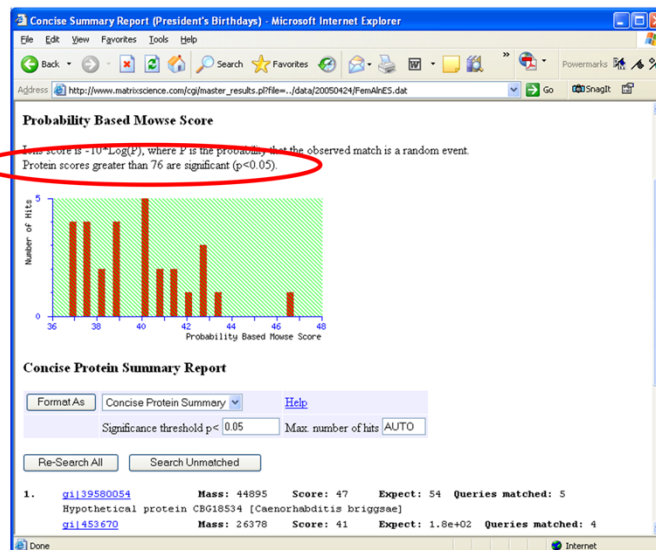


Scoring & Statistics

MASCOT

 **MATRIX
SCIENCE**

Probability based scoring



MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science



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



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%

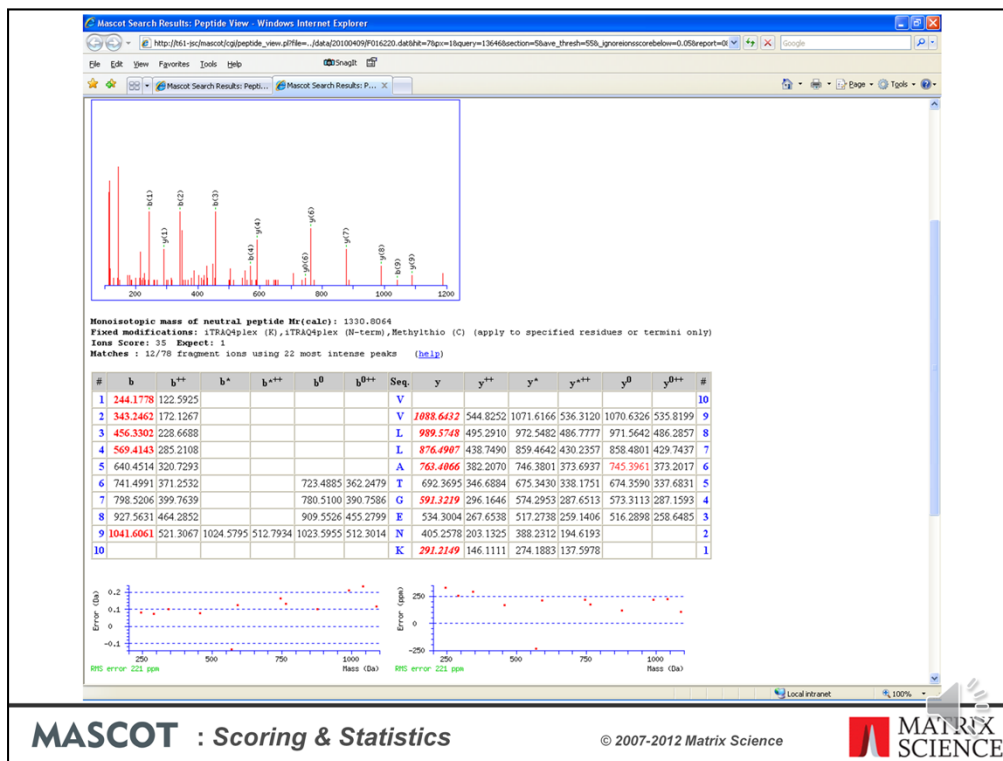
Why is probability based scoring important?

- 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



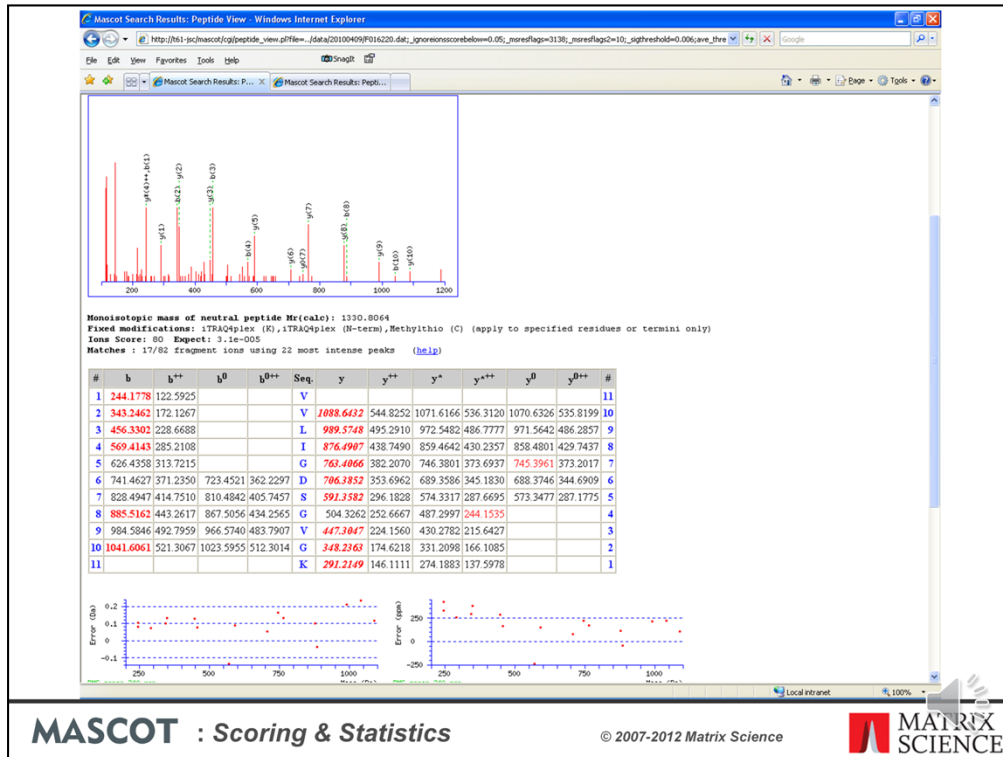
This match has a good number of matches to y and b ions, highlighted in red. All the major peaks above 200 Da seem to be labelled. Could such a good match have occurred by chance?

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.



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.

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

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.

RID=1057164027-011425-28152, - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Search Favorites History Powermarks

Address <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi> Go

```
>gi|10348033|gb|BE890074.1|BE890074 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 SEDFGVNEDLGDSAR 16
      SEDFGVNEDLGDSAR
Sbjct: 603 SEDFGVNEDLGDSAR 650
```

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

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

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

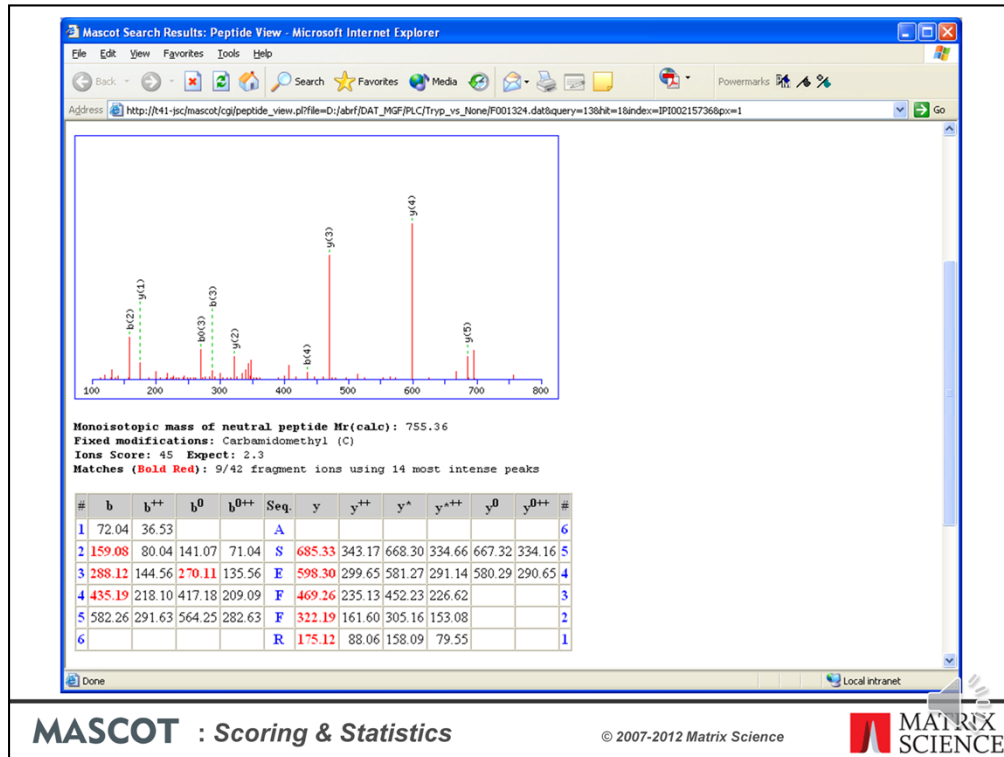
Done Internet

MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science

MATRIX
SCIENCE

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.



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

Address <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi> Go

>[gi|126792|sp|P22710|NPB1_HUMAN](#) C-myc promoter-binding protein (NPB-1) (NBP-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|ENO1_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|RAC1_DICDI](#) RAS-related protein rac1
 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

Done Internet

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

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 $-10\log_{10}(P)$, 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

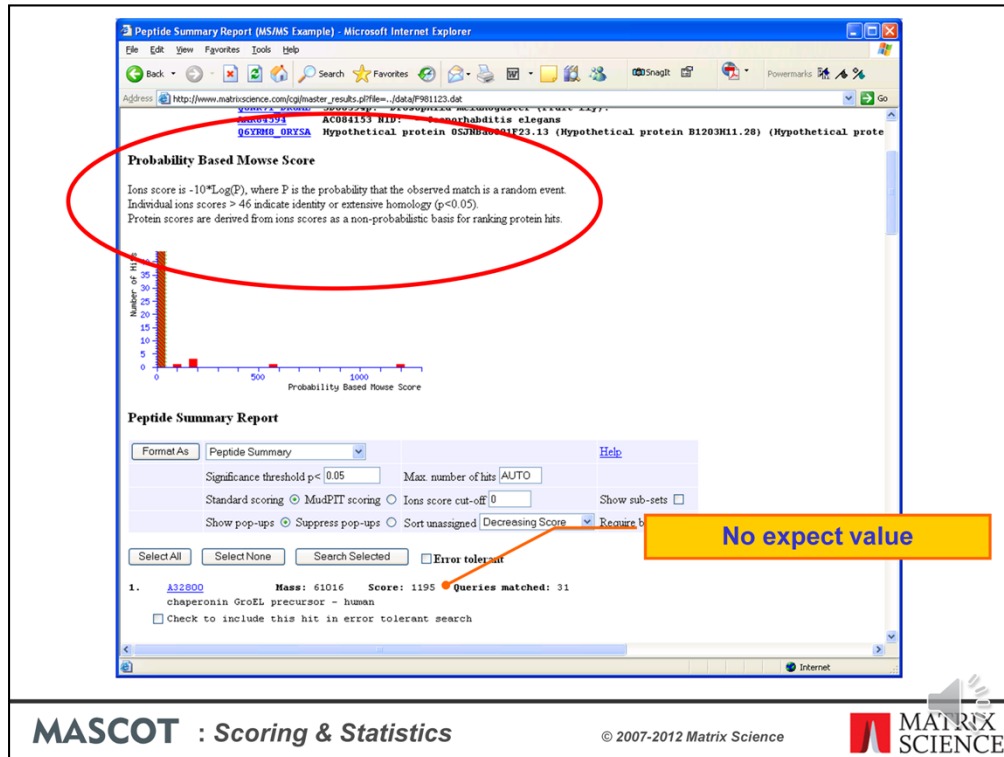


For a peptide mass fingerprint, there is just one score that matters: the protein score. This tells us whether the match is significant or not, and is determined by calculating the probability of getting the observed number of peptide mass matches if the protein sequence was random.

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.



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

$$P = 1 / (20 \times 500,000)$$

which is a score of

$$S = -10\log P = 70$$

MASCOT : Scoring & Statistics

© 2007-2012 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.

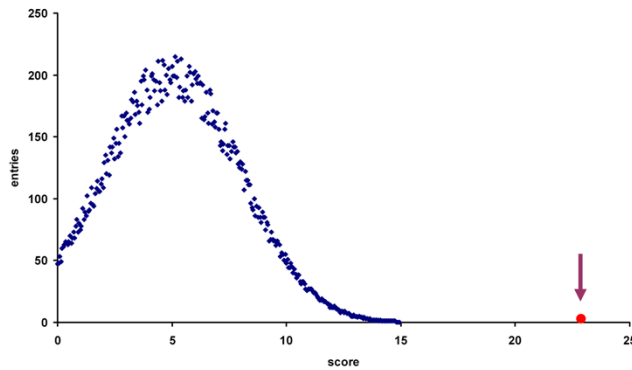
So, for example, if we are comfortable with a 1 in a 20 chance of getting a false positive match, and we are doing a PMF search of a database that contains 500,000 entries, we are looking for a probability of less than $1 / (20 \times 500,000)$ which is a Mascot score of 70

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

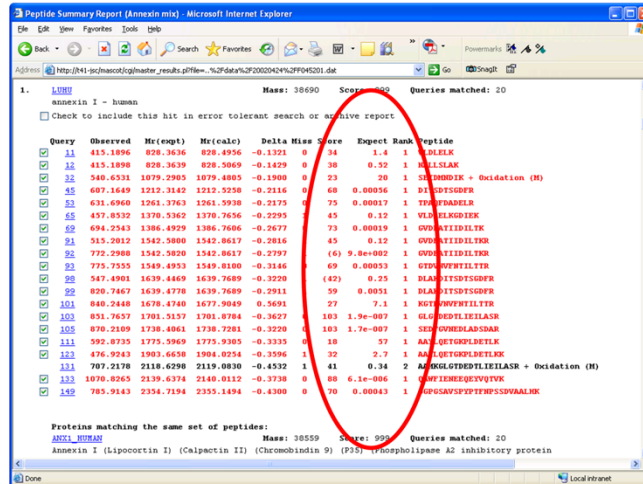
© 2007-2012 Matrix Science



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



Peptide Summary Report (Annexin.mst) - Microsoft Internet Explorer

Address: http://141-pc-mascot.cgl.mater.res.au:8080/~%2Fdata%2F20020424%2F040201.dat

1. **Annexin I - human** Mass: 36690 Score: 992 Queries matched: 20

☐ 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	LDLELK
12	415.1898	828.3639	828.5069	-0.1429	0	30	0.52	1	LDLISLAK
22	540.4551	1079.5905	1079.4805	-0.1100	0	23	20	1	LDLHDDIK + Oxidation (H)
45	607.1649	1212.3142	1212.5238	-0.2116	0	68	0.00056	1	LDLSTSDER
53	631.6960	1261.3763	1261.5938	-0.2175	0	75	0.00017	1	LDLSTSDER
65	457.8532	1370.5362	1370.7656	-0.2295	0	45	0.12	1	VLDLKGRIEK
69	694.2543	1386.4929	1386.7686	-0.2677	0	73	0.00019	1	LDVSTIDILTK
81	515.2012	1542.5000	1542.8617	-0.2616	0	45	0.12	1	LDVSTIDILTK
92	772.2988	1542.5020	1542.8617	-0.2797	0	(6)	9.8e-002	1	LDVSTIDILTK
93	775.7555	1549.4953	1549.8100	-0.3146	0	69	0.00053	1	LDVSTIDILTK
98	547.4901	1639.4469	1639.7689	-0.3220	0	(42)	0.25	1	LDVSTIDILTK
99	820.7467	1639.4778	1639.7689	-0.2911	0	59	0.0051	1	LDVSTIDILTK
101	840.2440	1678.4740	1677.9049	0.5691	0	27	7.1	1	LDVSTIDILTK
103	851.7657	1701.5157	1701.8784	-0.3627	0	103	1.9e-007	1	GLDSTIDILTK
105	870.2109	1738.4061	1738.7281	-0.3220	0	103	1.7e-007	1	SDVSTIDILTK
111	592.8735	1775.5969	1775.9305	-0.3335	0	10	57	1	AAVSTIDILTK
112	476.9243	1903.6658	1904.0254	-0.3596	0	32	5.7	1	AAVSTIDILTK
113	787.2170	2118.6298	2119.0830	-0.4532	0	41	0.34	2	AAVSTIDILTK
113	1070.8265	2139.6374	2140.0112	-0.3738	0	88	6.1e-006	1	LDVSTIDILTK
119	785.9143	2354.7194	2355.1494	-0.4300	0	70	0.00043	1	LDVSTIDILTK

Proteins matching the same set of peptides:

Annexin I Mass: 36559 Score: 990 Queries matched: 20

Annexin I (lipocortin I) (Calpain II) (Chromobindin 9) (P35) (Phospholipase A2 inhibitory protein)

MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science



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

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

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

score = 40 corresponds to $E = 0.5$

score = 50 corresponds to $E = 0.05$

score = 60 corresponds to $E = 0.005$

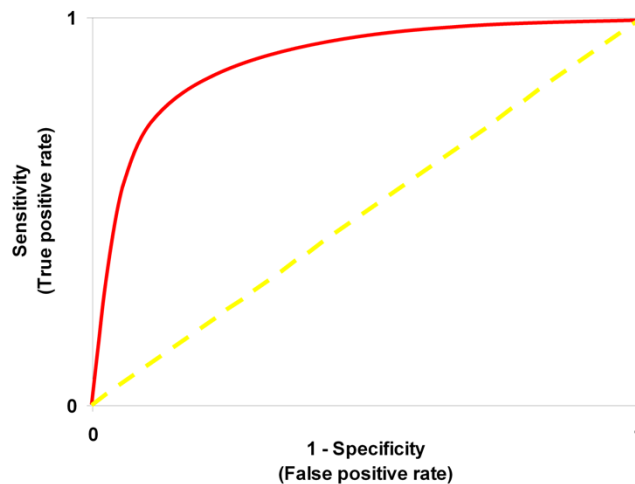
The expectation value does not contain new information. It can be derived directly from the score and the threshold. The advantage is that it tells you everything you need to know in a single number.

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

© 2007-2012 Matrix Science

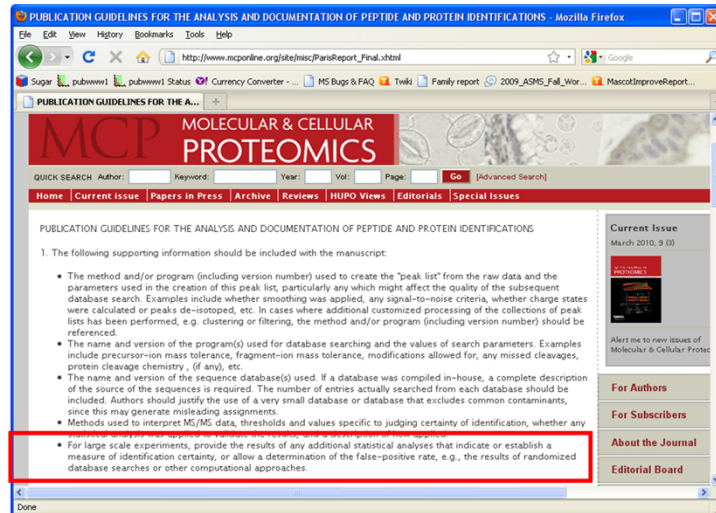


The most important attributes of a scoring scheme are sensitivity and specificity. That is, you want as many correct matches as possible, and as few incorrect matches as possible.

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



MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science



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)

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

The screenshot shows the Mascot MS/MS Ions Search web form. The 'Decoy' checkbox is highlighted with a red circle. The form includes fields for 'Your name', 'Email', 'Search title', 'Database(s)', 'Enzyme', 'Allow up to', 'Quantitation', 'Taxonomy', 'Fixed modifications', 'Variable modifications', 'Peptide tol. ±', 'Peptide charge', 'Data file', 'Data format', 'Instrument', 'Decoy', 'Precursor', 'Error tolerant', 'Report top', 'Start Search', and 'Reset Form'.

MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science



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.

The screenshot displays the Mascot 2.4 web interface. At the top, the browser address bar shows the URL: `bogong/mascot_2_4_0_64/cgi/master_results_2.pl?file=.,%2Fdata%2F20120519%2F001683.dat`. The main section is titled "Protein Family Summary". It includes a "Filter" button and several input fields: "Significance threshold p<" set to 0.05, "Max. number of families" set to AUTO, "Ions score or expect cut-off" set to 0, and "Dendrograms cut at" set to 0. There are also checkboxes for "Show Percolator scores" and a dropdown for "Preferred taxonomy" set to "All entries".

Below this, the "▼Decoy search summary (reversed protein sequences)" section is expanded. It shows peptide matches in a table with columns for "Peptide matches", "in iPRG_2012 in Decoy", and "FDR". The table has two rows: one for "above identity threshold" (3110 matches, 61 families, 1.96% FDR) and one for "above identity or homology threshold" (4247 matches, 127 families, 2.99% FDR). Each row has an "Adjust to" button and a dropdown set to 1%.

Below the table, it says "Decoy results are available in [the decoy report](#)." There are buttons for "Report Builder" and "Unassigned (11996)", and a "S permalink" link.

The "Protein families 1-10 (out of 599)" section shows a list of protein families. It includes a "10 per page" dropdown, a "1 2 3 4 5 6 ... 60 Next" pagination bar, and "Expand all" and "Collapse all" buttons. There is a search bar with "Accession" and "contains" dropdowns and a "Find" button.

The protein families listed are:

- 1 P00925: Enolase 2 OS=Saccharomyces cerevisiae (strain ...)
- 2 P00924: Enolase 1 OS=Saccharomyces cerevisiae (strain ...)
- P00549: Pyruvate kinase 1 OS=Saccharomyces cerevisiae ...
- 1 P40150: Heat shock protein SSB2 OS=Saccharomyces cer...
- 2 P11484: Heat shock protein SSB1 OS=Saccharomyces cer...

At the bottom of the interface, there is a footer with the text "MASCOT : Scoring & Statistics", "© 2007-2012 Matrix Science", and the "MATRIX SCIENCE" logo.

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, 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

The screenshot displays the Mascot search results interface. At the top, the browser address bar shows the URL: `bogong/mascot_2_4_0_64/cgi/master_results_2.pl?file=.,%2Fdata%2F20120519%2F001683.dat;_selected_fdr=1;`. The main section is titled "Protein Family Summary". Below this, there are several filters and settings: "Significance threshold p<" is set to 0.01245, "Max. number of families" is set to AUTO, "Ions score or expect cut-off" is 0, "Dendrograms cut at" is 0, "Show Percolator scores" is unchecked, and "Preferred taxonomy" is set to "All entries".

Below the filters, there is a section titled "▼Decoy search summary (reversed protein sequences)". It shows peptide matches in the iPRG_2012 In Decoy. The table has columns for "Peptide matches", "in iPRG_2012 In Decoy", and "FDR". The data is as follows:

Peptide matches	in iPRG_2012 In Decoy	FDR
- above identity threshold	2478	22 0.89%
- above identity or homology threshold	3422	34 0.99%

Below the table, it says "Decoy results are available in [the decoy report](#)". There are buttons for "Report Builder" and "Unassigned (12178)".

Below this, there is a section titled "Protein families 1-10 (out of 495)". It shows a list of protein families with their accession numbers and descriptions. The first family is P00925, which is Enolase 2 OS=Saccharomyces cerevisiae (strain ...). The second family is P00924, which is Enolase 1 OS=Saccharomyces cerevisiae (strain ...). The third family is P00549, which is Pyruvate kinase 1 OS=Saccharomyces cerevisiae

At the bottom of the screenshot, there is a footer with the text "MASCOT : Scoring & Statistics" and "© 2007-2012 Matrix Science". There is also a logo for "MATRIX SCIENCE".

The significance threshold has been automatically adjusted from 0.05 to 0.012.

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.

Select Summary Report (PRG)

Format As: **Select Summary (protein hits)** [Help](#)

Significance threshold $p < 0.01245$ Max. number of hits: **AUTO** Show Percolator scores: ☐

Standard scoring: ☐ MudPIT scoring ☒ Ions score or expect cut-off: **0** Show sub-sets: **0**

Show pop-ups: ☒ Suppress pop-ups ☐ Require bold red: ☐

Re-Search ☒ All queries ☐ Unassigned ☐ Below homology threshold ☐ Below identity threshold

1. **Q8TE96** Score: 67 Matches: 4 (3) Sequences: 2 (1)

Random sequence.

Query	Observed	Mr(expt)	Mr(calc)	p.p.m.	Miss	Score	Expect	Rank	Unique	Peptide
1036	431.7427	861.4709	861.4708	0.08	0	48	0.00057	1	U	R.LVQFEAR.Y 1035 1037
10272	778.8868	1555.7591	1555.7671	-5.12	0	2	13	5	U	R.DPPITDHYIPSPR.E

Proteins matching the same set of peptides:
Q8TE96-2 Score: 67 Matches: 4 (3) Sequences: 2 (1)

Random sequence.

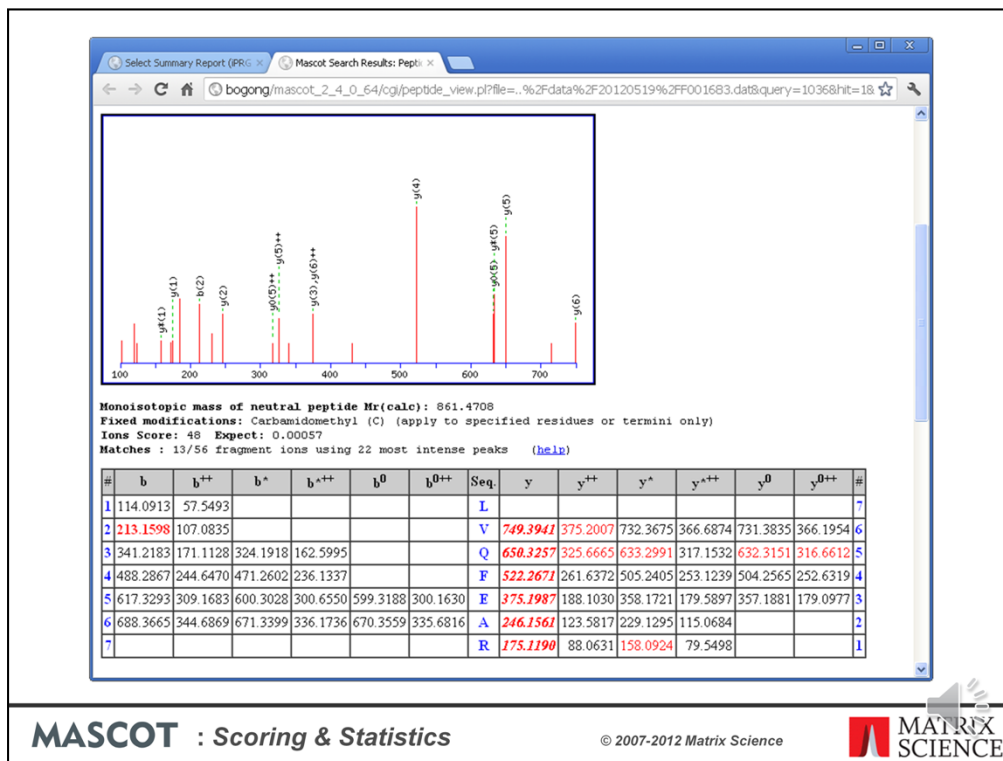
2. **Q03653** Score: 62 Matches: 6 (3) Sequences: 3 (1)

Random sequence.

Query	Observed	Mr(expt)	Mr(calc)	p.p.m.	Miss	Score	Expect	Rank	Unique	Peptide
133	404.2035	806.3925	806.3923	0.28	0	44	0.00079	1	U	R.ADEAVFR.Y 134 135 136
4474	529.7965	1057.5784	1057.5768	1.56	0	2	24	4	U	R.DSLSSIPALR.L
12192	569.2921	1704.8545	1704.8651	-6.16	3	8	4.5	1	U	R.RHNPINKEQITCK.L

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 48 and 44, with expect values well below 1%. If we click on the query number link to display the Peptide View of one of these matches ...

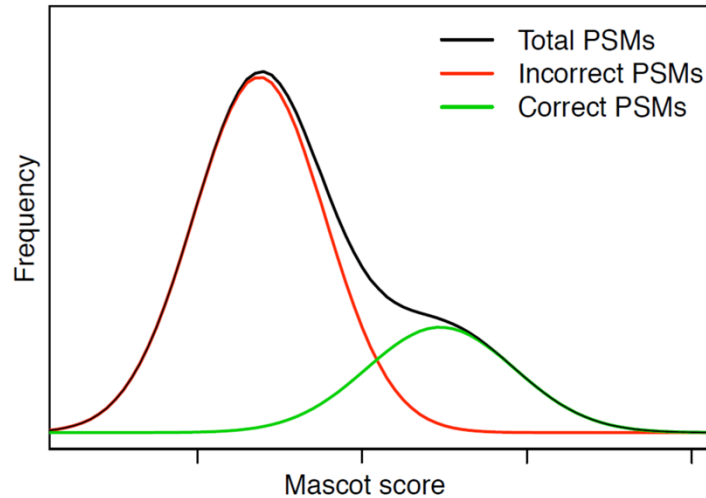


This is what it looks like. A pretty decent match from a decoy sequence. Tryptic peptide, no modifications, complete run of y ions, most of the larger peaks matched.

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



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

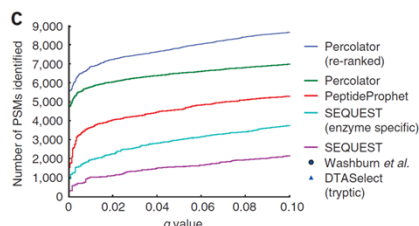


Anal. Chem. 2002, 74, 5383–5392

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 Käll¹, Jesse D Canterbury¹, Jason Weston², William Stafford Noble^{1,3} & Michael J MacCoss¹

MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science

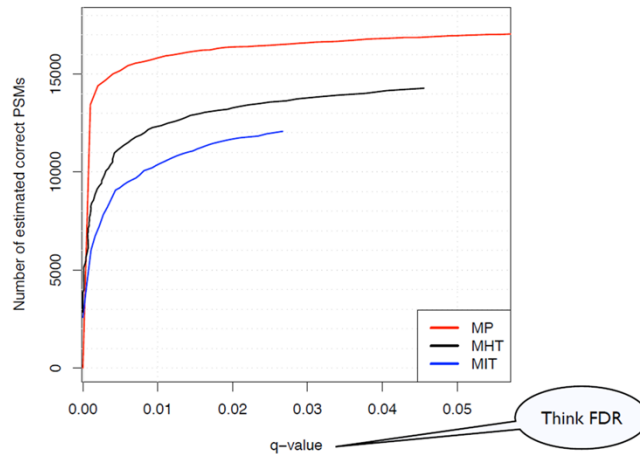


One of the first attempts to do this was Peptide Prophet from the ISB. This was and is popular for transforming Sequest scores into probabilities.

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 termini.

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).

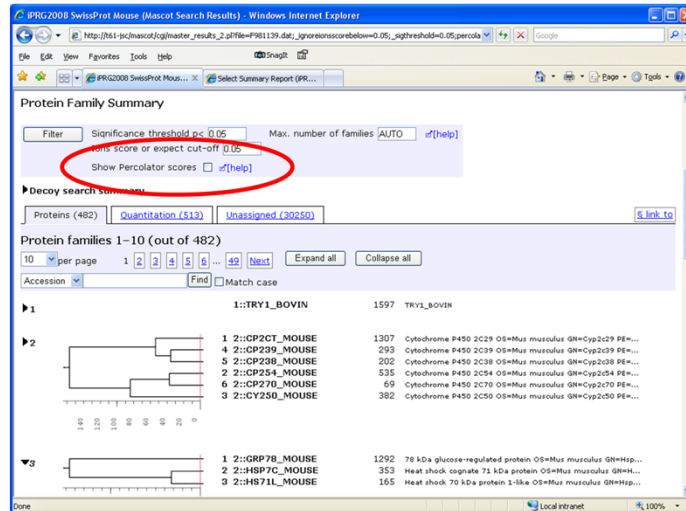
MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science



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



MASCOT : Scoring & Statistics

© 2007-2012 Matrix Science



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	▶1	U	■	■	K.VQVEYK.G
0.4581	0	18	0.015	▶1	U	■	■	K.VLEDSDLK.K
0.0517	0	21	0.0087	▶1	U	■	■	K.VLEDSDLK.K
0.2227	0	25	0.0031	▶1	U	■	■	K.IITINDQNR.L

Score > 13 indicates identity

MASCOT : Scoring & Statistics

© 2007-2012 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.