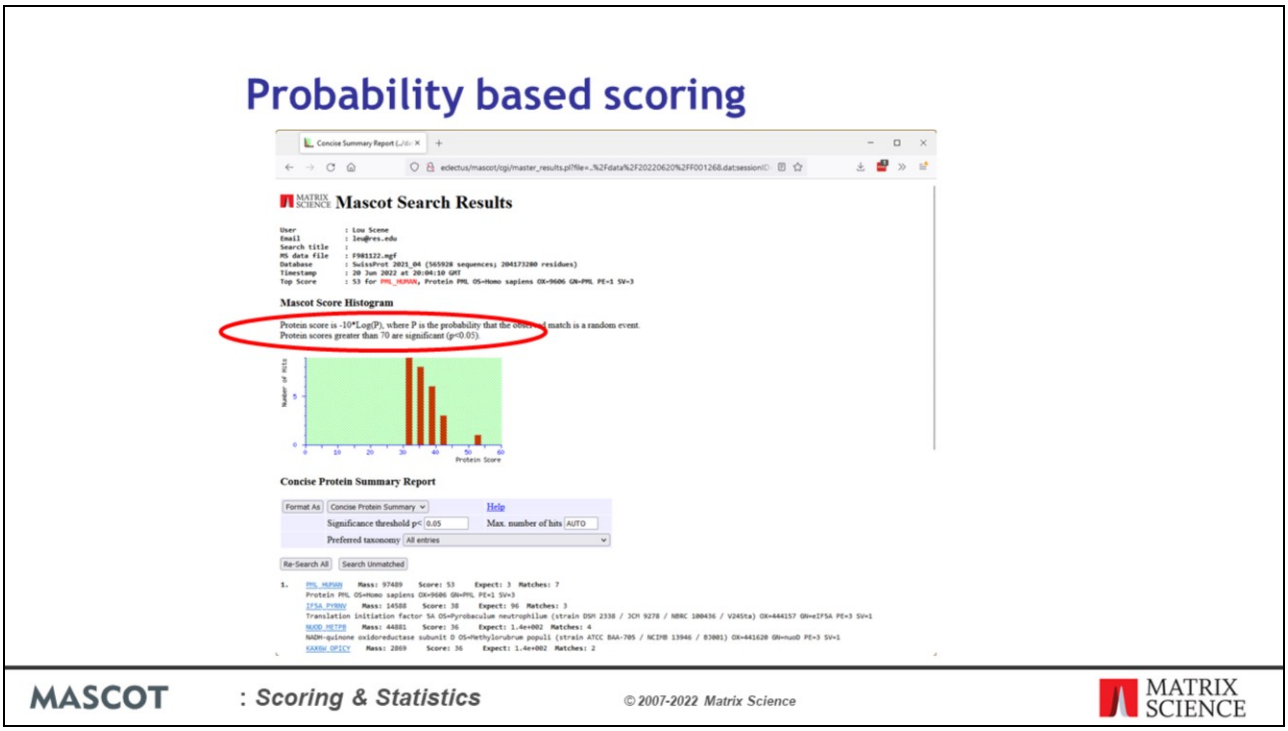
Scoring & Statistics



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 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

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%

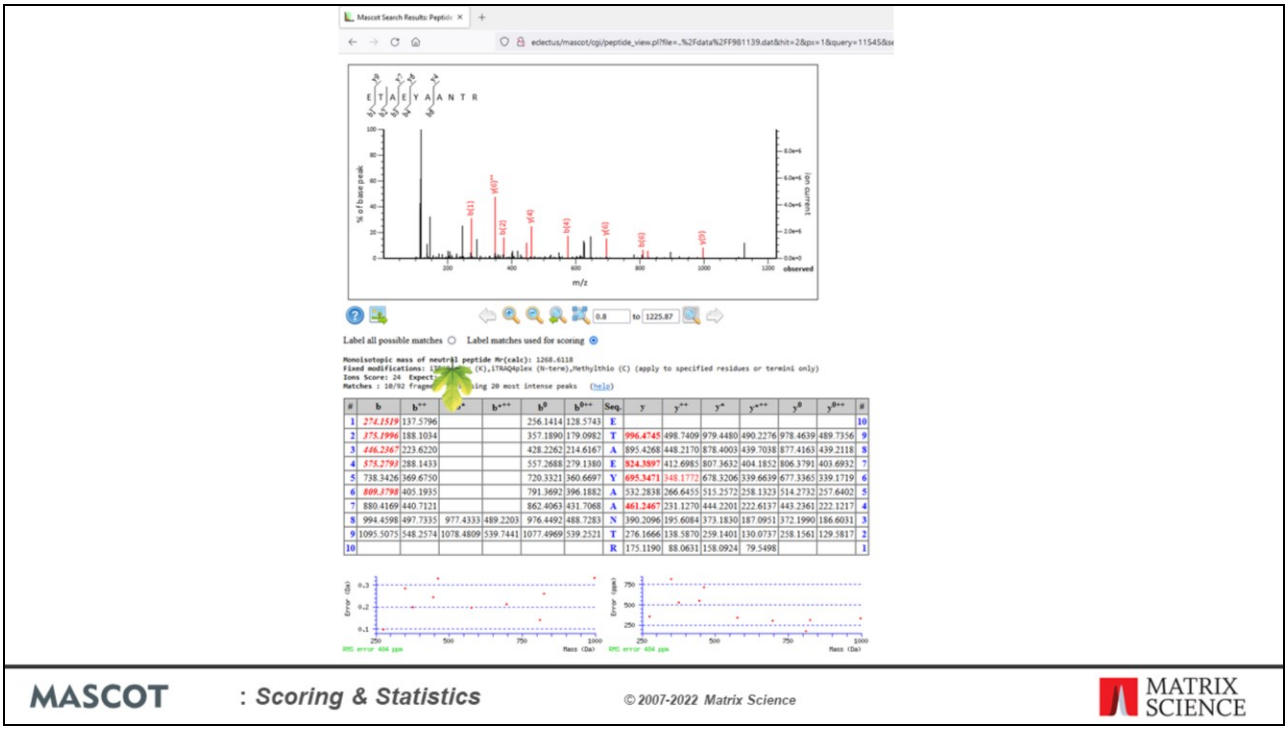
Why is probability based scoring important?

- 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



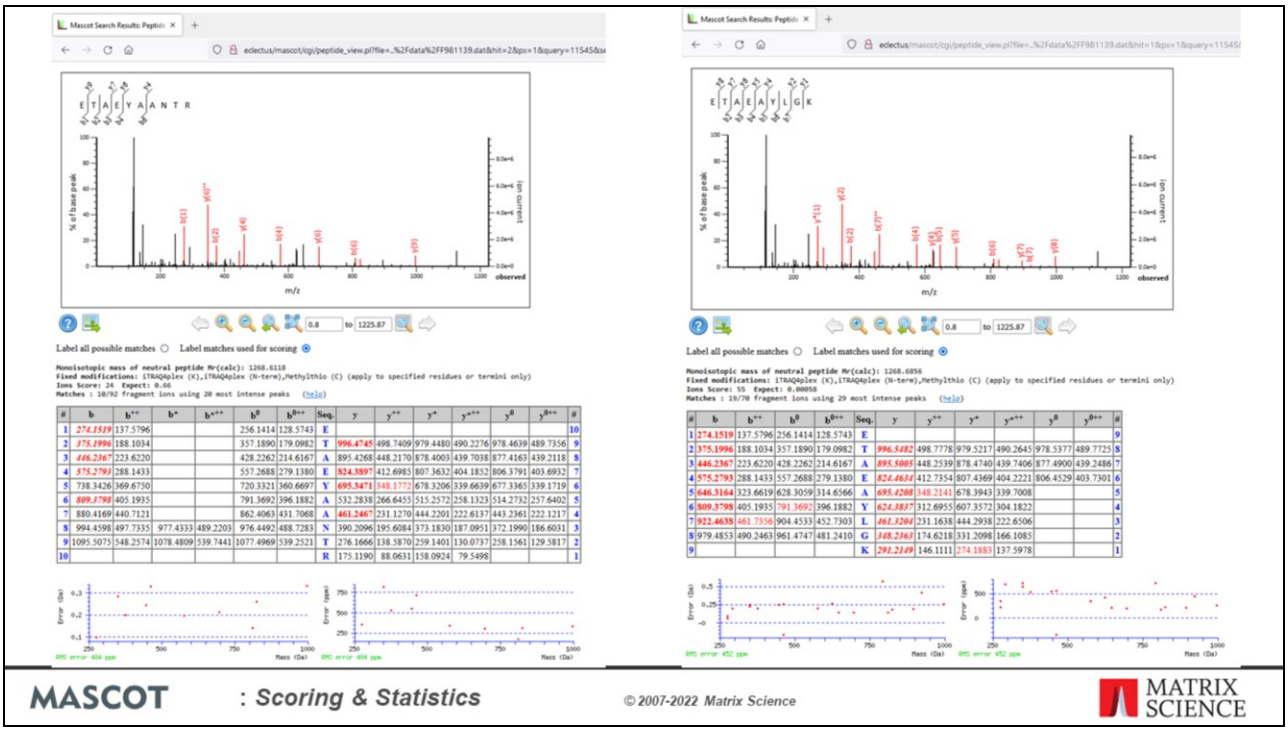
This match has a good number of matches to y and b ions, highlighted in red. Almost all the major peaks above 200 Da seem to be labelled. Could such a good match have occurred by chance?

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.



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

Can we calculate a probability that a match is correct?

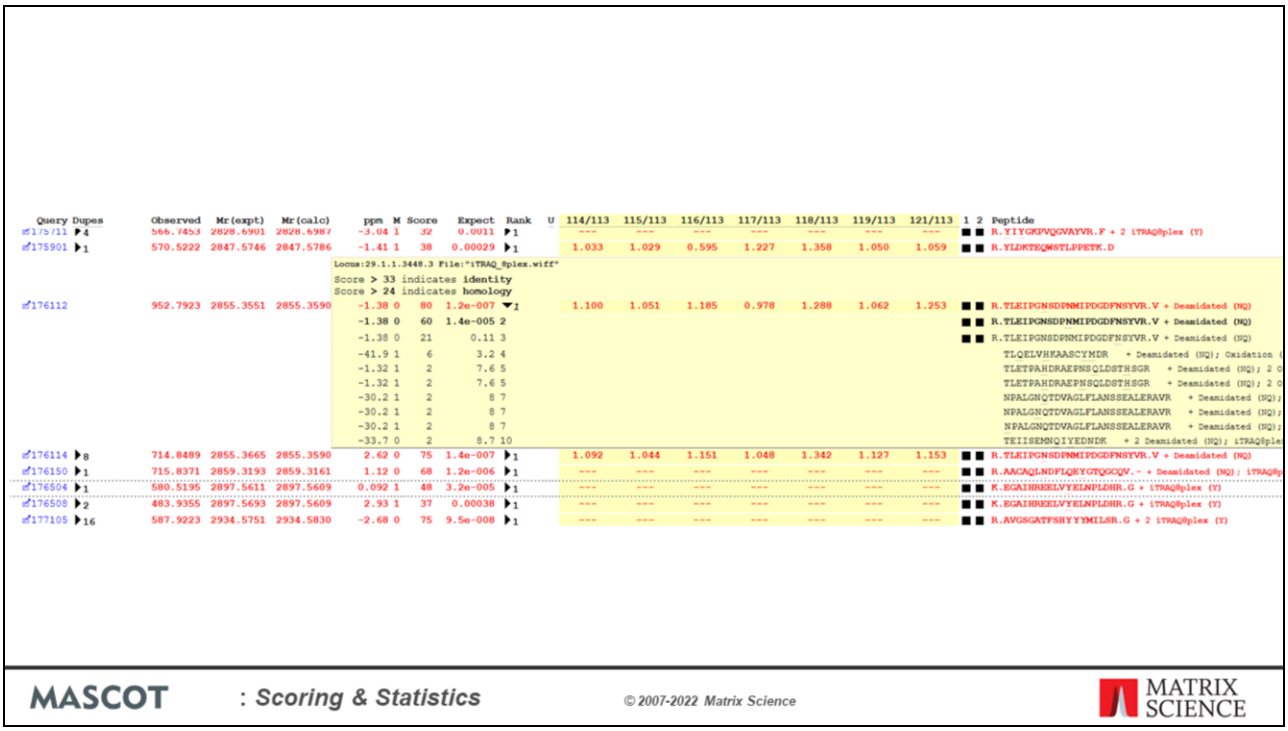
Yes, if it is a test sample and you know what the answer should be

- 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

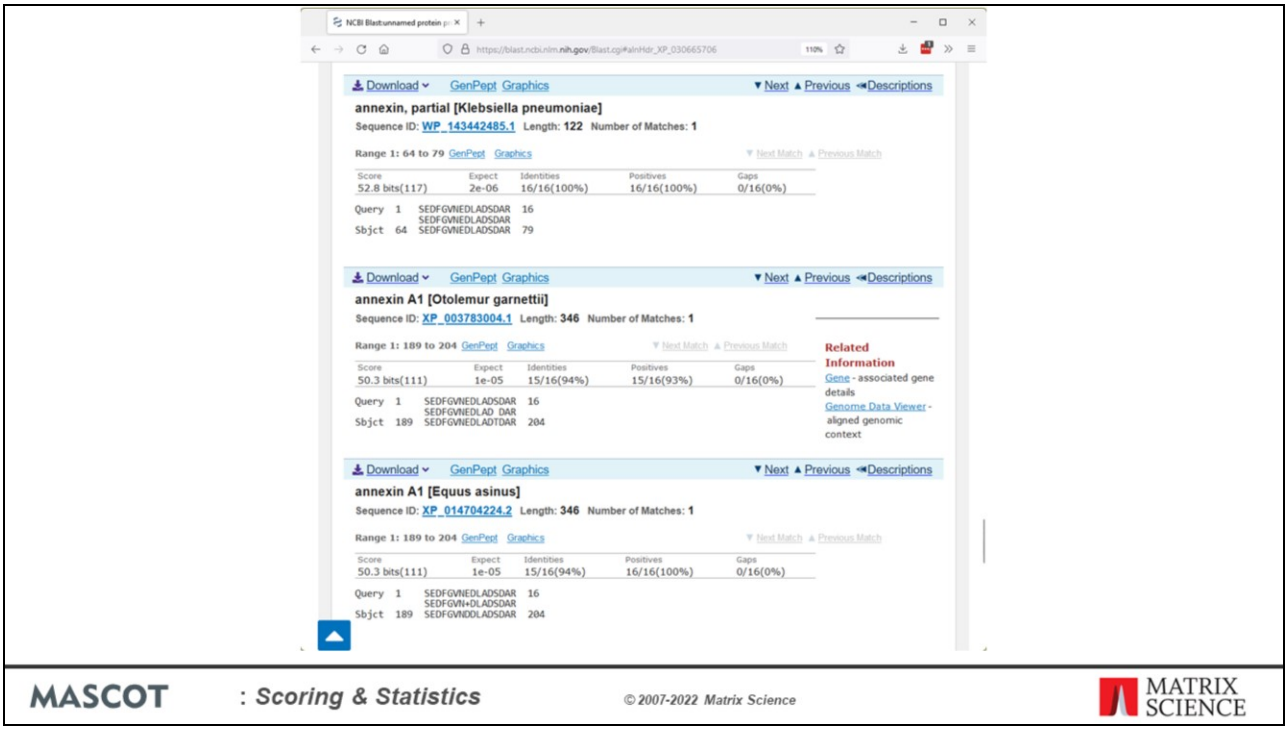
Probability based scoring calculates the probability that the match is random. This is, the probability that the match is meaningless. Many people ask whether we can report the probability that the match is correct. Is this possible?

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.



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.

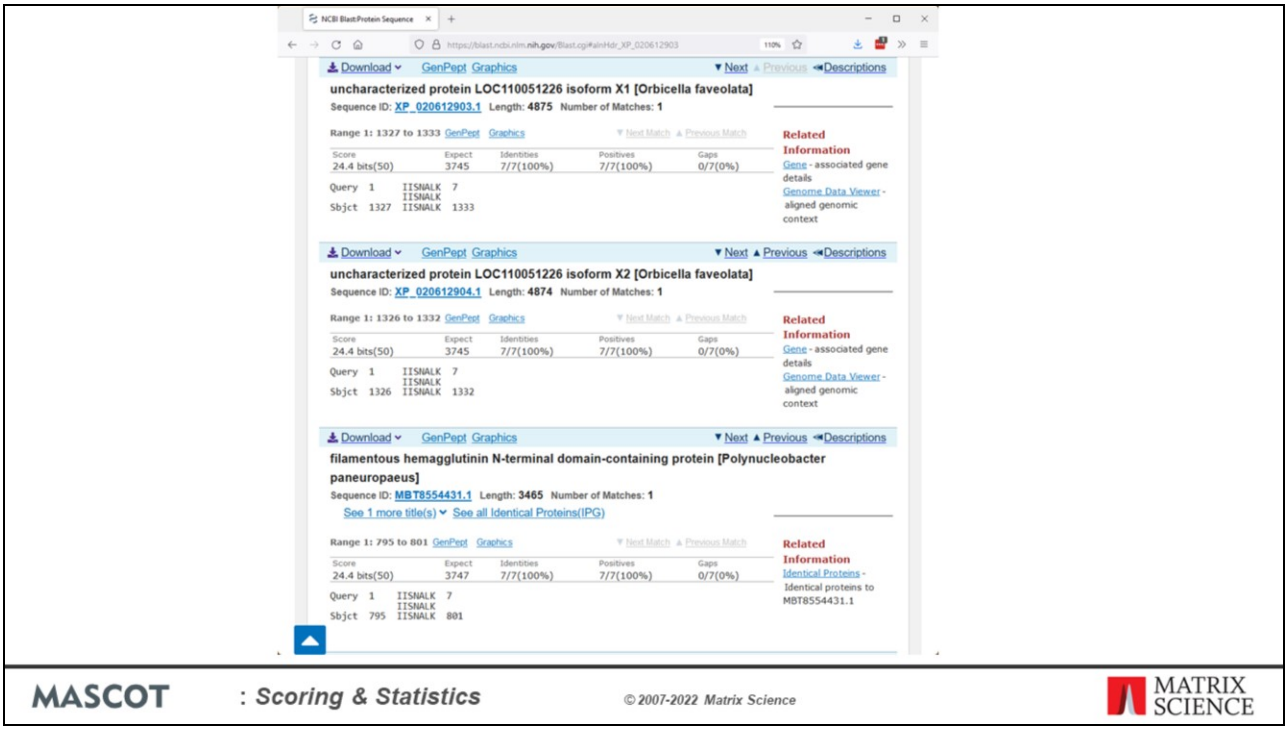


It is a similar situation in Blast, except that you have the luxury of seeing when you have a perfect identity match. Here, the identity match has an expectation value of 2E-6, which reminds us that it would be a random match if the database was 2 million times larger. The match with one different residue 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.

#	b	b^{++}	b*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
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.



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.
- 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.

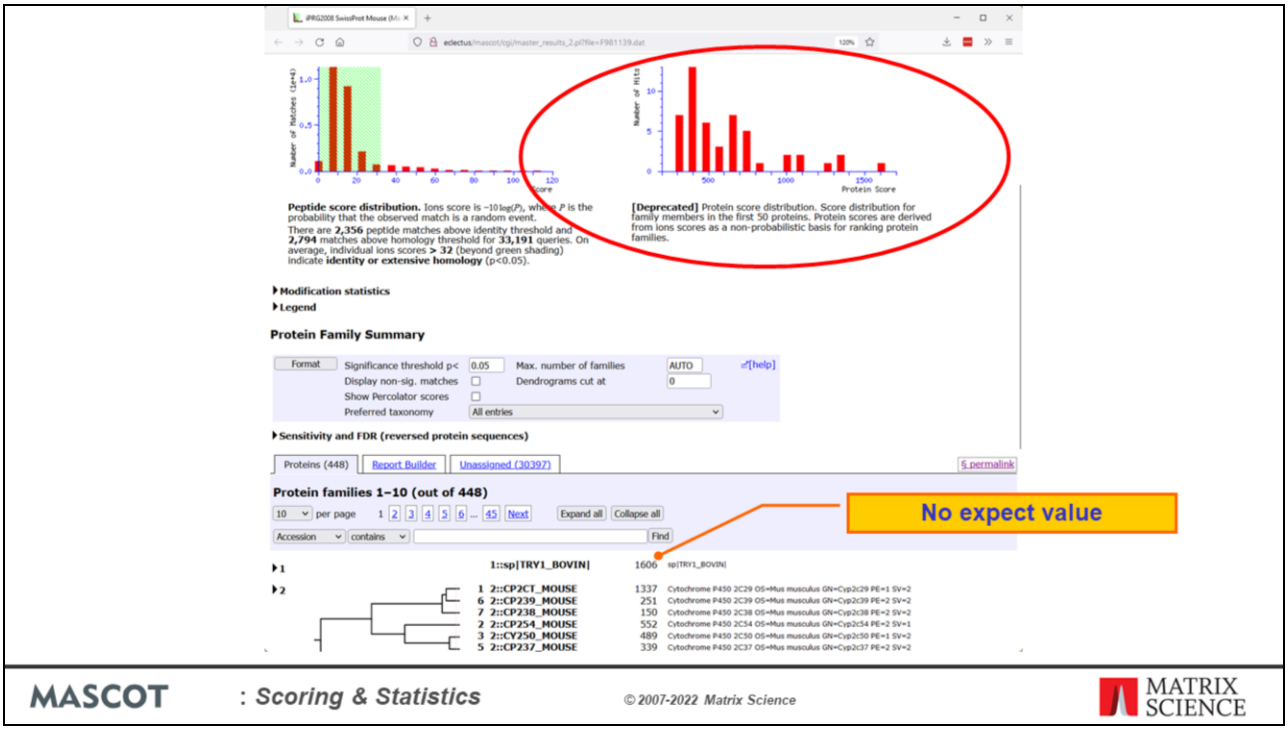
For an MS/MS search, the protein score is ***not*** statistically rigorous. It is just a way of ranking the protein hits

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 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.

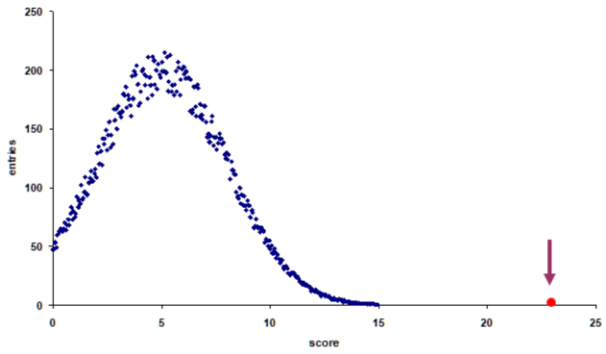
So, for example, if we are comfortable with a 1 in a 20 chance of getting a false positive match, and we are doing a PMF search of a database that contains 500,000 entries, we are looking for a probability of less than 1 / (20 x 500,000) which is a Mascot score of 70

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

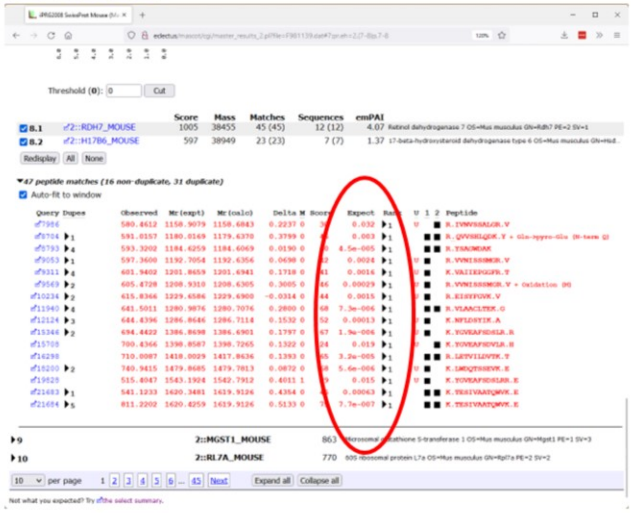


MASCOT : Scoring & Statistics MATRIX SCIENCE

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



MASCOT : Scoring & Statistics © 2007-2022 Matrix Science 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

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

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

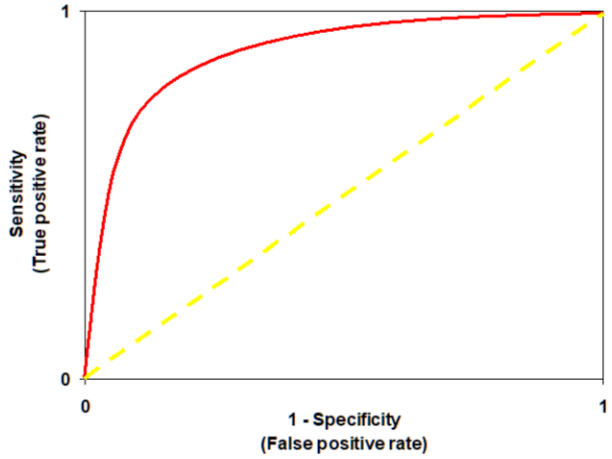
The expectation value does not contain new information. It can be derived directly from the score and the threshold. The advantage is that it tells you everything you need to know in a single number.

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



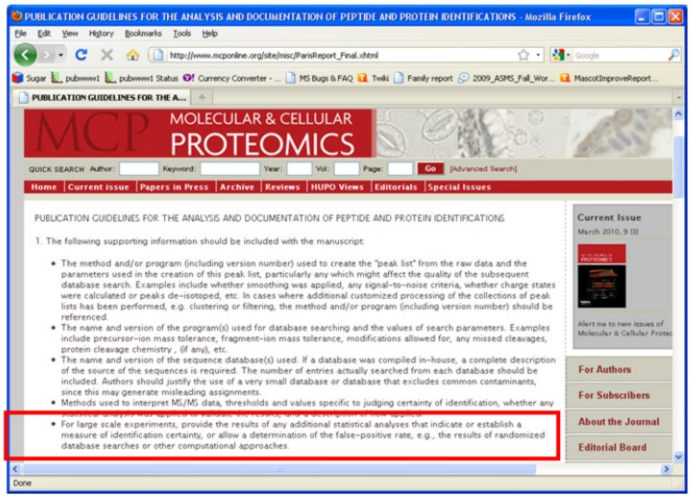
MASCOT : Scoring & Statistics MATRIX SCIENCE

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



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

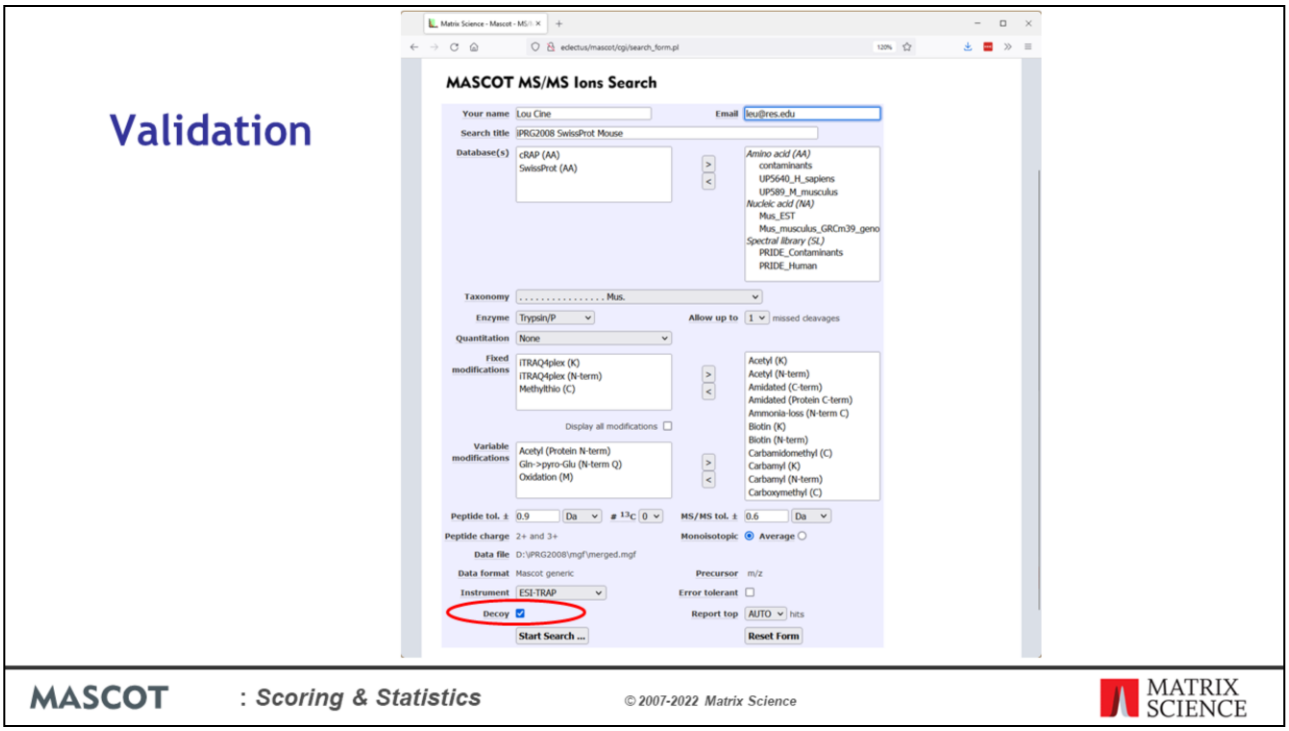
Gives a clear estimate of false discovery rate

- 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.

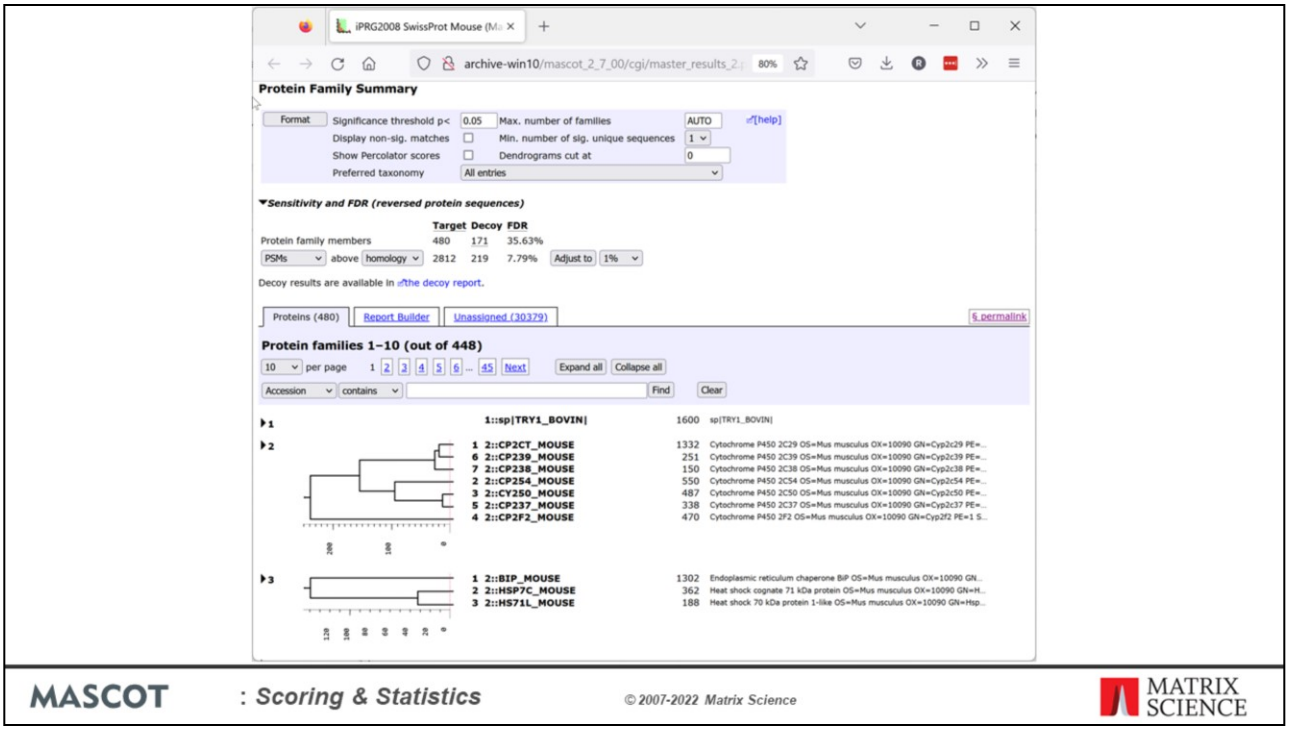
You'll read a lot of discussion in the literature about whether the decoy sequences should be reversed or randomised; whether to search a single database containing both target and decoy sequences or separate databases. I suggest the most important thing is to do a decoy search; any decoy search. What 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.

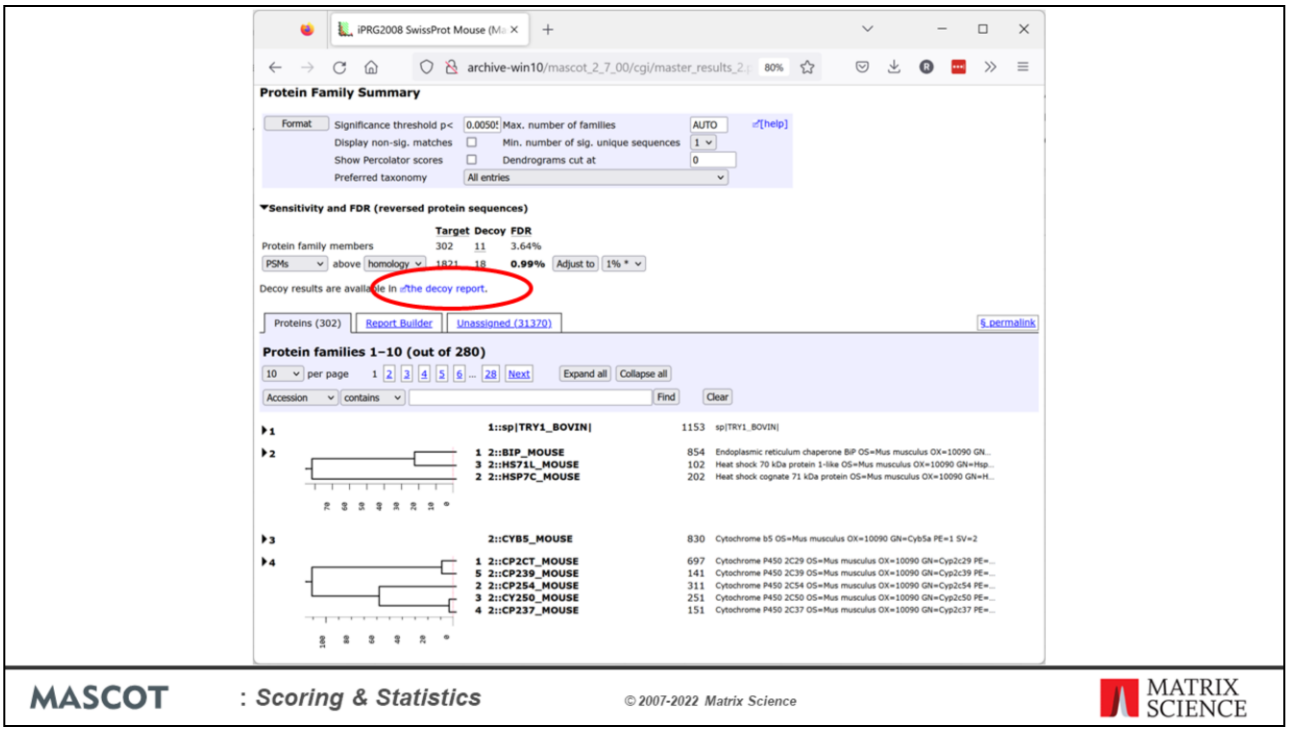


On our public web site there is a help page devoted to decoy database searches. It includes a download link to a utility program that allows you to create a randomised or reversed database. If you have an early version of Mascot, or if you want to verify the results from another search engine, you can 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.

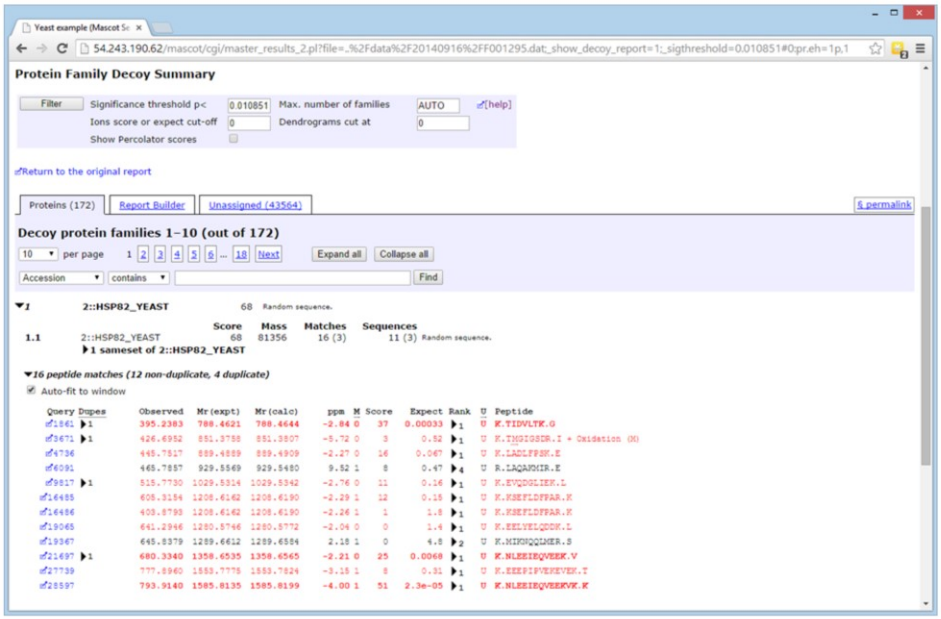


When the search is complete, the statistics for matches to the decoy sequences are reported in the result header. If you change the significance threshold, the numbers are recalculated. In Mascot 2.4 and later, there is a button to adjust the significance threshold so as to achieve a chosen FDR value. For example, if we choose 1% FDR using the homology threshold



The significance threshold has been automatically adjusted from 0.05 to 0.005054.

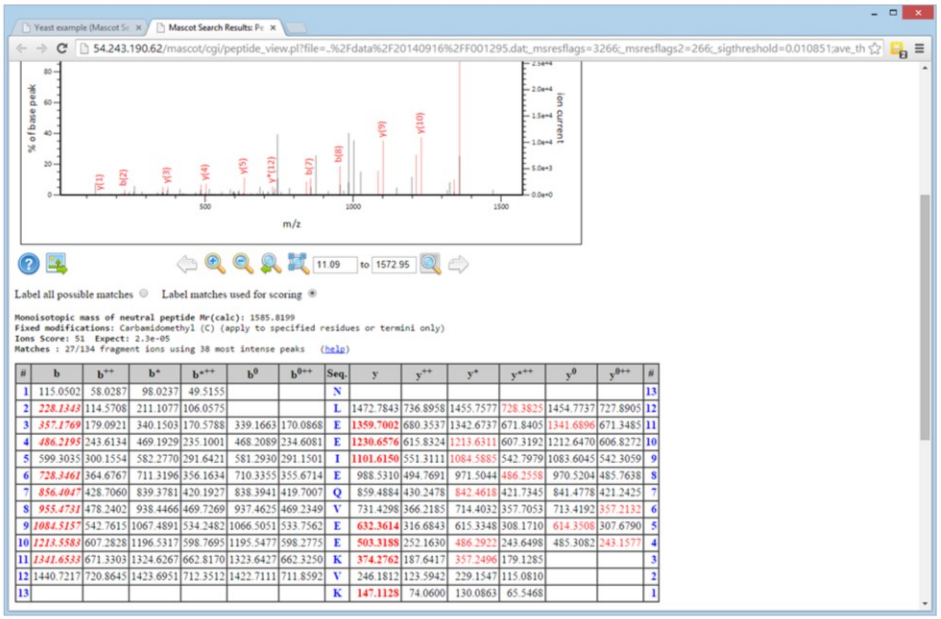
Why do we get these false positives? Do they reflect some defect in the search engine? Let's have a closer look. If you click the link here, then you will see the results from searching the randomised database.



MASCOT

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 …



#	b	b++	b*	b*++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
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

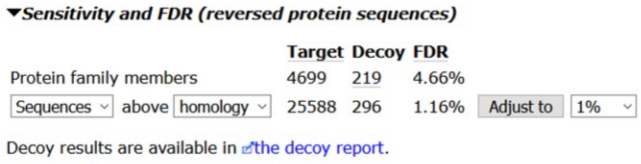
This is what it looks like. A pretty decent match from a decoy sequence. Tryptic peptide, no variable modifications, good run of b and y ions, most of the larger peaks matched.

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

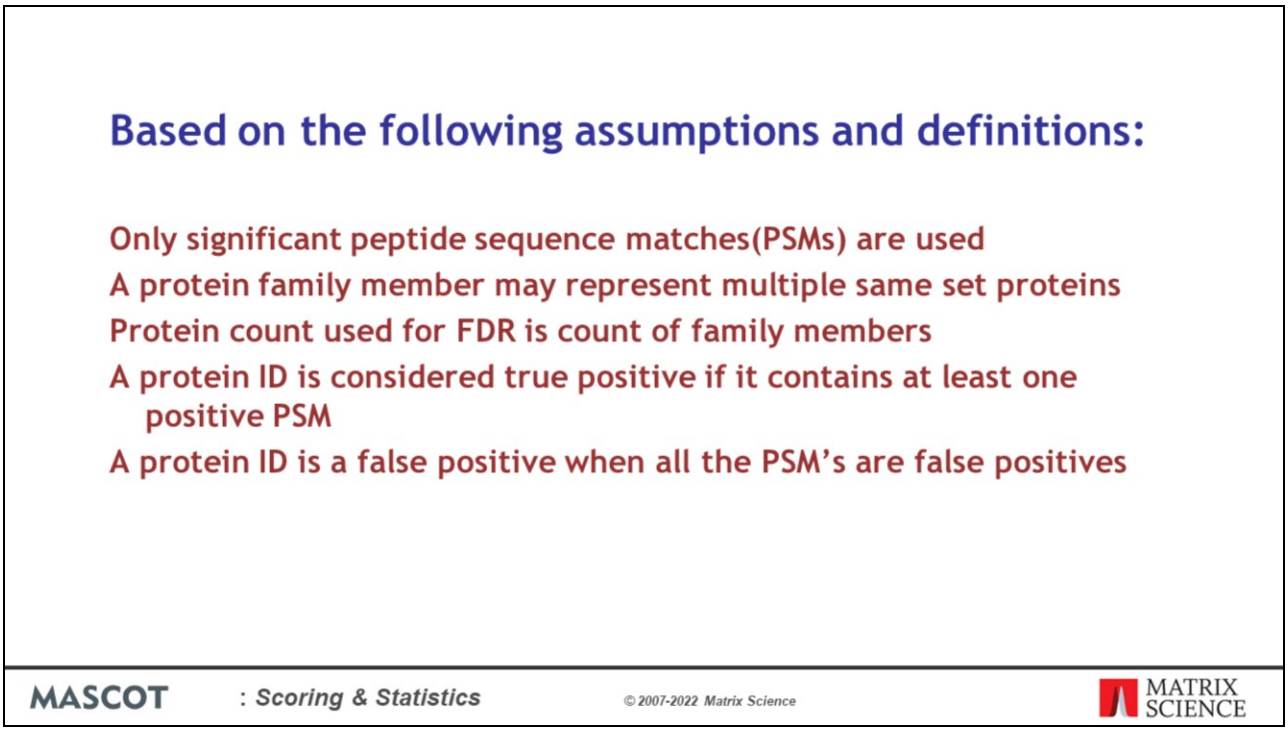
Automatically calculated when decoy option selected
Displayed in the decoy section:



MASCOT : Scoring & Statistics MATRIX SCIENCE

New in Mascot Server 2.7 is the Protein FDR value which is automatically calculated when the decoy search option is selected.

It is displayed above the peptide/PSM decoy results.



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.

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

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%?

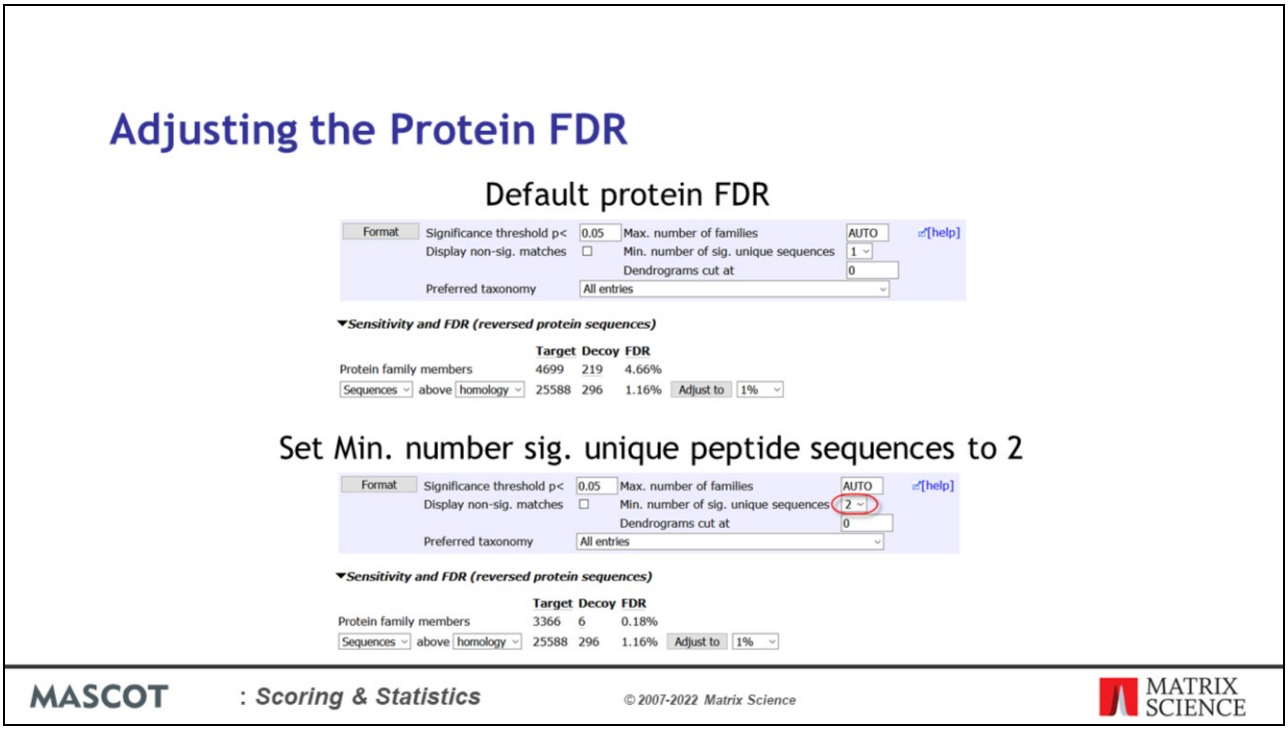
False PSMs distributed across 10 proteins

- 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%

MASCOT : Scoring & Statistics MATRIX SCIENCE

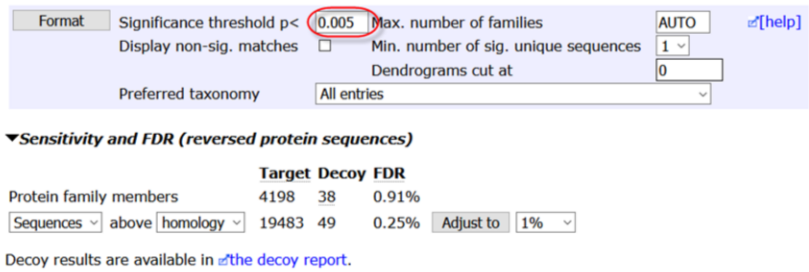
Imagine the and the search results show 500 target proteins and 10 decoy proteins. Does this mean protein FDR is 10/500 = 2%? No, it does not. We can assume the false PSMs in the target are distributed across 10 proteins, but some of these will also contain true PSMs, so should not be counted as false. 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%.



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 ~4.5%. If we want to lower the Protein FDR we can try adjusting the peptide FDR to 1%. In this case the default peptide FDR is already close to 1% so we can 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 0.18%.

Adjusting the Protein FDR



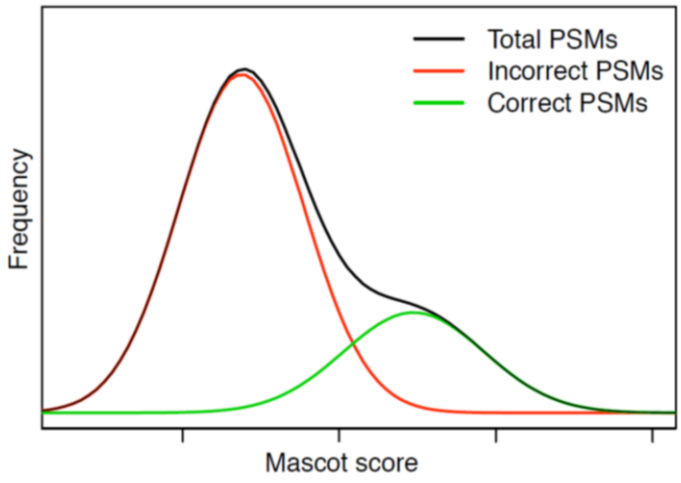
Alternatively if we can adjust the Significance threshold for the results. I took a guess and reduced it by a factor of 10 from the default values of 0.05 to 0.005 and clicked the format button. The resulting Protein FDR is approximately 1%.

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 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.

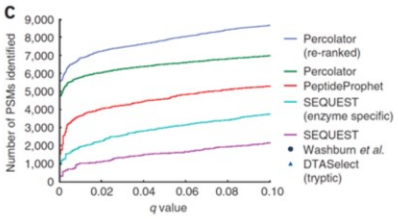
Sensitivity optimisation

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

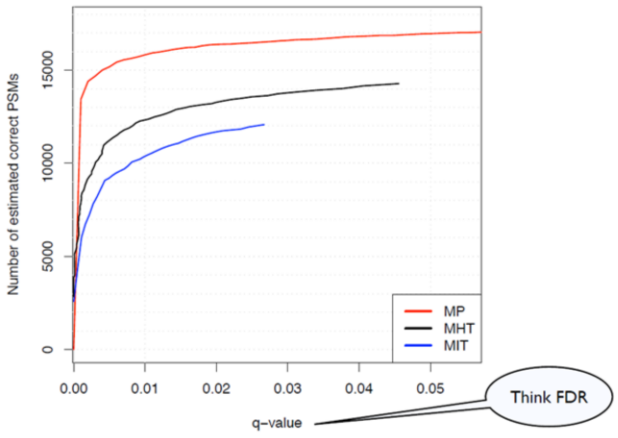
Lukas Käll[1], Jesse D Canterbury[1], Jason Weston[2], William Stafford Noble[1,3] & Michael J MacCoss[1]

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.

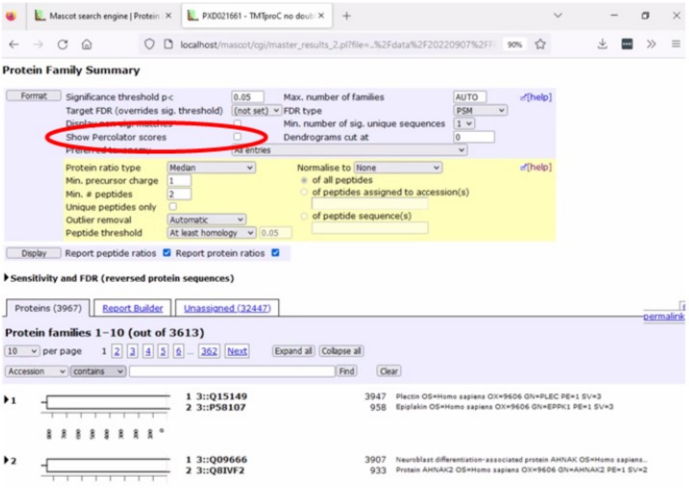
Sensitivity optimisation



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

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.

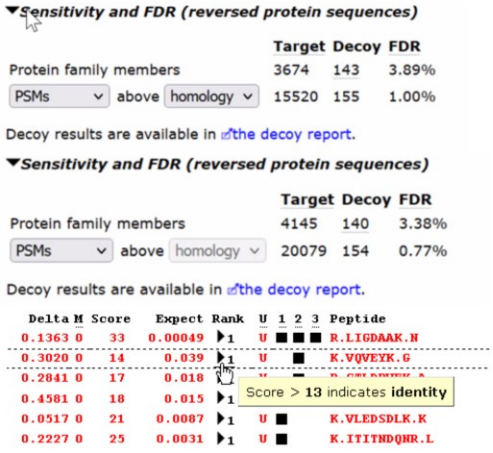
Sensitivity optimisation



MASCOT : Scoring & Statistics © 2007-2022 Matrix Science 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. In this example we take a medium sized search result.

Sensitivity optimisation



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.