MASCOT Search Results Interpretation

The Mascot protein identification program (Matrix Science, Ltd.) uses statistical methods to assess the validity of a match. MS/MS data is not ideal. That is, there are unassignable peaks (noise) and usually missing peaks in every spectrum, and even the assignable peaks have variable, and somewhat unpredictable peak intensity. Moreover, coincidental peptide matches occur, due to randomly similar peptide and peptide fragment masses. As the database size grows, the potential for random matches increases. Therefore, protein identification based on Mascot peptide MS/MS searches rely on a statistical argument that incorporates the number of peptides in the database which have the same mass (within user-defined search mass tolerance). Choice of database, or subset of sequences within a database (prokaryotes, or fungi, for instance), and user-defined mass tolerances for the search affects the scoring. You assign the database to be considered in your search, and this should be specified in your manuscript methods, etc. Contemporary publishing guidelines can also require the inclusion of a decoy database search for certain types of data. A decoy database is one that contains sequences not expected to include anything except random matches, such as a reversed sequence database of the same size as the database that was searched, and results provide a false positive rate to include with your data. Mascot offers this option which can be included while setting up searches.

The strength of a peptide match is based initially on the coincidence of masses; 1) the precursor (peptide) mass and 2), MS/MS fragment masses present in the scan, coinciding with the predicted masses of peptides (and peptide fragment masses) calculated on the basis of the sequence from a peptide present in a protein database accession. The statistical strength of the match is negatively affected by the occurrence of unassignable peaks (presumably noise), and by the number of peptides in the database that coincidentally have the same peptide mass, within user-definable search tolerance. Most search programs use the peptide mass first, to select a subset of peptides from the database that have the correct mass, and then the algorithm compares the MS/MS peak masses to a set of virtual/predicted fragment masses, generated in silico from the sequence of each peptide in the subset, one by one. Identified protein database entries are presented in descending order in the search results, based on scores. A protein that receives more matching MS/MS scans in the search receives a higher overall score and is ranked higher in the search results. The top listed match is often the most abundant protein in the sample.

Peptide Ions Score:

Mascot assigns a raw score for a peptide match, which takes into account the number of peaks in the spectrum that match predicted fragments based on a peptide sequence in the database. As noted above, the number of unassigned peaks (noise) in the spectrum is taken into account, because random peaks due to noise can enhance or produce a coincidental, random match. The assigned “ions score” can then be compared to a threshold score value which Mascot provides in search results. The threshold value is actually derived by a well defined statistical equation (shown below). Although the exact ions score calculation is a proprietary method, the number (or percentage) of the more intense (taller) peaks in a spectrum that correspond to predicted peptide fragment values is an important consideration in the scoring algorithm, which is reasonable because minor peaks are often just detector noise.

Reference the following publication in their thesis or peer reviewed publication.

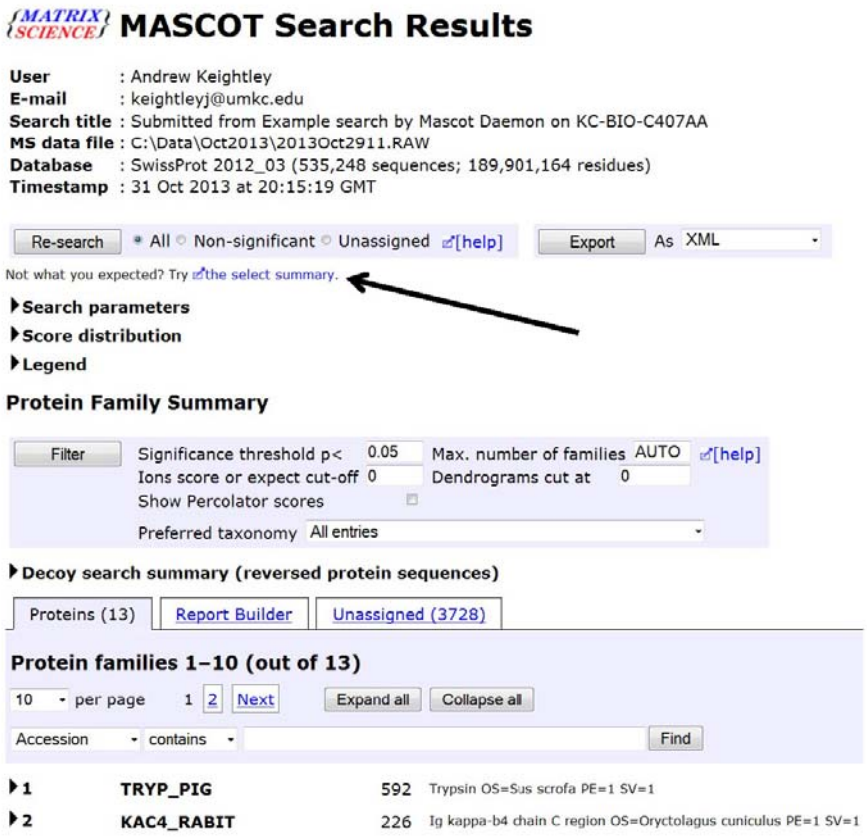
This paper describes the Mascot statistical approach for protein identification:

(Electrophoresis. 1999 Dec;20(18):3551-67).

MASCOT Search Results Interpretation

Opening Search results:
You may open Mascot MS/MS ions search results from a direct link provided in email, or from within the search program (not described, but you may learn how to run searches by request). The first screen that shows up is the Protein Family Summary (example shown below), which has a protein group dendrograms results format for viewing results in terms of protein family, which you may be interested in exploring. There are a few key lines at the top showing User, Search title, LC tandem MS data file searched (your data), the database searched (and the number of protein sequences in the database), and a timestamp. You may expand right facing triangles to view Search parameters (and Score distribution, Legend), or the protein family dendrograms.

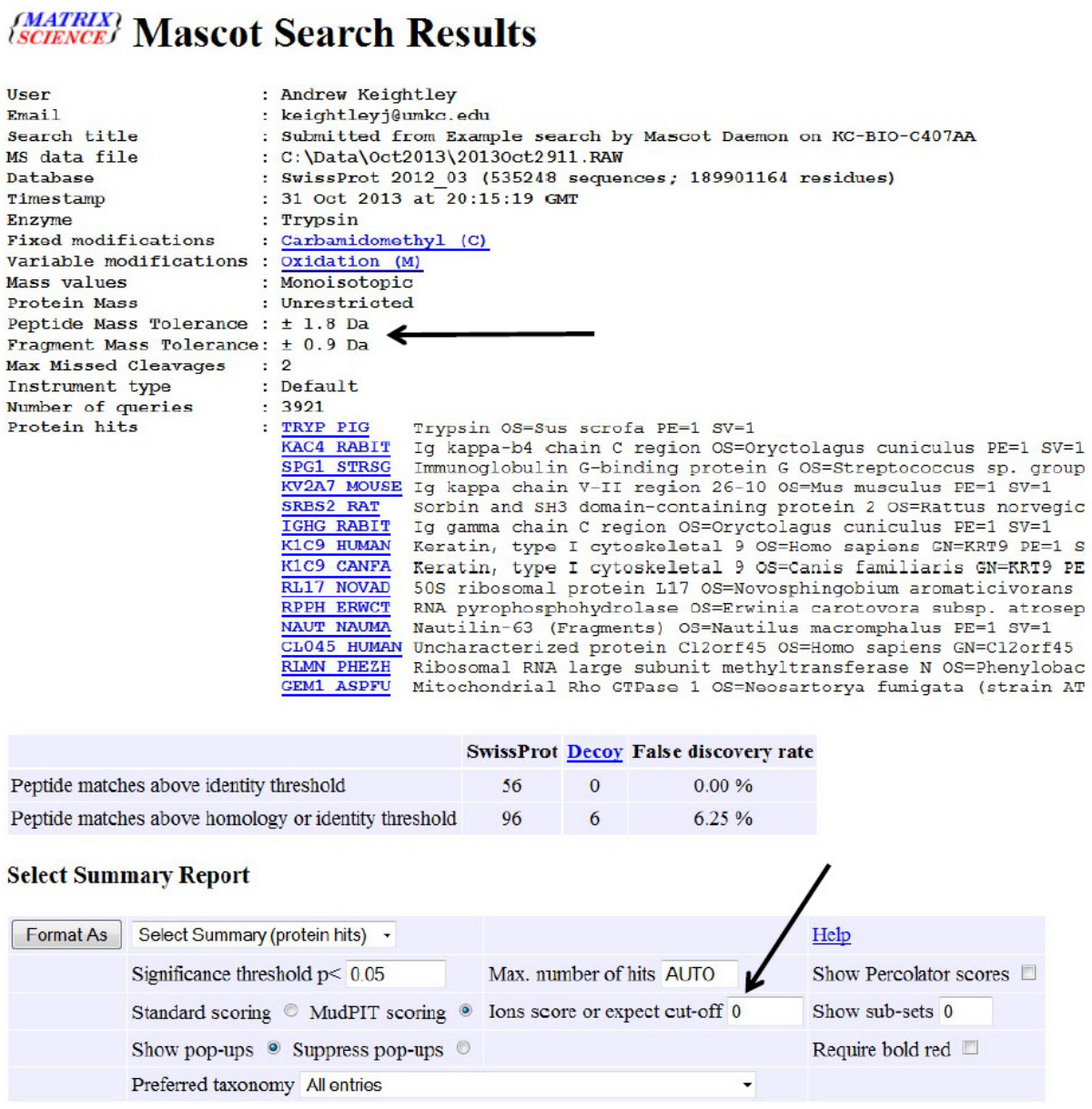
For the purposes of this search results interpretation guide, click on the "the select summary" (arrow points to this in the figure below). This will open the select summary, which we will discuss in the following sections below.



The Select Summary page opens (next page). In addition to the information shown on the Protein Family formatted report, Select Summary Report lists additional search parameters, including the enzyme you specified for the search (Trypsin is typically used), fixed and variable modifications (if selected), the peptide mass tolerance and fragment mass tolerance used in the search, missed cleavage (sometimes trypsin doesn't hydrolyze peptides next to acidic residues, prolines, etc), instrument type, and the number of queries (this is the number of MS/MS scans that were selected for searching: a pre-filter that limits the search to "good" MS/MS data).

MASCOT Search Results Interpretation

The first thing to do is limit the MS/MS matches to reasonable ones by increasing the ions score cut-off value from zero to at least 20. The lower arrow points to this. Highlight the field and replace the zero with 20, and click the format as button. It takes minute to reload.



Mascot Search Results

User	: Andrew Keightley
Email	: keightleyj@umkc.edu
Search title	: Submitted from Example search by Mascot Daemon on KC-BIO-C407AA
MS data file	: C:\Data\Oct2013\20130ct2911.RAW
Database	: SwissProt 2012_03 (535248 sequences; 189901164 residues)
Timestamp	: 31 Oct 2013 at 20:15:19 GMT
Enzyme	: Trypsin
Fixed modifications	: Carbamidomethyl (C)
Variable modifications	: Oxidation (M)
Mass values	: Monoisotopic
Protein Mass	: Unrestricted
Peptide Mass Tolerance	: ± 1.8 Da
Fragment Mass Tolerance	: ± 0.9 Da
Max Missed Cleavages	: 2
Instrument type	: Default
Number of queries	: 3921
Protein hits	: TRYP_PIG Trypsin OS=Sus scrofa PE=1 SV=1
	KAC4_RABIT Ig kappa-b4 chain C region OS=Oryctolagus cuniculus PE=1 SV=1
	SPG1_STRSG Immunoglobulin G-binding protein G OS=Streptococcus sp. group
	KV2A7_MOUSE Ig kappa chain V-II region 26-10 OS=Mus musculus PE=1 SV=1
	SRBS2_RAT Sorbin and SH3 domain-containing protein 2 OS=Rattus norvegic
	IGHG_RABIT Ig gamma chain C region OS=Oryctolagus cuniculus PE=1 SV=1
	K1C9_HUMAN Keratin, type I cytoskeletal 9 OS=Homo sapiens GN=KRT9 PE=1 S
	K1C9_CANFA Keratin, type I cytoskeletal 9 OS=Canis familiaris GN=KRT9 PE
	RL17_NOVAD 50S ribosomal protein L17 OS=Novosphingobium aromaticivorans
	RPPH_ERWCT RNA pyrophosphohydrolase OS=Erwinia carotovora subsp. atrosep
	NAUT_NAUMA Nautilin-63 (Fragments) OS=Nautilus macromphalus PE=1 SV=1
	CL045_HUMAN Uncharacterized protein C12orf45 OS=Homo sapiens GN=C12orf45
	RLMN_PHEZH Ribosomal RNA large subunit methyltransferase N OS=Phenylobac
	GEM1_ASPFU Mitochondrial Rho GTPase 1 OS=Neosartorya fumigata (strain AT

	SwissProt	Decoy	False discovery rate
Peptide matches above identity threshold	56	0	0.00 %
Peptide matches above homology or identity threshold	96	6	6.25 %

Select Summary Report

Format As	Select Summary (protein hits)		Help
	Significance threshold p< 0.05	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
	Preferred taxonomy All entries		

Did you notice that a decoy database search was included in this search? After the page is reloaded with increased ions score cutoff, you can begin validation of search results. If your purpose is to simply identify the most abundant proteins in the sample, and there are one or a few proteins matches with high scores (1000 or more), you may just print the first couple of pages of the Select Summary. But if you need to decide whether weaker identifications are real, you need to validate them by inspecting individual peptide MS/MS matches to see if they are acceptable.

MASCOT Search Results Interpretation

Do this by looking at the statistical information (compare threshold value to ions score, and by inspecting the MS/MS ions matching in peptide view (we'll go over this).

The identified proteins begin just below what is shown on the image on the previous page (example, next page). Threshold values are shown when you mouse over the query number for a peptide match: This is described below.
Read the following sections. If you have search results, you may follow this using your results.

Threshold Value:
Mascot offers this statement in the help section online on the Mascot Server.

> "In Mascot, the score for an MS/MS match is based on the absolute probability that the observed match between the experimental data and the database is a random event. The reported score is -10Log(P). So, during a search, if 1.5 x 10^5 peptides fell within the mass tolerance window about the precursor mass, and the significance threshold was chosen to be 0.05 (a 1 in 20 chance of a false positive), this would translate to a score threshold of 65."

The ions score for a peptide match incorporates matched fragment ions, the intensity of the matched ions (low intensity matched fragments might be noise), how many missing fragments are there, is there a lot of noise in the spectrum, etc). But the threshold is calculated simply based on how many peptides in the database have the mass in question (within your search tolerance).
Here is a summary of the actual calculation for the Threshold value:

$$\textbf{Threshold value (64.77) = (-10) times log (P), where P= } \frac{1}{20(150{,}000)}$$

SO, if there are more peptides in the database that have the same mass (within search tolerance), the threshold is pushed higher, and a given peptide match needs a higher "score" from Mascot to reach the confidence level of 95% (a 1 in 20 chance of being a random match).
For instance, if there are 3.0×10^5 peptides,

$$\textbf{Threshold value (67.78) = (-10) times log (P), where P= } \frac{1}{20(300{,}000)}$$

What if there are fewer peptides in the database that have the same mass?

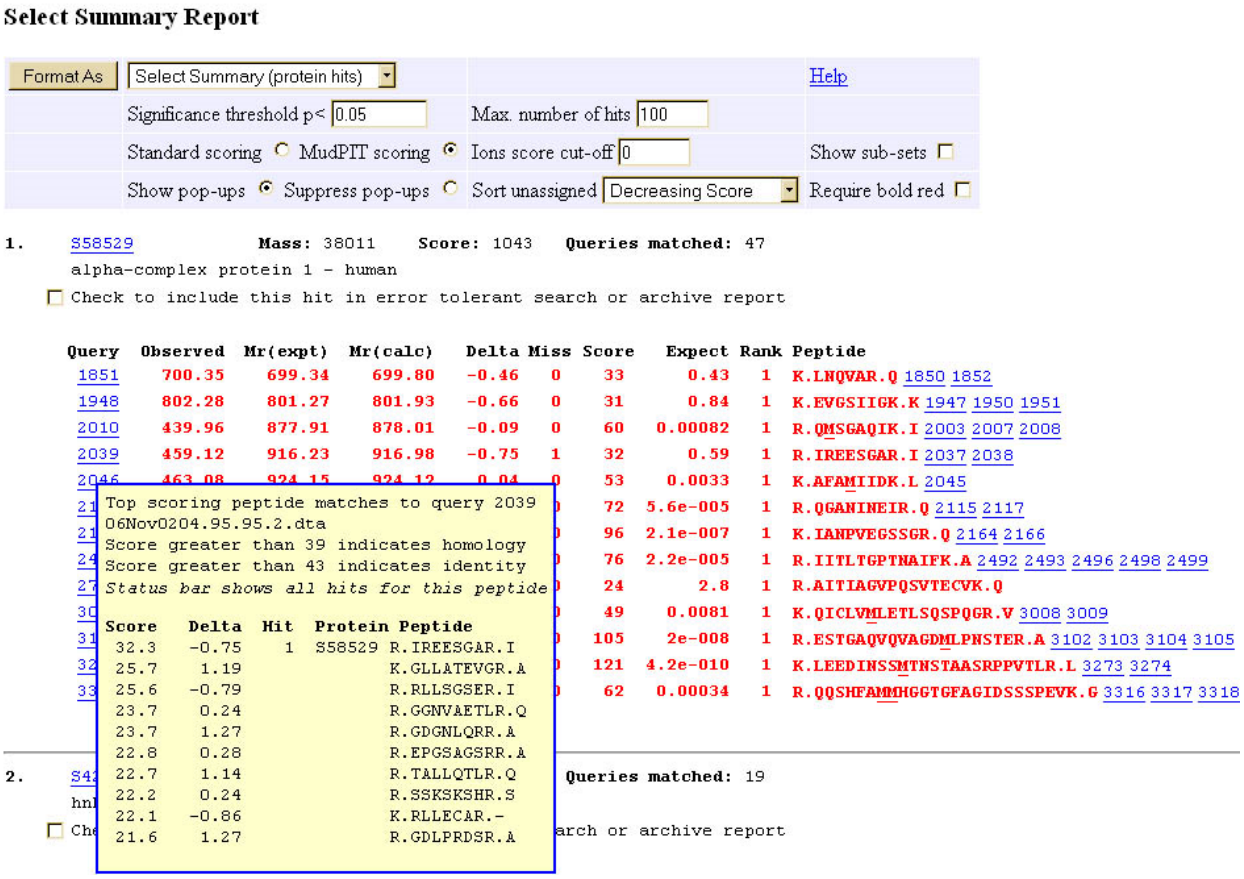
$$\textbf{Threshold value (54.77) = (-10) times log (P), where P= } \frac{1}{20(15{,}000)}$$

That's a lower threshold-fewer potential random matches results in a lower statistical threshold.

Threshold values are calculated by Mascot, and appear when you "mouse over" the query number in search results main page. Below is a cropped screen shot from a Main Search Results Page:

The example below is from a different search, but the search had a good match for a protein called alpha-complex protein 1 – human (score of 1043- Can you find that score?). Do you see where it shows that 47 queries were matched from the data to this protein? FYI- each separate matched MS/MS scan is a query in the search, shown as a blue number (link to peptide view).

MASCOT Search Results Interpretation



This example is formatted as Select Summary (protein hits). Do you see where this is selected? There is an alternative format called peptide summary as well. Significance threshold is 0.05 (default). The yellow box was invoked by mousing over (not clicking) the blue hyperlink query 2039. In the yellow box, Mascot shows the dta file (scan number) that the match is based on (the query), and lists the top scoring peptides from the search. The threshold is 43 "Score greater than 43 indicates identity". This peptide received the best score in the database, and yet it only got a score of 32. It may actually be the correct match, but one would not base an identification on this alone. In this case, the score for many other matching peptides DO meet or exceed the threshold value. Overall protein score of 1043 for this database entry is described by Mascot as being essentially the sum of the peptide scores. The other statement regarding "homology" (greater than 39) is addressed in the Mascot help text. The 'homology' threshold is not always provided. EACH blue query number represents a different scan! There are thousands of scans in some data files, hundreds in others. More than one scan (as you can see) can match the peptide. The blue hyperlink query number at the left is the best scoring of all of the scans for that peptide (with others listed at the right, if they exist), and THAT best score is the one shown in the score column. Inspect this partial screenshot carefully-find the scores in the yellow box. Which score corresponds to the score for this peptide in the red bold list of peptides. Can you find the score? Can you find other query matches for the peptide (blue links -numbers)?
Note: You can change the threshold calculation from the search results page, though the default value (0.05) appears to be acceptable by most journals/editors. To do this, just enter a different value and click the format as button. You can also go to a peptide view to validate a peptide match by clicking on the blue query number of interest.

MASCOT Search Results Interpretation

About the threshold calculation:
By mathematical definition, a smaller database will yield lower threshold values and therefore, potentially false confidence based on the simple fact that there were fewer peptides with the same mass (within tolerance). Therefore, you must disclose the database used (listed on the search result). You don't necessarily have to 'defend' your choices, but you should define them in your methods. Searching a larger database will therefore make it harder to achieve matches that meet or exceed the threshold. NOTE: Overall protein score is essentially the summation of the peptide scores, but matches are often described in terms of how many matching peptides were found, which meet or exceed the threshold. The choice is yours, but if the aim is to publish, look at other papers in the same journal for guidance. You may be surprised at some of the statements you find in the literature. The important thing is that you need to be able to defend your protein assignments. Therefore, it is important to understand what you are claiming, regardless of what you may see in print!
For reporting purposes, you can use some variation of the following statement:
"Protein identifications in table 1 contain at least two peptide matches that meet or exceed the threshold values for 95% confidence level (a 1 in 20 chance that the match is random)."

Expect value:
There is an additional useful statistic that Mascot provides called the "expect" value for each peptide match, seen under the Expect header in the Main Mascot Search Results page. Matrix Science states,

> "…each ions score in an MS/MS search, is accompanied by an expectation value. This is the number of matches with equal or better scores that are expected to occur by chance alone. It is directly equivalent to the E-value in a Blast search result. For a score that is exactly on the default significance threshold, ($p<0.05$), the expectation value is also 0.05. Increase the score by 10 and the expectation value drops to 0.005. The lower the expectation value, the more significant the score."

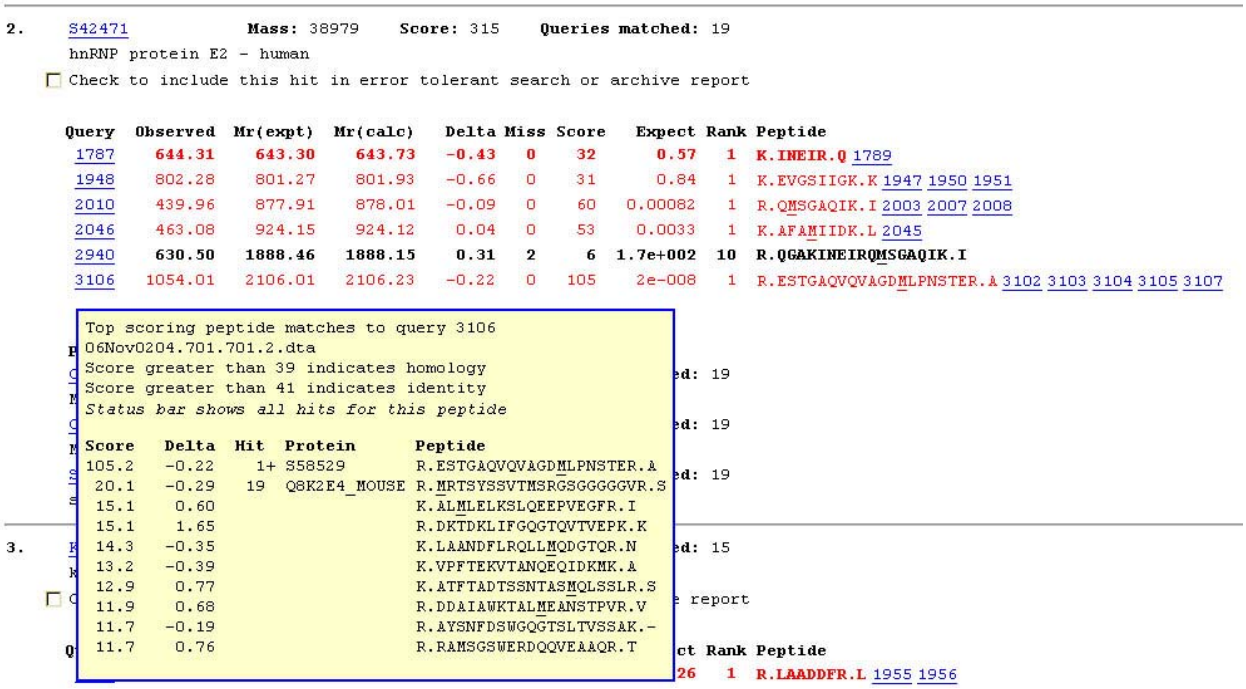
Refer initially to the partial screenshot for the protein identified called alpha-complex protein 1 – human on the previous page while reading this next section.

Red, Black, bold, not bold- what does this mean?

Red and Bold peptides listed under Peptide in the summary in alpha-complex protein 1 are the best matching peptide (red) in the database for the MS/MS scan, AND this is the first time this peptide appears in the Mascot Search results list (bold). The fact that a peptide does or does not meet or exceed the threshold is a separate question than the fact that it is the best peptide match in the database for a given scan (the query). Note that the peptide IREESGAR is the best matching peptide in the database for the scan (query 2039), and yet it does not meet or exceed the threshold. In the example below, a related protein which has some peptides in common with the number one hit is shown.This match was number 2 after alpha-complex protein 1 in that search. There are more unique peptide matches in alpha-complex protein 1 that make it the clear best match, but one might look for the same peptides found in a related protein further down the list that indicate the presence of another isoform sharing a conserved peptide. Do you see the red text (NOT bold) matches in the second protein match (below)? These already

MASCOT Search Results Interpretation

appeared in the first protein match above, so they are not bold. Do you see that they are bold in the 1st match (alpha-complex protein 1)? And note that the second protein identified has a novel peptide match (INEIR). It is red and bold. If that is a good match, both proteins may be present. The black text peptide is not the best peptide in the database for the indicated query (MS/MS scan), but it is the first time it appears in the list, so it is bold. SO, red peptide=best peptide in database for the scan, bold peptide=first time appears in the list. Black peptide= not the best peptide in the database for the scan, and 'not' bold,=not the first time the peptide appears in the list. When you are scanning down through the results, Red bold peptides DO get your attention, so it is helpful. Note that the best peptide match in the yellow box below refers to the first (best matching) protein in the list (S58529), though it is present in the second protein as well! Apparently the protein field in the yellow box is filled with the best protein match.



2. S42471 **Mass:** 38979 **Score:** 315 **Queries matched:** 19

hnRNP protein E2 - 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
1787	644.31	643.30	643.73	-0.43	0	32	0.57	1	K.INEIR.Q 1789
1948	802.28	801.27	801.93	-0.66	0	31	0.84	1	K.EVGSIIGK.K 1947 1950 1951
2010	439.96	877.91	878.01	-0.09	0	60	0.00082	1	R.QMSGAQIK.I 2003 2007 2008
2046	463.08	924.15	924.12	0.04	0	53	0.0033	1	K.AFAMIIDK.L 2045
2940	630.50	1888.46	1888.15	0.31	2	6	1.7e+002	10	R.QGAKINEIRQMSGAQIK.I
3106	1054.01	2106.01	2106.23	-0.22	0	105	2e-008	1	R.ESTGAQVQVAGDMLPNSTER.A 3102 3103 3104 3105 3107

Top scoring peptide matches to query 3106
06Nov0204.701.701.2.dta
Score greater than 39 indicates homology
Score greater than 41 indicates identity
Status bar shows all hits for this peptide

Score	Delta	Hit	Protein	Peptide
105.2	-0.22	1+	S58529	R.ESTGAQVQVAGDMLPNSTER.A
20.1	-0.29	19	Q8K2E4_MOUSE	R.MRTSYSSVTMSRGSGGGGGVR.S
15.1	0.60			K.ALMLELKSLQEEPVEGFR.I
15.1	1.65			R.DKTDKLIFGQGTQVTVEPK.K
14.3	-0.35			K.LAANDFLRQLLMQDGTQR.N
13.2	-0.39			K.VPFTEKVTANQEQIDKMK.A
12.9	0.77			K.ATFTADTSSNTASMQLSSLR.S
11.9	0.68			R.DDAIAWKTALMEANSTPVR.V
11.7	-0.19			R.AYSNFDSWGQGTSLTVSSAK.-
11.7	0.76			R.RAMSGSWERDQQVEAAQR.T

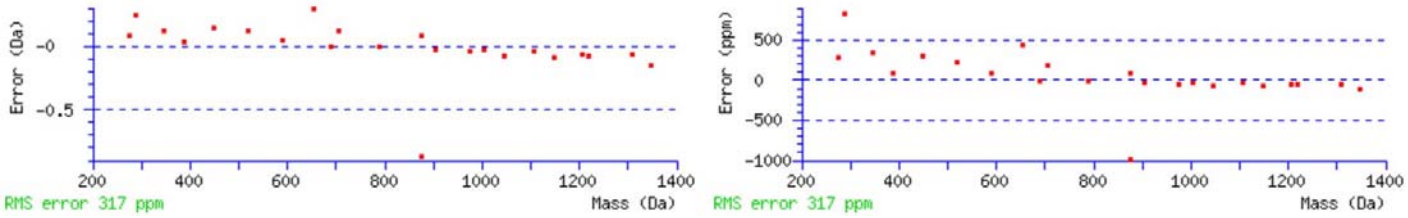
3. R.LAADDFR.L 1955 1956

Error bias is another very useful way to evaluation the quality of a peptide MS/MS match

Error Bias:

Mass Spectrometers have different mass accuracy (ability to correctly assign mass). In fact, error bias is consistent on any given mass spectrometer. Masses assigned fall along a calibration curve which is distinct and recognizable. This is very useful in validating peptide matches resulting from database searches. The user-definable mass tolerance of a search (+/- 0.5 Da, for instance) should be set based on the type of mass spectrometer used, and knowledge of the performance in the accuracy of mass assignments for that instrument. Error bias and adherence to a mass spectrometer's calibration curve is NOT considered in calculating the ions score. The peaks either are, or are not within the tolerances you have set for the search. However, the error bias is shown in the Peptide view screen as a plot of simple subtracted differences in assigned versus calculated mass for each matching peak. Since error bias is very consistent for each instrument, this plot provides extremely useful visible evidence for validation. Peptide view is invoked by clicking once on a query number from the Main Mascot Search Results page, or on the ions score hyperlink from the Protein View page. The error bias is reflected in a fairly constant error curve

visible in the plot. Look at this plot for good MS/MS matches (high scoring peptide matches) to become familiar with the error bias that your mass spectrometer is typically generating. This can be useful when validating the less convincing matches, particularly when you have few peptides matched to the data. Randomly distributed error, rather than an ordered cluster following the error bias, should cause concern. Not every peptide MS/MS match has a nice cluster of scattered error following the calibration curve like this example, but this illustrates the concept.



Mascot shows the absolute error in Daltons (left), which is the subtracted difference between the assigned mass and the calculated mass of each matching ion peak in the MS/MS spectrum. The graph on the right shows the difference in parts per million (a mass proportional plot). Note the outlying assigned peak at about 875Da. This may be due to a peak (detector noise) which randomly fell within search mass tolerance of a predicted peak for this peptide MS/MS which was assigned to that peak. Mascot autoscales the y-range to fit the point with largest error (about 1 Dalton (-1) on the left, and about 1000 ppm on the right).

In summary

There are a few key statistics presented by Mascot to help you evaluate protein identifications. Mascot search results main page shows the best scoring proteins on top, and summarizes the peptide matches that form the basis of the identifications right there with the protein. An overwhelmingly high protein score with dozens of red/bold peptide matches can be immediately believed, though you may need to look in other protein matches below that one for isoform/spliceform specific peptide matches, if they exist. These would be red/bold peptides that stand out clearly. Also, don't forget the obvious things like molecular weight, species (remember that many species have VERY similar protein sequences). Are peptides that should appear in the data if cut with trypsin absent (between 5 and 25 amino acids in length)? Observe where the peptides are in the entry. Are the matched peptides spread throughout the protein entry, are they all in the N-terminal or C-terminal region only? Processing? Degradation? And remember that database accessions include signal peptides that are removed, so they won't be present.

More modestly scoring proteins are the ones that need scrutiny. A couple of good peptide matches that exceed threshold are usually enough. Three is much better. Check the Threshold values to see which peptides meet or exceed them in their score. Look at the Expect values, which should be less than zero. Follow the link to the Protein View and follow the ions score links to see the Peptide View, (or open the Peptide view by clicking the query number from the main search results page). Look at the error analysis to see that the mass assignments are clustered along the calibration curve for the instrument. If you obtain more than one peptide whose scores meet or exceed threshold value, expect values are low, with adherence to the typical mass assignment bias in the error analysis graphic, you probably have an identification.