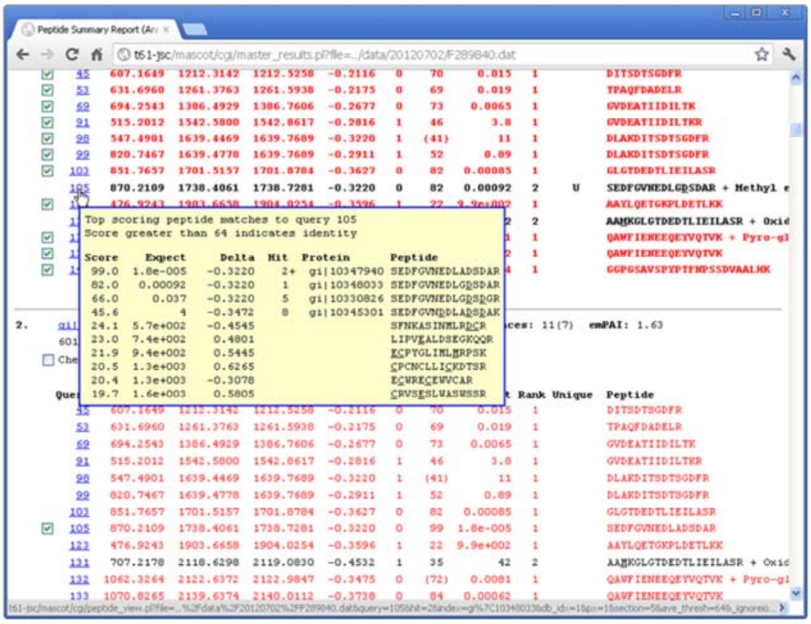
- Matches to the expected protein sequences are defined to be correct
- Matches to other sequences are defined to be wrong

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

MASCOT : *Scoring & Statistics* © 2007-2012 Matrix Science MATRIX SCIENCE

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.



Peptide Summary Report (Ar
t61-jsc/mascot/cgi/master_results.pl?file=../data/20120702/F289840.dat

```
☑  45   607.1649  1212.3142  1212.5258  -0.2116  0   70     0.015   1      DITSDTSGDFR
☑  53   631.6960  1261.3763  1261.5938  -0.2175  0   69     0.019   1      TPAQFDADELR
☑  69   694.2543  1386.4929  1386.7606  -0.2677  0   73     0.0065  1      GVDEATIIDILTK
☑  91   515.2012  1542.5800  1542.8617  -0.2816  1   46       3.8   1      GVDEATIIDILTKR
☑  98   547.4901  1639.4469  1639.7689  -0.3220  1  (41)       11   1      DLAKDITSDTSGDFR
☑  99   820.7467  1639.4778  1639.7689  -0.2911  1   52      0.89   1      DLAKDITSDTSGDFR
☑  103  851.7657  1701.5157  1701.8784  -0.3627  0   82   0.00085   1      GLGTDEDTLIEILASR
   105  870.2109  1738.4061  1738.7281  -0.3220  0   82   0.00092   2   U  SEDFGVNEDLGDSDAR + Methyl e
☑  1    476.9243  1903.6658  1904.0254  -0.3596  1   22  9.9e+002   1      AAYLQETGKPLDETLKK
   1                                                             2   2      AAMKGLGTDEDTLIEILASR + Oxid
☑  1                                                             1   1      QAWFIENEEQEYVQTVK + Pyro-gl
☑  1                                                             2   1      QAWFIENEEQEYVQTVK
☑  1                                                             4   1      GGPGSAVSPYPTFNPSSDVAALHK
```

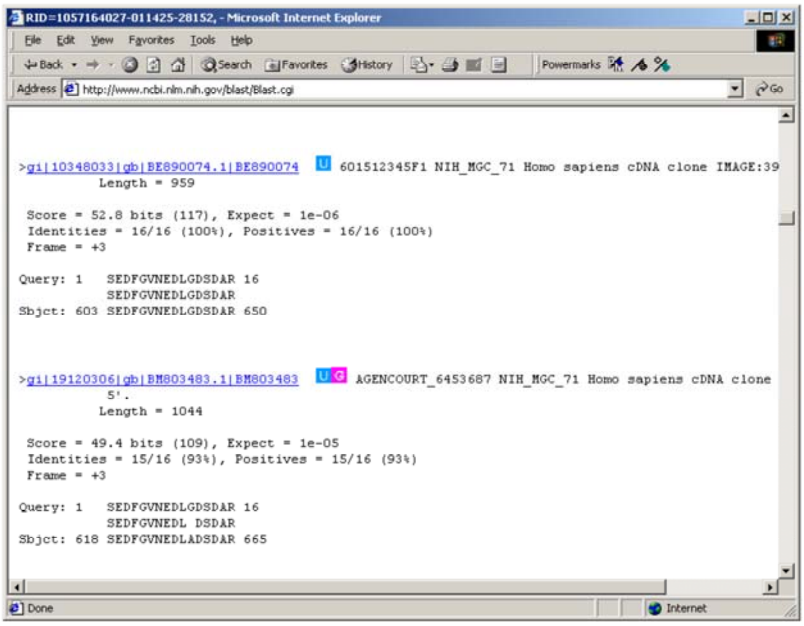
```
Top scoring peptide matches to query 105
Score greater than 64 indicates identity

Score    Expect     Delta  Hit  Protein        Peptide
 99.0  1.8e-005   -0.3220   2+  gi|10347940  SEDFGVNEDLADSDAR
 82.0   0.00092   -0.3220   1   gi|10348033  SEDFGVNEDLGDSDAR
 66.0     0.037   -0.3220   5   gi|10330826  SEDFGVNEDLGDSDGR
 45.6         4   -0.3472   8   gi|10345301  SEDFGVNDDLADSDAK
 24.1  5.7e+002   -0.4545                    SFNKASINMLRDCR
 23.0  7.4e+002    0.4801                    LIPVEALDSEGKQQR
 21.9  9.4e+002    0.5445                    ECPYGLIMLMRPSK
 20.5  1.3e+003    0.6265                    CPCNCLLICKDTSR
 20.4  1.3e+003   -0.3078                    ECWRECEWVCAR
 19.7  1.6e+003    0.5805                    CRVSESLWASWSSR
```

```
2.  gi|                                               ces: 11(7)   emPAI: 1.63
    601
    □ Che

    Que                                                     t  Rank Unique  Peptide
    45   607.1649  1212.3142  1212.5258  -0.2116  0   70     0.015   1      DITSDTSGDFR
    53   631.6960  1261.3763  1261.5938  -0.2175  0   69     0.019   1      TPAQFDADELR
    69   694.2543  1386.4929  1386.7606  -0.2677  0   73    0.0065   1      GVDEATIIDILTK
    91   515.2012  1542.5800  1542.8617  -0.2816  1   46       3.8   1      GVDEATIIDILTKR
    98   547.4901  1639.4469  1639.7689  -0.3220  1  (41)       11   1      DLAKDITSDTSGDFR
    99   820.7467  1639.4778  1639.7689  -0.2911  1   52      0.89   1      DLAKDITSDTSGDFR
    103  851.7657  1701.5157  1701.8784  -0.3627  0   82   0.00085   1      GLGTDEDTLIEILASR
☑   105  870.2109  1738.4061  1738.7281  -0.3220  0   99  1.8e-005   1      SEDFGVNEDLADSDAR
    123  476.9243  1903.6658  1904.0254  -0.3596  1   22  9.9e+002   1      AAYLQETGKPLDETLKK
    131  707.2178  2118.6298  2119.0830  -0.4532  1   35        42   2      AAMKGLGTDEDTLIEILASR + Oxid
    132 1062.3264  2122.6372  2122.9847  -0.3475  0  (72)   0.0081   1      QAWFIENEEQEYVQTVK + Pyro-gl
    133 1070.8265  2139.6374  2140.0112  -0.3738  0   84   0.00062   1      QAWFIENEEQEYVQTVK
```

t61-jsc/mascot/cgi/peptide_view.pl?file=...%2Fdata%2F20120702%2FF289840.dat&query=105&hit=2&index=gi%7C10348033&db_idx=1&px=1&section=5&ave_thresh=64&_ignoreio...

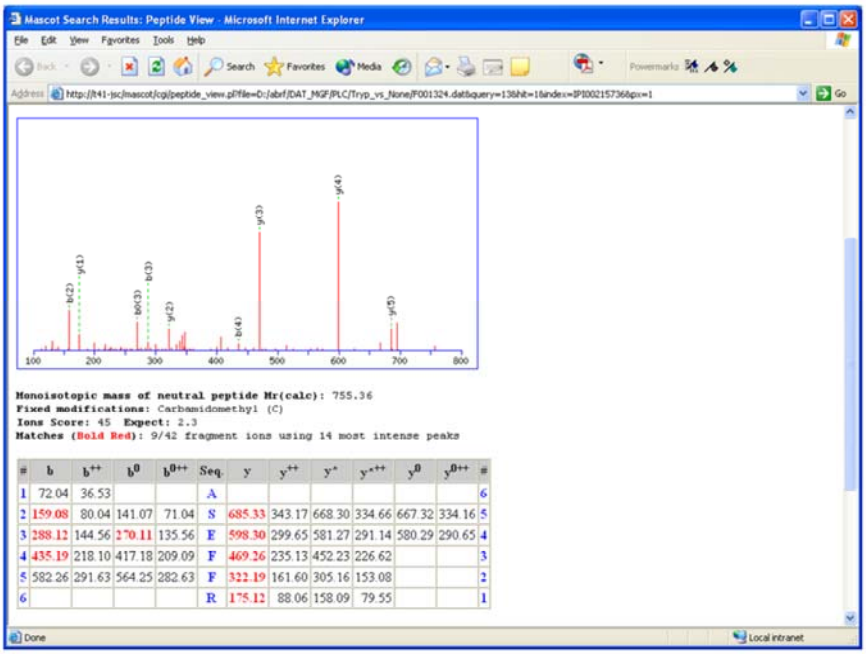
This is a typical MS/MS search result, where we see a series of high scoring homologous peptides. The sequences of the top four matches are very similar, and their expectation values vary from random through to very unlikely to be random. The best match has an expectation value of 2E-5. However, we cannot 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.


RID=1057164027-011425-28152, - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Powermarks
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
Done
Internet

```
>gi|10348033|gb|BE890074.1|BE890074  U 601512345F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:39
          Length = 959

 Score = 52.8 bits (117), Expect = 1e-06
 Identities = 16/16 (100%), Positives = 16/16 (100%)
 Frame = +3

Query: 1   SEDFGVNEDLGDSDAR 16
            SEDFGVNEDLGDSDAR
Sbjct: 603 SEDFGVNEDLGDSDAR 650


>gi|19120306|gb|BM803483.1|BM803483  UG AGENCOURT_6453687 NIH_MGC_71 Homo sapiens cDNA clone
            5'.
          Length = 1044

 Score = 49.4 bits (109), Expect = 1e-05
 Identities = 15/16 (93%), Positives = 15/16 (93%)
 Frame = +3

Query: 1   SEDFGVNEDLGDSDAR 16
            SEDFGVNEDL DSDAR
Sbjct: 618 SEDFGVNEDLADSDAR 665
```

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 1E-6, which reminds us that it would be a random match if the database was a 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.

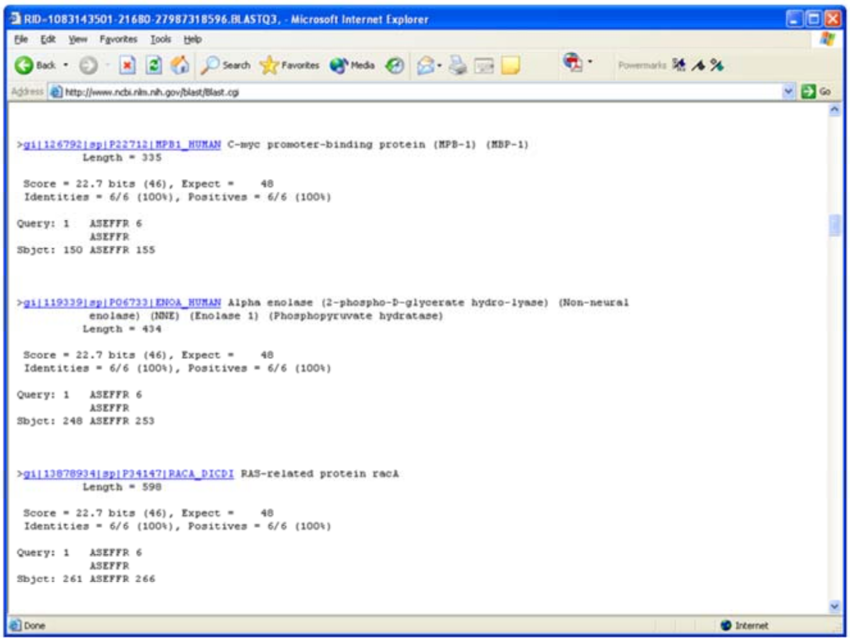

Mascot Search Results: Peptide View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Powermarks
Address http://t41-jsc/mascot/cgi/peptide_view.pl?file=D:/abrf/DAT_MGF/PLC/Tryp_vs_None/F001324.dat&query=13&hit=1&index=IPI00215736&px=1 Go
b(2)
y(1)
b0(3)
b(3)
y(2)
b(4)
y(3)
y(4)
y(5)
100
200
300
400
500
600
700
800
Done
Local intranet

Monoisotopic mass of neutral peptide Mr(calc): 755.36
Fixed modifications: Carbamidomethyl (C)
Ions Score: 45 **Expect:** 2.3
Matches (Bold Red): 9/42 fragment ions using 14 most intense peaks

#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	72.04	36.53			A							6
2	**159.08**	80.04	141.07	71.04	S	**685.33**	343.17	668.30	334.66	667.32	334.16	5
3	**288.12**	144.56	**270.11**	135.56	E	**598.30**	299.65	581.27	291.14	580.29	290.65	4
4	**435.19**	218.10	417.18	209.09	F	**469.26**	235.13	452.23	226.62			3
5	582.26	291.63	564.25	282.63	F	**322.19**	161.60	305.16	153.08			2
6					R	**175.12**	88.06	158.09	79.55			1

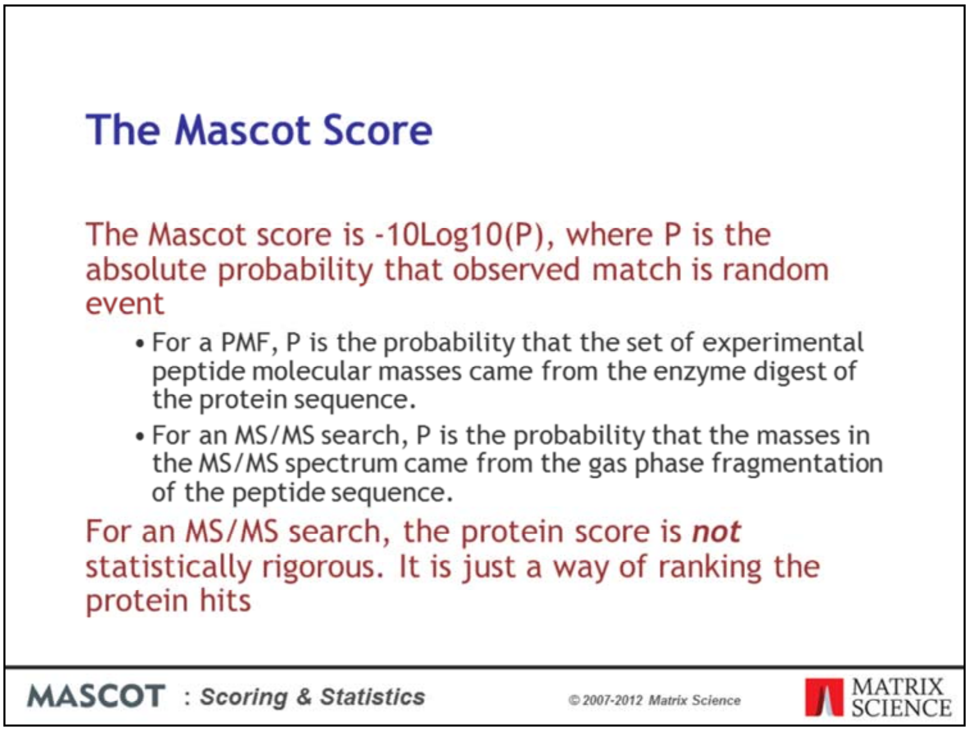
If we are doing probability based matching, we are not scoring the quality of the spectrum, we are scoring whether the match is random or not.

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.


RID-1083143501-21680-27987318596.BLASTQ3, - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Powermarks
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
Done
Internet

```
>gi|126792|sp|P22712|MPB1_HUMAN C-myc promoter-binding protein (MPB-1) (MBP-1)
          Length = 335

 Score = 22.7 bits (46), Expect =    48
 Identities = 6/6 (100%), Positives = 6/6 (100%)

Query: 1   ASEFFR 6
           ASEFFR
Sbjct: 150 ASEFFR 155


>gi|119339|sp|P06733|ENOA_HUMAN Alpha enolase (2-phospho-D-glycerate hydro-lyase) (Non-neural
           enolase) (NNE) (Enolase 1) (Phosphopyruvate hydratase)
          Length = 434

 Score = 22.7 bits (46), Expect =    48
 Identities = 6/6 (100%), Positives = 6/6 (100%)

Query: 1   ASEFFR 6
           ASEFFR
Sbjct: 248 ASEFFR 253


>gi|13878934|sp|P34147|RACA_DICDI RAS-related protein racA
          Length = 598

 Score = 22.7 bits (46), Expect =    48
 Identities = 6/6 (100%), Positives = 6/6 (100%)

Query: 1   ASEFFR 6
           ASEFFR
Sbjct: 261 ASEFFR 266
```

The situation in a Blast search is identical. Even though this is a perfect identity match, the expectation value is 48. 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.
- 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.

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

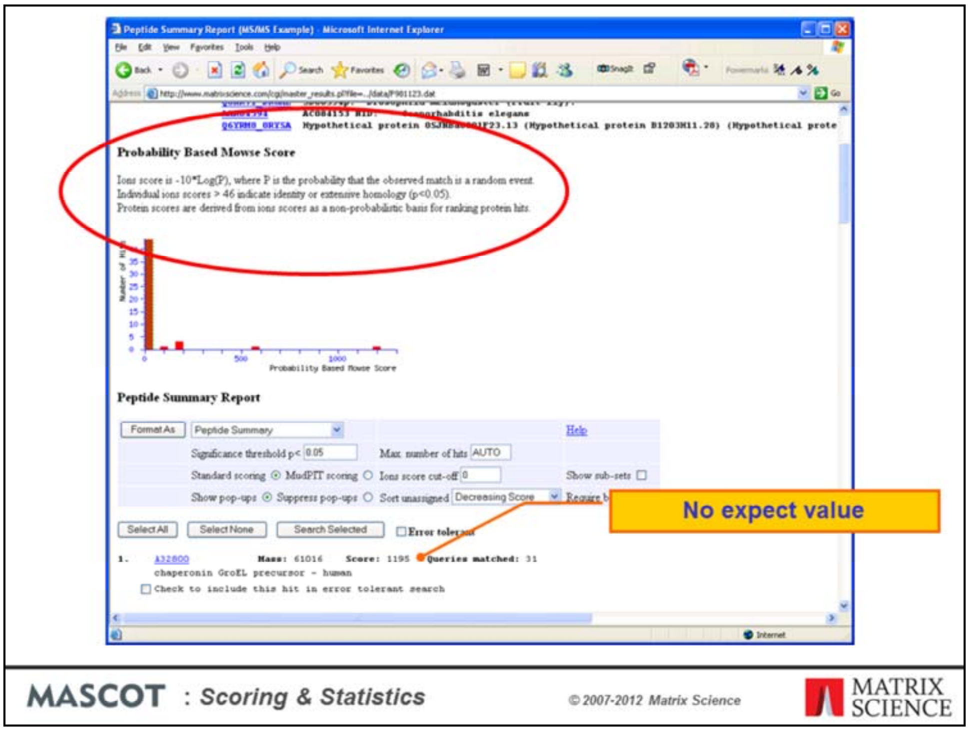
MASCOT : *Scoring & Statistics* © 2007-2012 *Matrix Science* MATRIX SCIENCE

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 random.

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 Summary Report (MS/MS Example) - Microsoft Internet Explorer

Address http://www.matrixscience.com/cgi/master_results.pl?file=../data/F981123.dat

```
ACO84153 HID:     Caenorhabditis elegans
Q6YRM8_ORYSA   Hypothetical protein OSJNBa0001F23.13 (Hypothetical protein B1203H11.20) (Hypothetical prote
```

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 46 indicate identity or extensive homology (p<0.05).
Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.

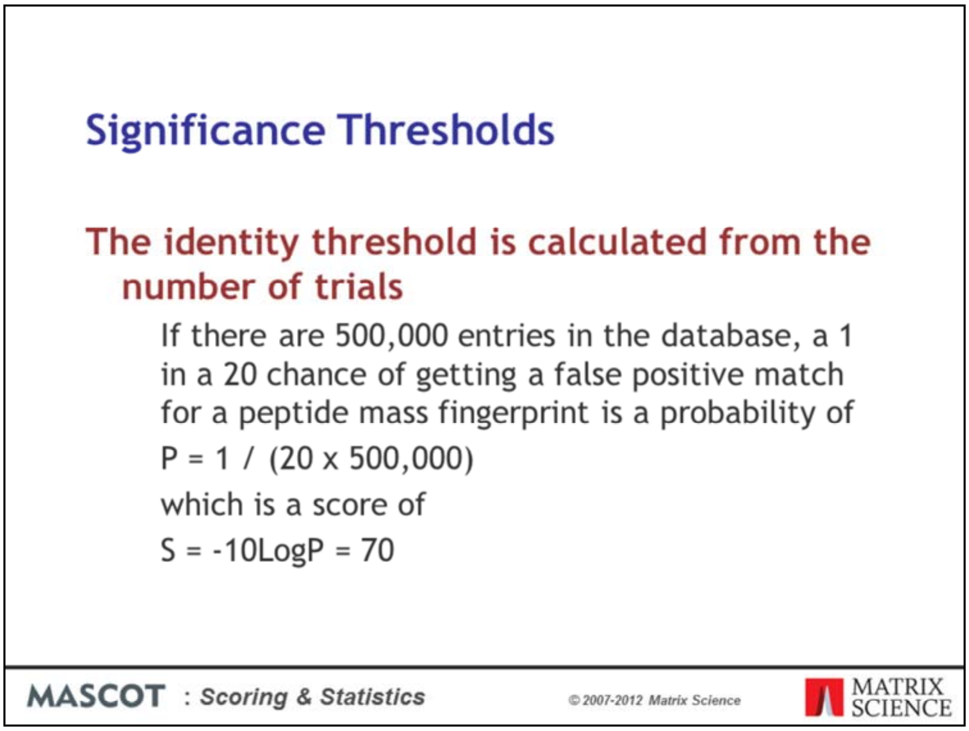
Peptide Summary Report

Format As: Peptide Summary
Significance threshold p< 0.05
Max. number of hits AUTO
Standard scoring ● MudPIT scoring ○
Ions score cut-off 0
Show sub-sets □
Show pop-ups ● Suppress pop-ups ○
Sort unassigned Decreasing Score
Require b
□ Error tolerant

```
1.    A32800          Mass: 61016    Score: 1195    Queries matched: 31
      chaperonin GroEL precursor - human
   □ Check to include this hit in error tolerant search
```

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

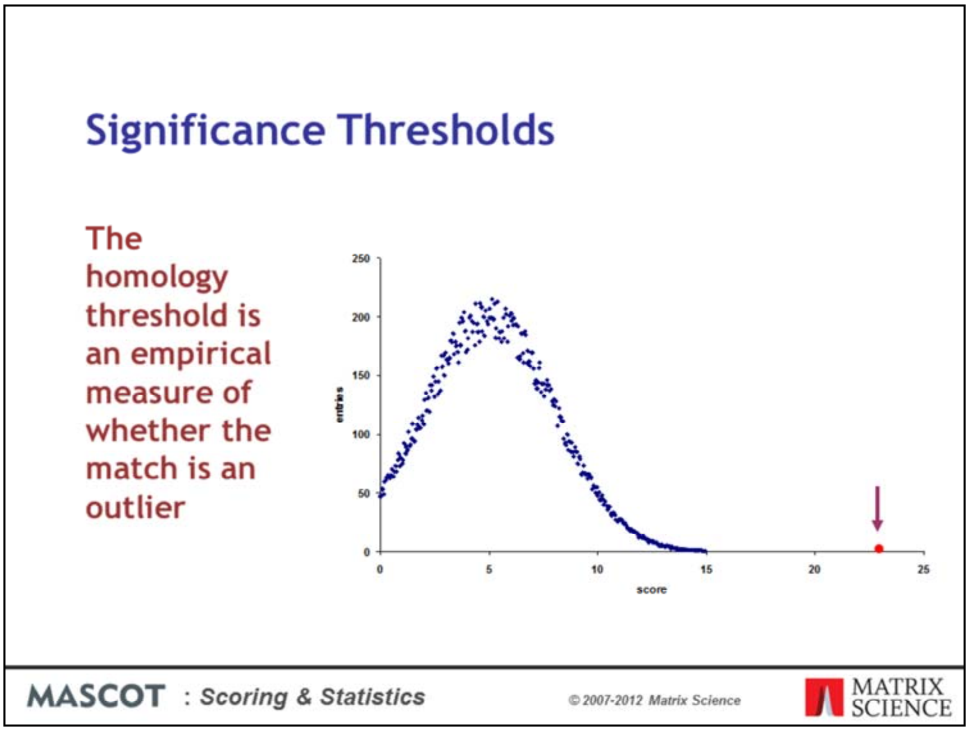
MASCOT : *Scoring & Statistics* © 2007-2012 *Matrix Science* MATRIX SCIENCE

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*

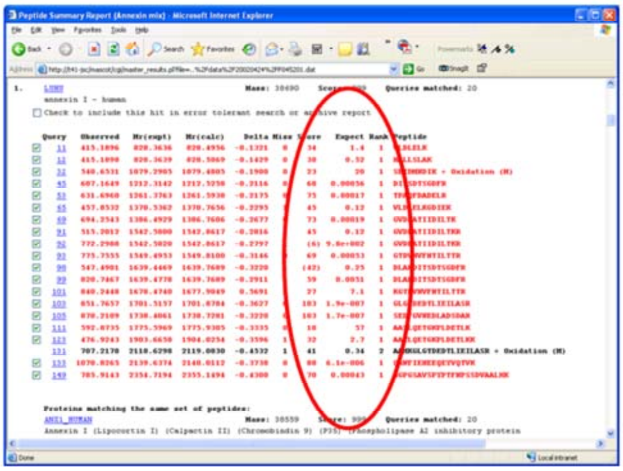
© 2007-2012 Matrix Science

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.

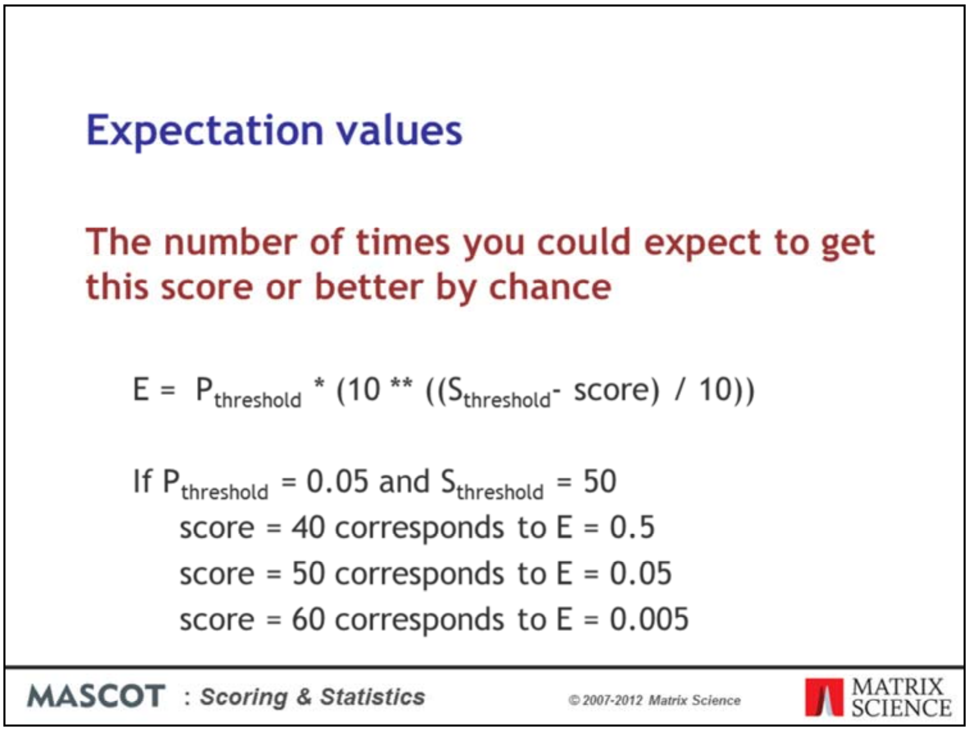
Expectation values



```
1.    LIHU                                  Mass: 38690    Score: 909    Queries matched: 20
      annexin I - human
  □ Check to include this hit in error tolerant search or archive report

      Query   Observed   Mr(expt)   Mr(calc)    Delta Miss Score  Expect Rank Peptide
  ☑    11    415.1896   828.3636   828.4956   -0.1321   0    34      1.4    1  LDLELR
  ☑    12    415.1890   828.3639   828.5069   -0.1429   0    30     0.32    1  NALLSLAK
  ☑    32    540.6531  1079.2905  1079.4805   -0.1900   0    23       20    1  SEIDMNDIK + Oxidation (M)
  ☑    45    607.1649  1212.3142  1212.5258   -0.2116   0    68  0.00056    1  DITSDTSGDFR
  ☑    53    631.6960  1261.3763  1261.5938   -0.2175   0    75  0.00017    1  TPAQFDADELR
  ☑    65    457.8532  1370.5362  1370.7656   -0.2295   0    45     0.12    1  VLDLELKGDIEK
  ☑    69    694.2543  1386.4929  1386.7606   -0.2677   0    73  0.00019    1  GVDEATIIDILTK
  ☑    91    515.2012  1542.5800  1542.8617   -0.2816   0    45     0.12    1  GVDEATIIDILTKR
  ☑    92    772.2988  1542.5820  1542.8617   -0.2797   0   (6) 9.8e+002   1  GVDEATIIDILTKR
  ☑    93    775.7555  1549.4953  1549.8100   -0.3146   0    69  0.00053    1  GTDVNVFNTILTTR
  ☑    98    547.4901  1639.4469  1639.7689   -0.3220   0  (42)     0.25    1  DLAKDITSDTSGDFR
  ☑    99    820.7467  1639.4778  1639.7689   -0.2911   0    59   0.0051    1  DLAKDITSDTSGDFR
  ☑   101    840.2448  1678.4740  1677.9049    0.5691   0    27      7.1    1  KGTDVNVFNTILTTR
  ☑   103    851.7657  1701.5157  1701.8784   -0.3627   0   103 1.9e-007    1  GLGTDEDTLIEILASR
  ☑   105    870.2109  1738.4061  1738.7281   -0.3220   0   103 1.7e-007    1  SEIDMNDIKAFYQK
  ☑   111    592.8735  1775.5969  1775.9305   -0.3335   0    18       57    1  AAYLQETGKPLDETLK
  ☑   123    476.9243  1903.6658  1904.0254   -0.3596   1    32      2.7    1  AAYLQETGKPLDETLKK
      131    707.2178  2118.6298  2119.0830   -0.4532   1    41     0.34    2  AMKGLGTDEDTLIEILASR + Oxidation (M)
  ☑   133   1070.8265  2139.6374  2140.0112   -0.3738   0    88 6.1e-006    1  GNWTIEHEEQETVQTVK
  ☑   149    785.9143  2354.7194  2355.1494   -0.4300   0    70  0.00043    1  GGPGSAVSPYPTFNPSSDVAALHK

      Proteins matching the same set of peptides:
      ANX1_HUMAN                            Mass: 38559    Score: 909    Queries matched: 20
      Annexin I (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhibitory protein
```

Peptide Summary Report (Annexin mix) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://141-pc/mascot/cgi/master_results.pl?file=../data%2F20020424%2FF045201.dat
Go
Done
Local intranet

In Mascot 2.0, we also started displaying 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

MASCOT : *Scoring & Statistics*

© 2007-2012 Matrix Science

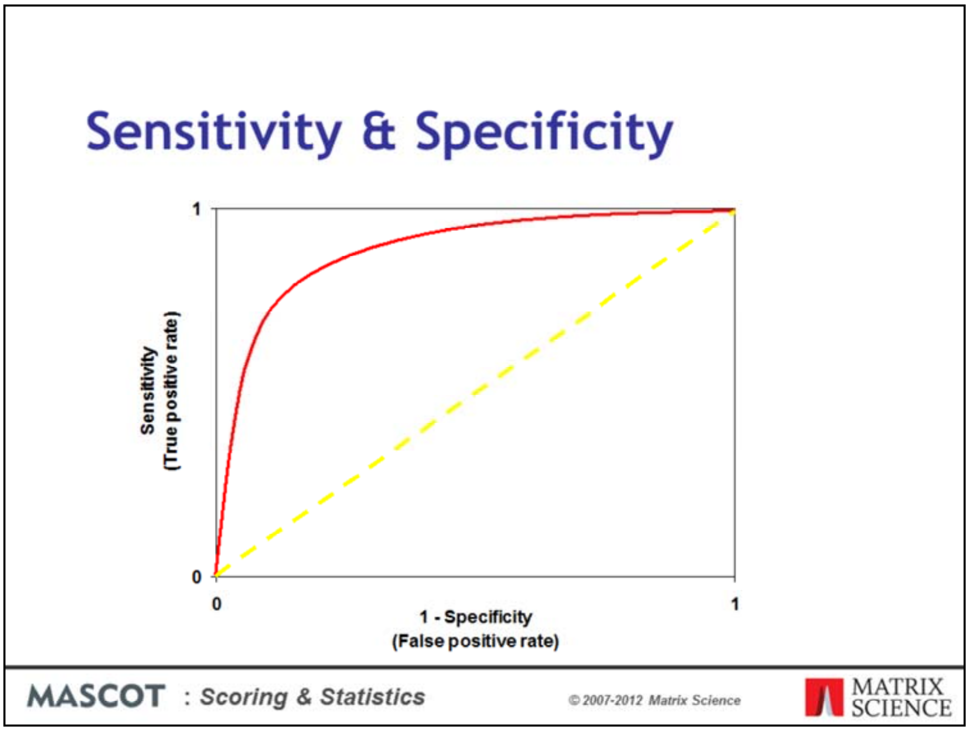
MATRIX SCIENCE

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*

© 2007-2012 Matrix Science

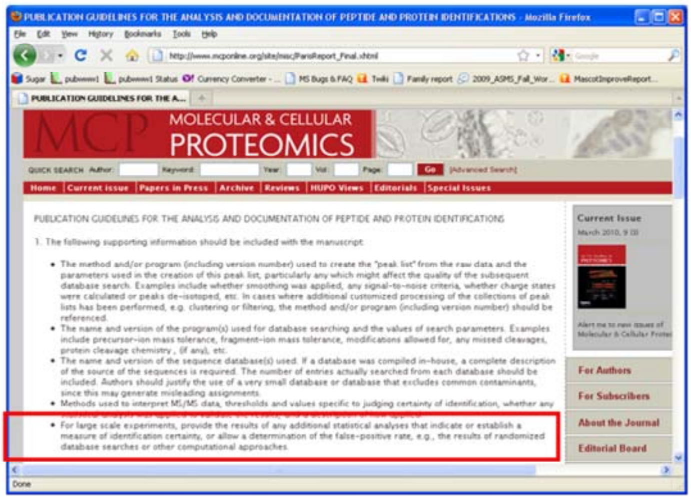
MATRIX SCIENCE

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 Prote
For Authors
For Subscribers
About the Journal
Editorial Board
Done

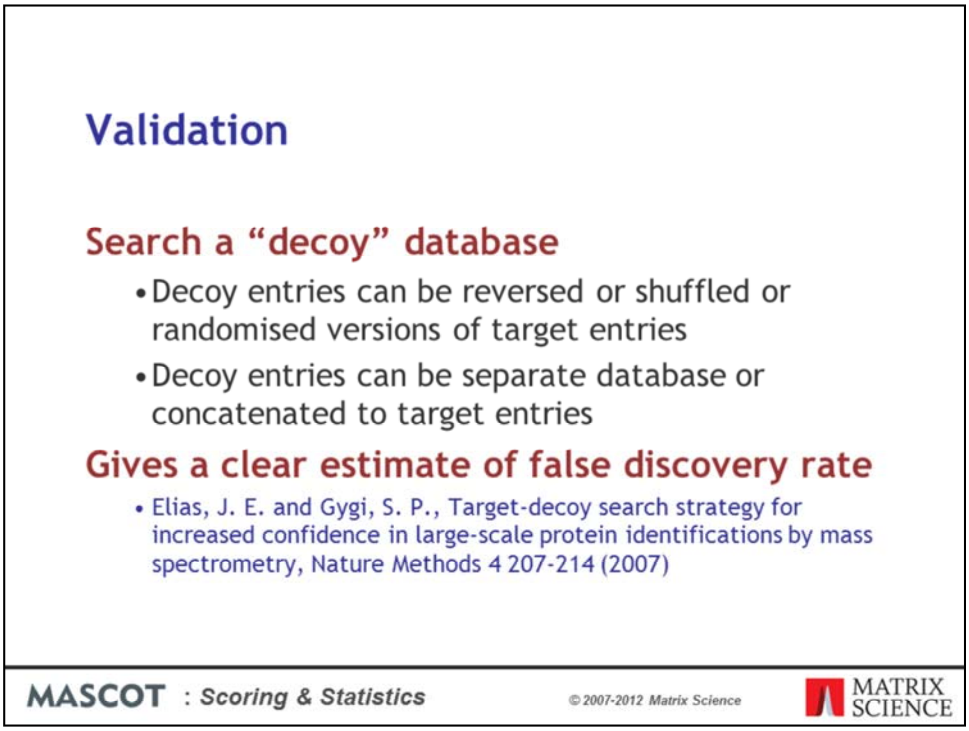
PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATIONS

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
- 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 has just recently been revised.

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)

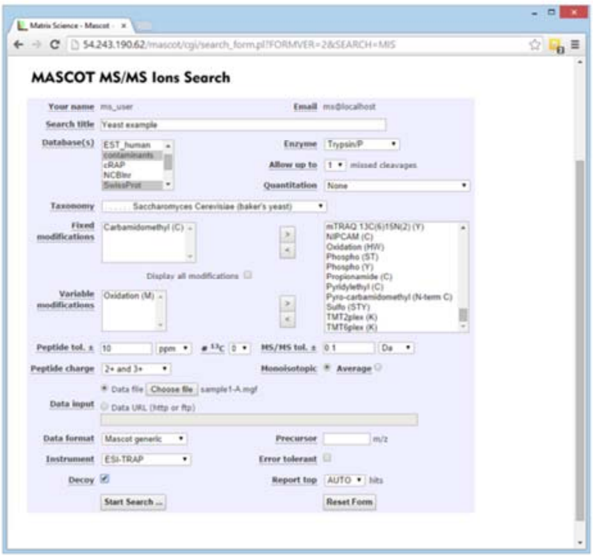
MASCOT : *Scoring & Statistics* © 2007-2012 *Matrix Science* MATRIX SCIENCE

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, 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

54.243.190.62/mascot/cgi/search_form.pl?FORMVER=2&SEARCH=MIS

MASCOT MS/MS Ions Search

Your name: ms_user

Email: ms@localhost

Search title: Yeast example

Database(s): EST_human, contaminants, cRAP, NCBInr, SwissProt

Enzyme: Trypsin/P

Allow up to: 1 missed cleavages

Quantitation: None

Taxonomy: . . . . . Saccharomyces Cerevisiae (baker's yeast)

Fixed modifications: Carbamidomethyl (C)

Display all modifications □

Variable modifications: Oxidation (M)

Peptide tol. ± 10 ppm

^{13}C 0

MS/MS tol. ± 0.1 Da

Peptide charge: 2+ and 3+

Monoisotopic ● Average ○

Data input: ● Data file Choose file sample1-A.mgf ○ Data URL (http or ftp)

Data format: Mascot generic

Precursor: m/z

Instrument: ESI-TRAP

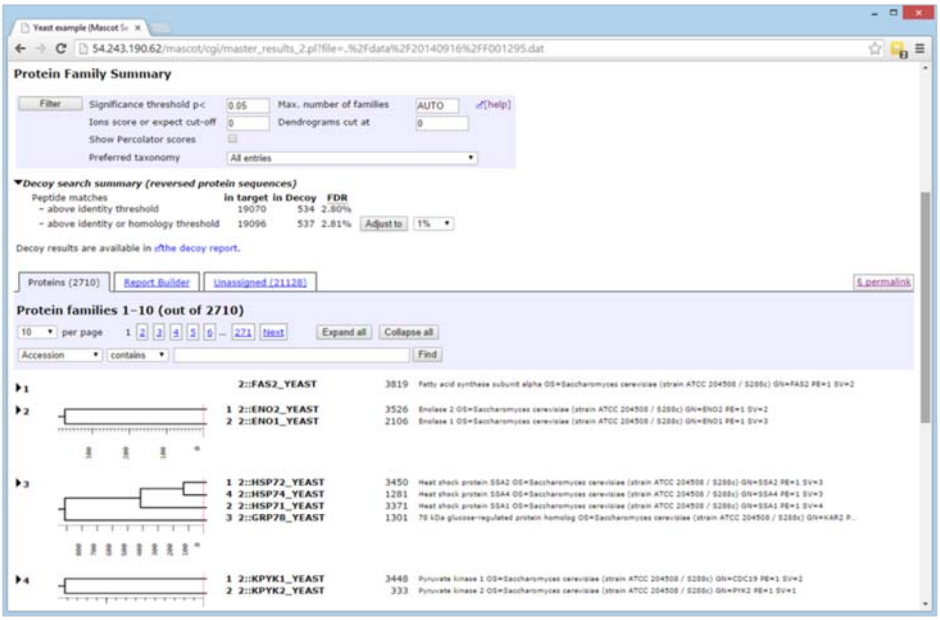
Error tolerant □

Decoy ☑

Report top AUTO hits

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. 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.

Because more and more people wish to perform decoy searches routinely, we've added this into Mascot as a built-in part of the search. 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.



54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20140916%2FF001295.dat

Protein Family Summary

Filter

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

Ions score or expect cut-off 0 Dendrograms cut at 0

Show Percolator scores □

Preferred taxonomy All entries

▼*Decoy search summary (reversed protein sequences)*

Peptide matches	in target	in Decoy	FDR
- above identity threshold	19070	534	2.80%
- above identity or homology threshold	19096	537	2.81%

Adjust to 1%

Decoy results are available in the decoy report.

Proteins (2710) | Report Builder | Unassigned (21128) | permalink

Protein families 1-10 (out of 2710)

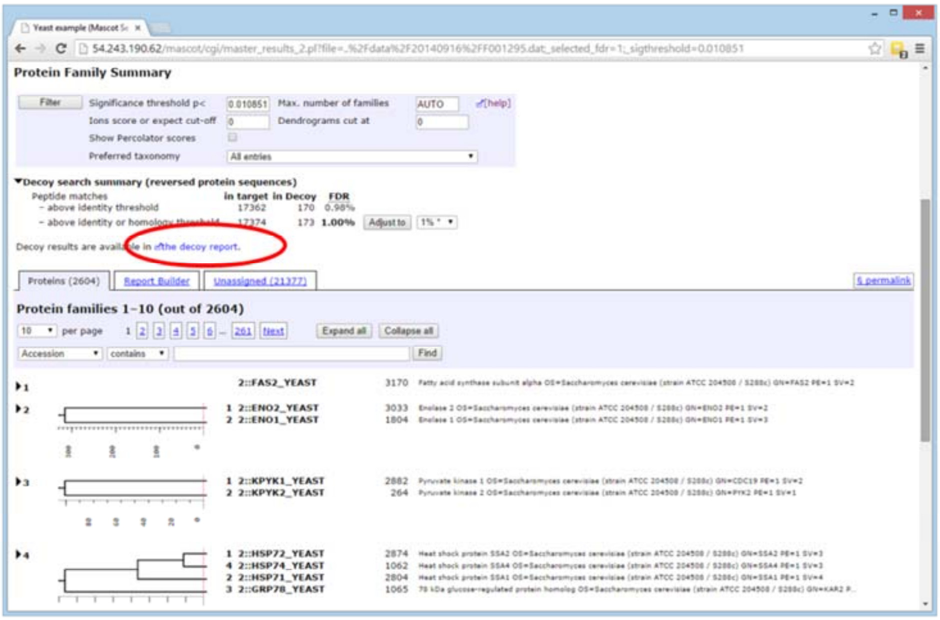
10 per page 1 2 3 4 5 6 .. 271 Next Expand all Collapse all

Accession contains Find

▶1		2::FAS2_YEAST	3819	Fatty acid synthase subunit alpha OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=FAS2 PE=1 SV=2
▶2	1	2::ENO2_YEAST	3526	Enolase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=ENO2 PE=1 SV=2
	2	2::ENO1_YEAST	2106	Enolase 1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=ENO1 PE=1 SV=3
▶3	1	2::HSP72_YEAST	3450	Heat shock protein SSA2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SSA2 PE=1 SV=3
	4	2::HSP74_YEAST	1281	Heat shock protein SSA4 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SSA4 PE=1 SV=3
	2	2::HSP71_YEAST	3371	Heat shock protein SSA1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SSA1 PE=1 SV=4
	3	2::GRP78_YEAST	1301	78 kDa glucose-regulated protein homolog OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=KAR2 P...
▶4	1	2::KPYK1_YEAST	3448	Pyruvate kinase 1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=CDC19 PE=1 SV=2
	2	2::KPYK2_YEAST	333	Pyruvate kinase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=PYK2 PE=1 SV=1

MASCOT : *Scoring & Statistics* © 2007-2012 *Matrix Science* MATRIX SCIENCE

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. In Mascot 2.4 and later, there is a button to adjust the significance threshold so as to achieve a chosen FDR value. For example, if we choose 1% FDR using the homology threshold



Yeast example (Mascot S
54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20140916%2FF001295.dat_selected_fdr=1;_sigthreshold=0.010851

Protein Family Summary

Filter

Significance threshold p< 0.010851 Max. number of families AUTO [help]

Ions score or expect cut-off 0 Dendrograms cut at 0

Show Percolator scores □

Preferred taxonomy All entries

▾Decoy search summary (reversed protein sequences)

Peptide matches	in target	in Decoy	FDR
– above identity threshold	17362	170	0.98%
– above identity or homology threshold	17374	173	**1.00%**

Decoy results are available in the decoy report.

Protein families 1–10 (out of 2604)

▸1		2::FAS2_YEAST	3170	Fatty acid synthase subunit alpha OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=FAS2 PE=1 SV=2
▸2	1	2::ENO2_YEAST	3033	Enolase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=ENO2 PE=1 SV=2
	2	2::ENO1_YEAST	1804	Enolase 1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=ENO1 PE=1 SV=3
▸3	1	2::KPYK1_YEAST	2882	Pyruvate kinase 1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=CDC19 PE=1 SV=2
	2	2::KPYK2_YEAST	264	Pyruvate kinase 2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=PYK2 PE=1 SV=1
▸4	1	2::HSP72_YEAST	2874	Heat shock protein SSA2 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SSA2 PE=1 SV=3
	4	2::HSP74_YEAST	1062	Heat shock protein SSA4 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SSA4 PE=1 SV=3
	2	2::HSP71_YEAST	2804	Heat shock protein SSA1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=SSA1 PE=1 SV=4
	3	2::GRP78_YEAST	1065	78 kDa glucose-regulated protein homolog OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) GN=KAR2 P...

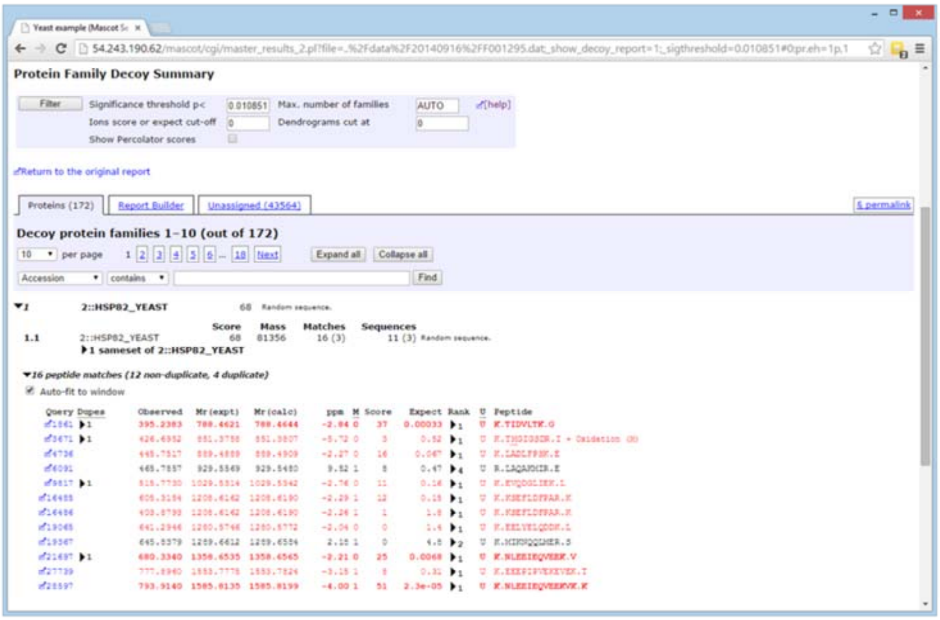
MASCOT : *Scoring & Statistics*

© 2007-2012 Matrix Science

MATRIX SCIENCE

The significance threshold has been automatically adjusted from 0.05 to 0.01.

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 link here, then you will see the results from searching the randomised database.



Yeast example (Mascot S

54.243.190.62/mascot/cgi/master_results_2.pl?file=..%2Fdata%2F20140916%2FF001295.dat;_show_decoy_report=1;_sigthreshold=0.010851#0:pr.eh=1p.1

Protein Family Decoy Summary

Filter | Significance threshold p< 0.010851 | Max. number of families AUTO | [help]

Ions score or expect cut-off 0 | Dendrograms cut at 0

Show Percolator scores □

Return to the original report

Proteins (172) | Report Builder | Unassigned (43564) | permalink

Decoy protein families 1–10 (out of 172)

10 per page 1 2 3 4 5 6 … 18 Next | Expand all | Collapse all

Accession | contains | Find

▼1 **2::HSP82_YEAST** 68 Random sequence.

		Score	Mass	Matches	Sequences	
1.1	2::HSP82_YEAST	68	81356	16 (3)	11 (3)	Random sequence.
	▶1 sameset of 2::HSP82_YEAST					

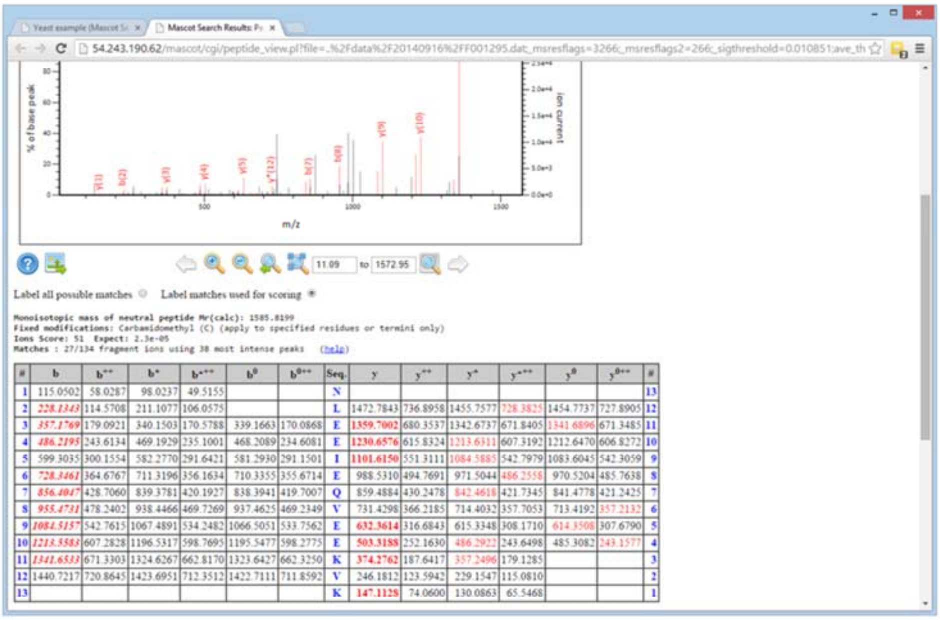
▼*16 peptide matches (12 non-duplicate, 4 duplicate)*

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
1561	▶1	395.2383	788.4621	788.4644	-2.94	0	37	0.00033	▶1	U	K.TIDVLTK.G
3671	▶1	426.6952	851.3758	851.3807	-5.72	0	3	0.52	▶1	U	K.TMDIGSSR.I + Oxidation (M)
4736		445.7517	889.4889	889.4909	-2.27	0	16	0.067	▶1	U	K.LADLFPSK.E
6091		465.7857	929.5569	929.5480	9.52	1	8	0.47	▶4	U	R.LAQAMMIR.E
9817	▶1	515.7730	1029.5314	1029.5342	-2.76	0	11	0.16	▶1	U	K.EVQDGLIEK.L
16485		605.3154	1208.6162	1208.6190	-2.29	1	12	0.15	▶1	U	K.KSEFLDFFAR.K
16486		403.8799	1208.6162	1208.6190	-2.26	1	1	1.8	▶1	U	K.KSEFLDFFAR.K
19065		641.2946	1280.5746	1280.5772	-2.04	0	0	1.4	▶1	U	K.EELYELQDDK.L
19367		645.8379	1289.6612	1289.6584	2.18	1	0	4.8	▶2	U	K.MIRPQQLMER.S
21697	▶1	680.3340	1358.6535	1358.6565	-2.21	0	25	0.0068	▶1	U	K.NLEEIEQVEEK.V
27739		777.8960	1553.7778	1553.7824	-3.15	1	8	0.31	▶1	U	K.EEEPIPVEKKVEK.T
28597		793.9140	1585.8135	1585.8199	-4.00	1	51	2.3e-05	▶1	U	K.NLEEIEQVEEKVK.K

MASCOT : *Scoring & Statistics* © 2007-2012 Matrix Science MATRIX SCIENCE

The results from the matches to the randomised sequences are saved in new sections 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%. If we click on the query number link to display the Peptide View of one of these matches …



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

MASCOT : *Scoring & Statistics*

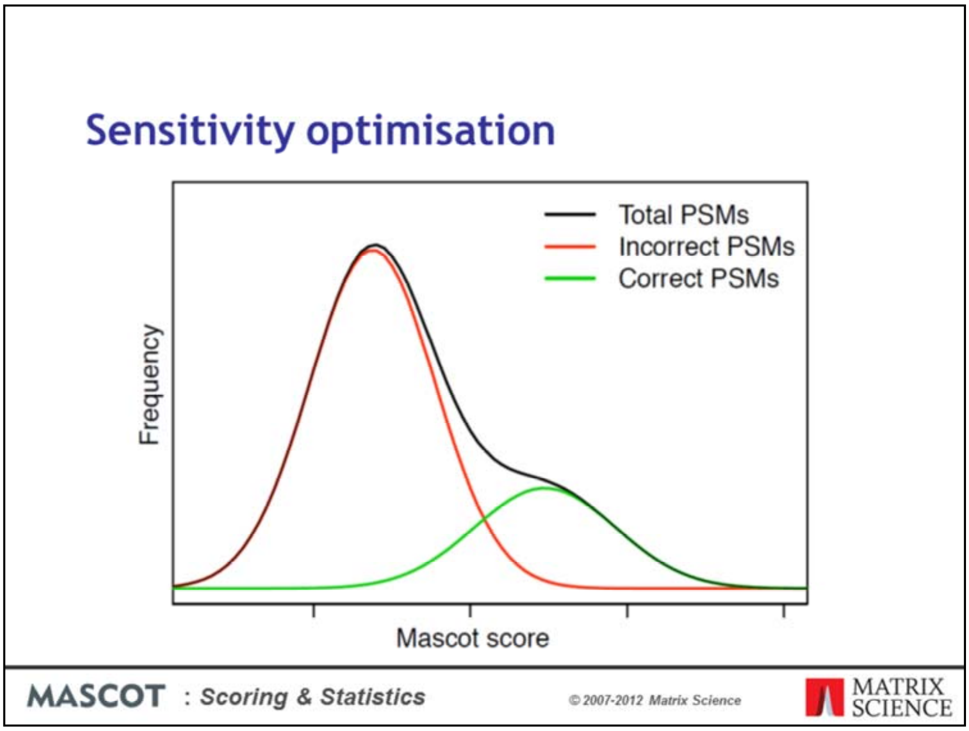
© 2007-2012 Matrix Science

MATRIX SCIENCE

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.

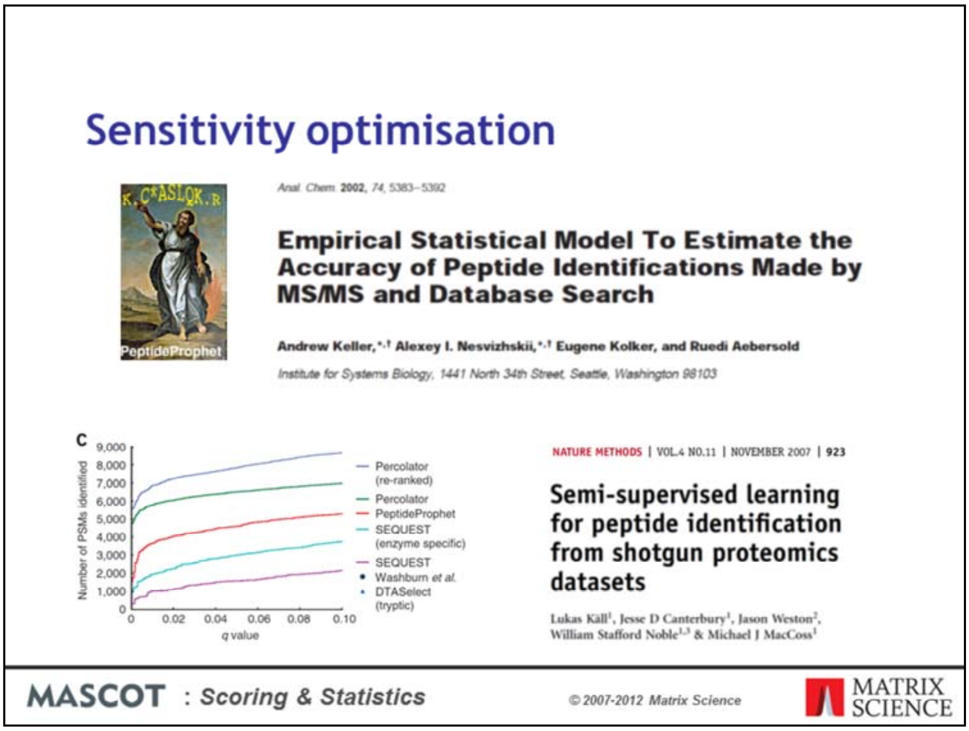


Sensitivity optimisation

MASCOT : *Scoring & Statistics* © 2007-2012 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

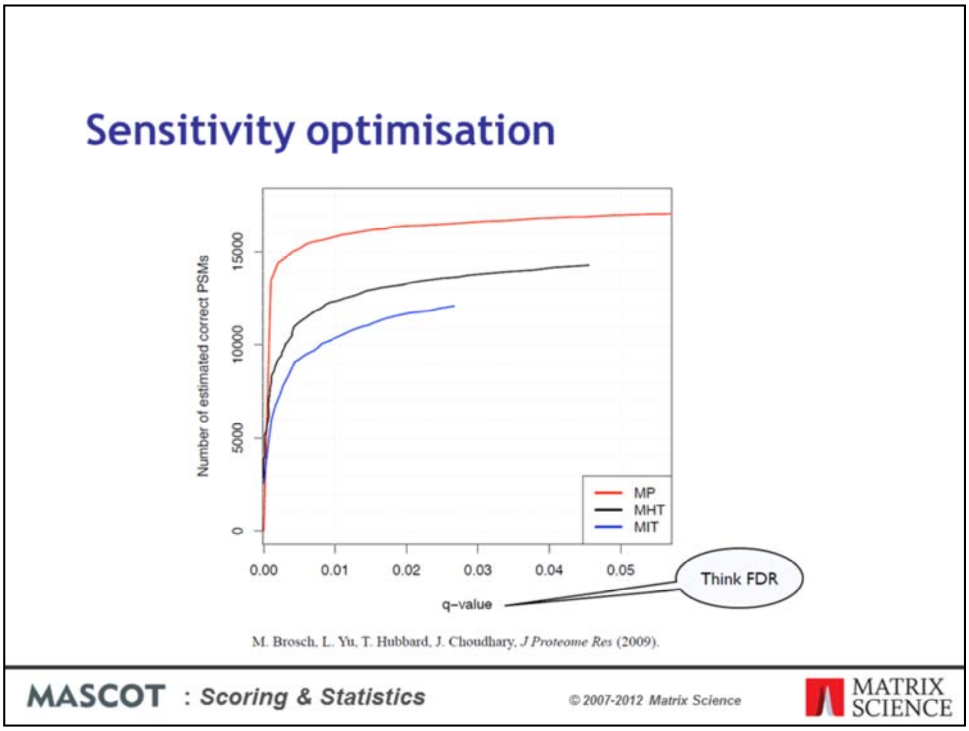
Semi-supervised learning for peptide identification from shotgun proteomics datasets

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.

A more recent development has been to use the matches from a decoy database as negative examples for the classifier. 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.



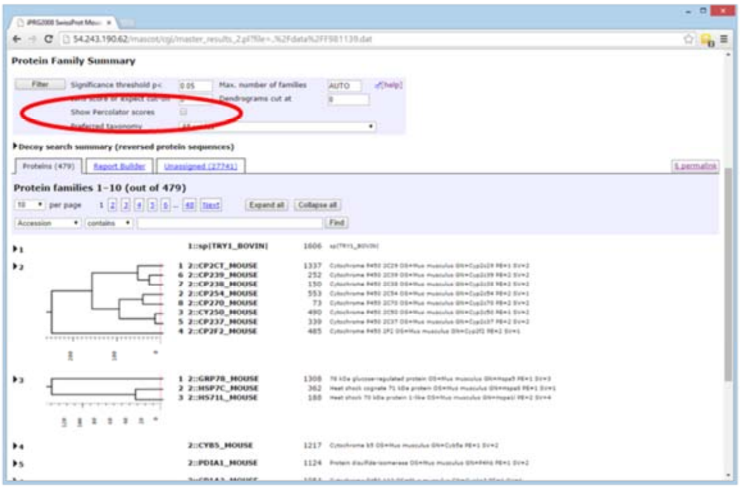
Sensitivity optimisation

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

MASCOT : *Scoring & Statistics* © 2007-2012 Matrix Science MATRIX SCIENCE

This 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



iPRG2008 SwissProt Mou: ×
54.243.190.62/mascot/cgi/master_results_2.pl?file=../%2Fdata%2FF981139.dat

Protein Family Summary

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

Ions score or expect cut-off 0 | Dendrograms cut at 0

Show Percolator scores □

Preferred taxonomy All entries

▸ Decoy search summary (reversed protein sequences)

Proteins (479) | Report Builder | Unassigned (27741) | permalink

Protein families 1–10 (out of 479)

10 per page 1 2 3 4 5 6 ... 48 Next | Expand all | Collapse all

Accession | contains | Find

▸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 252 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 553 Cytochrome P450 2C54 OS=Mus musculus GN=Cyp2c54 PE=2 SV=1
8 2::CP270_MOUSE 73 Cytochrome P450 2C70 OS=Mus musculus GN=Cyp2c70 PE=2 SV=2
3 2::CY250_MOUSE 490 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
4 2::CP2F2_MOUSE 485 Cytochrome P450 2F2 OS=Mus musculus GN=Cyp2f2 PE=2 SV=1

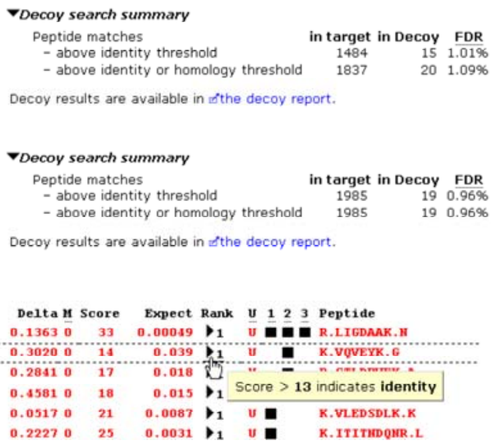
▸3
1 2::GRP78_MOUSE 1308 78 kDa glucose-regulated protein OS=Mus musculus GN=Hspa5 PE=1 SV=3
2 2::HSP7C_MOUSE 362 Heat shock cognate 71 kDa protein OS=Mus musculus GN=Hspa8 PE=1 SV=1
3 2::HS71L_MOUSE 188 Heat shock 70 kDa protein 1-like OS=Mus musculus GN=Hspa1l PE=2 SV=4

▸4 2::CYB5_MOUSE 1217 Cytochrome b5 OS=Mus musculus GN=Cyb5a PE=1 SV=2

▸5 2::PDIA1_MOUSE 1124 Protein disulfide-isomerase OS=Mus musculus GN=P4hb PE=1 SV=2

The developers of Percolator have kindly agreed to allow us to distribute and install Percolator as part of Mascot 2.3 and later. 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. This is the example search that is linked from the MS/MS Summary report help page

Sensitivity optimisation



▾*Decoy search summary*

Peptide matches	in target	in Decoy	FDR
- above identity threshold	1484	15	1.01%
- above identity or homology threshold	1837	20	1.09%

Decoy results are available in the decoy report.

▾*Decoy search summary*

Peptide matches	in target	in Decoy	FDR
- above identity threshold	1985	19	0.96%
- above identity or homology threshold	1985	19	0.96%

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]	U		■		[illegible]
0.4581	0	18	0.015	▸1					
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**

MASCOT : *Scoring & Statistics* © 2007-2012 Matrix Science MATRIX SCIENCE

Using the Mascot homology threshold for a 1% false discovery rate, there are 1837 peptide matches. Re-scoring with Percolator gives a useful increase to 1985 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.