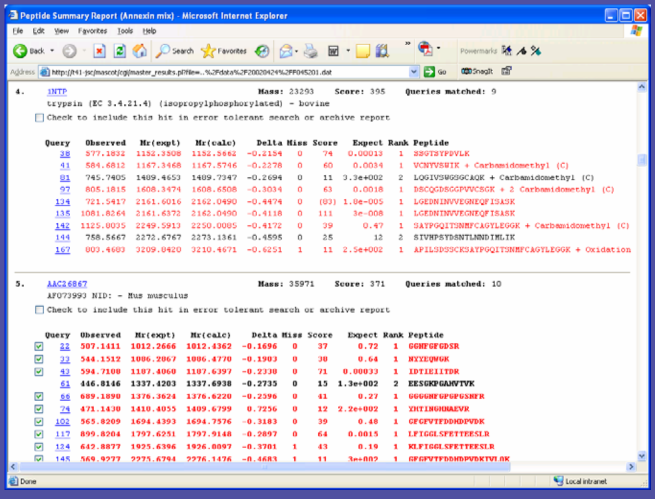
Mascot Search Results FAQ

ASMS 2005

MATRIX SCIENCE

Why are some peptide matches shown in red or bold face?



Peptide Summary Report (Annexin mix) - Microsoft Internet Explorer

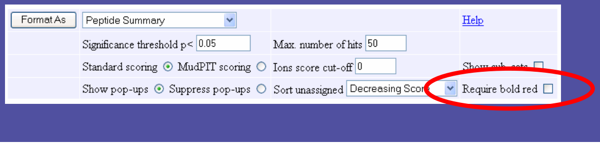
Address http://t41-jsc/mascot/cgi/master_results.pl?file=...%2Fdata%2F20020424%2FF045201.dat

```
4.     1NTP                                    Mass: 23293    Score: 395    Queries matched: 9
       trypsin (EC 3.4.21.4) (isopropylphosphorylated) - bovine
     □ Check to include this hit in error tolerant search or archive report

       Query  Observed   Mr(expt)   Mr(calc)    Delta Miss Score   Expect Rank Peptide
          38  577.1832  1152.3508  1152.5662  -0.2154    0    74  0.00013    1 SSGTSYPDVLK
          41  584.6812  1167.3468  1167.5746  -0.2278    0    60   0.0034    1 VCNYVSWIK + Carbamidomethyl (C)
          81  745.7405  1489.4653  1489.7347  -0.2694    0    11  3.3e+002   2 LQGIVSWGSGCAQK + Carbamidomethyl (C)
          97  805.1815  1608.3474  1608.6508  -0.3034    0    63   0.0018    1 DSCQGDSGGPVVCSGK + 2 Carbamidomethyl (C)
         134  721.5417  2161.6016  2162.0490  -0.4474    0  (83) 1.0e-005   1 LGEDNINVVEGNEQFISASK
         135 1081.8264  2161.6372  2162.0490  -0.4118    0   111    3e-008   1 LGEDNINVVEGNEQFISASK
         142 1125.8035  2249.5913  2250.0085  -0.4172    0    39     0.47   1 SAYPGQITSNMFCAGYLEGGK + Carbamidomethyl (C)
         144  758.5667  2272.6767  2273.1361  -0.4595    0    25       12   2 SIVHPSYDSNTLNNDIMLIK
         167  803.4683  3209.8420  3210.4671  -0.6251    1    11  2.5e+002   1 APILSDSSCKSAYPGQITSNMFCAGYLEGGK + Oxidation

5.     AAC26867                                Mass: 35971    Score: 371    Queries matched: 10
       AF073993 NID: - Mus musculus
     □ Check to include this hit in error tolerant search or archive report

       Query Observed   Mr(expt)   Mr(calc)    Delta Miss Score   Expect Rank Peptide
     ☑    22 507.1411  1012.2666  1012.4362  -0.1696    0    37     0.72    1 GGHFGFGDSR
     ☑    33 544.1512  1086.2867  1086.4770  -0.1903    0    38     0.64    1 NYYEQWGK
     ☑    43 594.7108  1187.4060  1187.6397  -0.2338    0    71  0.00033    1 IDTIEIITDR
          61 446.8146  1337.4203  1337.6938  -0.2735    0    15  1.3e+002   2 EESGKPGAHVTVK
     ☑    66 689.1890  1376.3624  1376.6220  -0.2596    0    41     0.27    1 GGGGNFGPGPGSNFR
     ☑    74 471.1430  1410.4055  1409.6799   0.7256    0    12  2.2e+002   1 YHTINGHNAEVR
     ☑   102 565.8209  1694.4393  1694.7576  -0.3183    0    39     0.48    1 GFGFVTFDDHDPVDK
     ☑   117 899.8204  1797.6251  1797.9148  -0.2897    0    64   0.0015    1 LFIGGLSFETTEESLR
     ☑   124 642.8877  1925.6396  1926.0097  -0.3701    1    43     0.19    1 KLFIGGLSFETTEESLR
     ☑   145 569.9277  2275.6794  2276.1476  -0.4683    1    11    3e+002    1 GFGFVTFDDHDPVDKIVLQK
```

Done

Why are some peptide matches shown in red or bold face?

- Red indicates the top scoring peptide match for this spectrum
  - Not necessarily a significant match
- Bold indicates the first time any match to this spectrum has appeared in the report
  - So, if a protein hit *doesn't* have any bold red matches, all the assigned spectra have better scoring matches elsewhere or the same matches have already appeared in the report, assigned to higher scoring protein(s)



Peptide Summary

Significance threshold p< 0.05 Max. number of hits 50

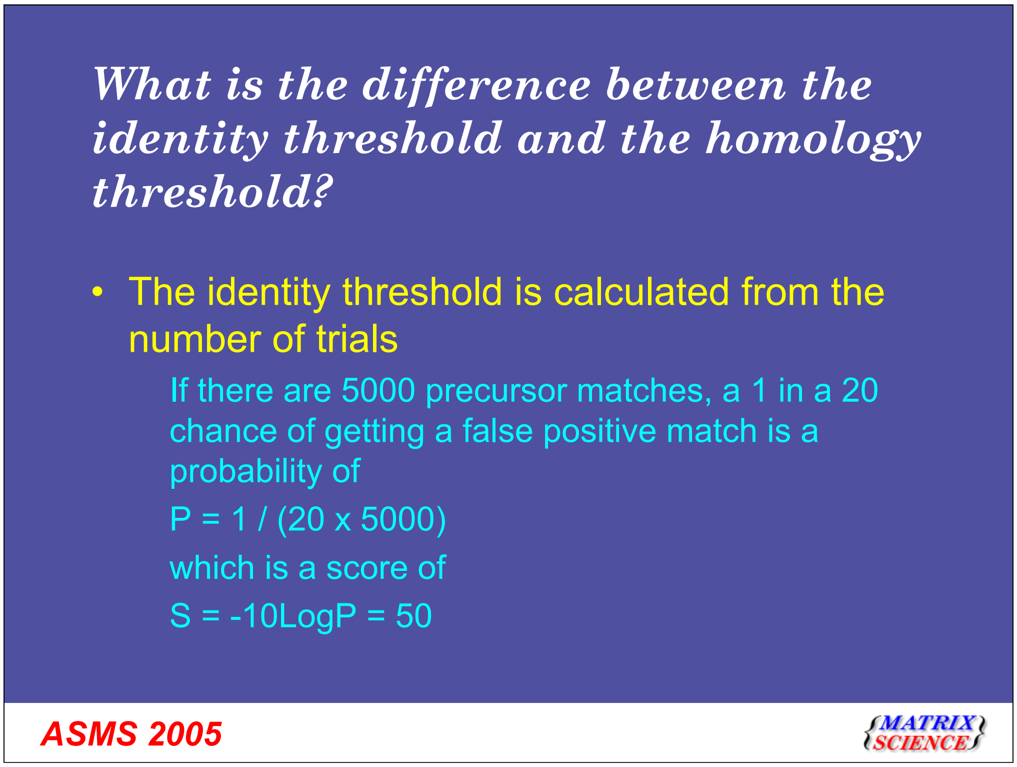
Standard scoring ● MudPIT scoring ○ Ions score cut-off 0 Show sub-sets □

Show pop-ups ● Suppress pop-ups ○ Sort unassigned Decreasing Score Require bold red □

ASMS 2005

MATRIX SCIENCE

Interpretation of the results from an LC-MS/MS search can be complex, because it is not always clear which peptide "belongs" to which protein. The use of red and bold typefaces is intended to highlight the most logical assignment of peptides to proteins. The first time a match to a spectrum appears in the report, it is shown in bold face. Whenever the top scoring peptide match for a spectrum appears, it is shown in red. This means that peptide matches which are both bold and red are the most likely assignments of the best matches. Conversely, if a protein hit *doesn't* have any bold red matches, all the assigned spectra have better scoring matches elsewhere or the same matches have already appeared in the report, assigned to higher scoring protein(s). This means that the protein hit is likely to be spurious, and would collapse into a higher scoring hit except for the presence of one or more weak, noisy matches. Such hits can be filtered out of the report by ticking the 'require bold red' checkbox.



What is the difference between the identity threshold and the homology threshold?

- The identity threshold is calculated from the number of trials

  If there are 5000 precursor matches, a 1 in a 20 chance of getting a false positive match is a probability of

  P = 1 / (20 x 5000)

  which is a score of

  S = -10LogP = 50

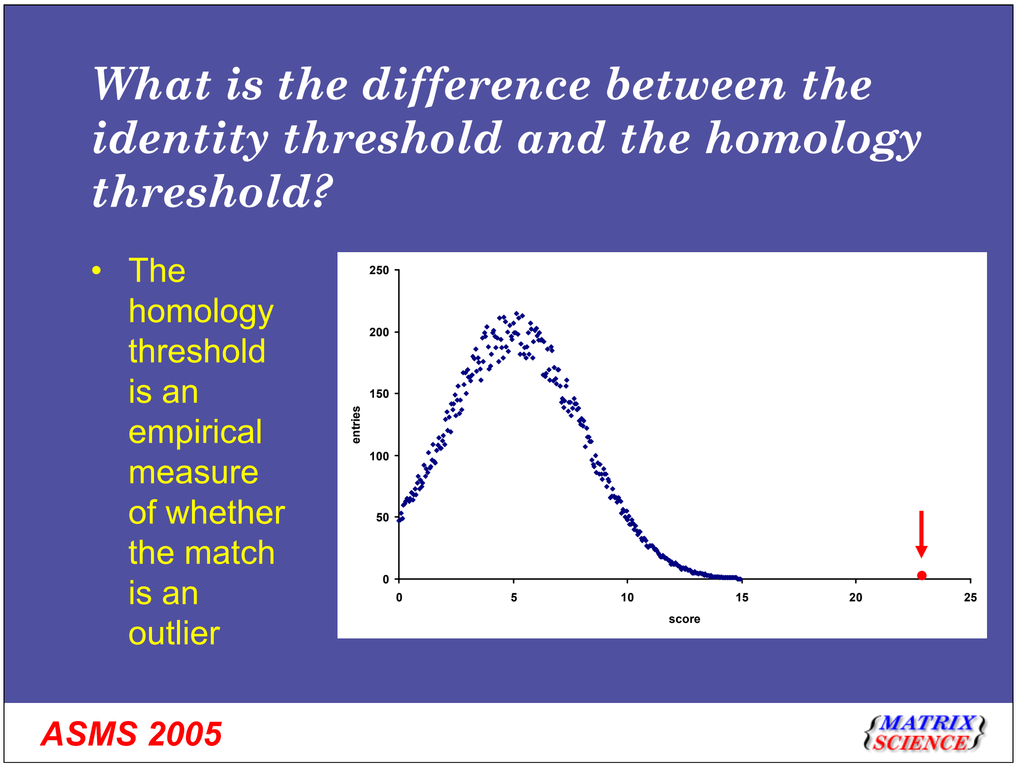
ASMS 2005

MATRIX SCIENCE

If a score is a true probability, then assigning a significance threshold is very simple ... its just a function of the number of trials - the number of times we test for a match.

If we are comfortable with a 1 in a 20 chance of getting a false positive match, and we are doing a MS/MS search of a database that contains 5000 peptides that fit to the precursor molecular weight , then we are looking for a probability of less than 1 / (20 x 5000) which is a Mascot score of 50

If we could only tolerate a false positive rate of 1 in 200 then the threshold would be 60, 1 in 2000 70, etc.



What is the difference between the identity threshold and the homology threshold?

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

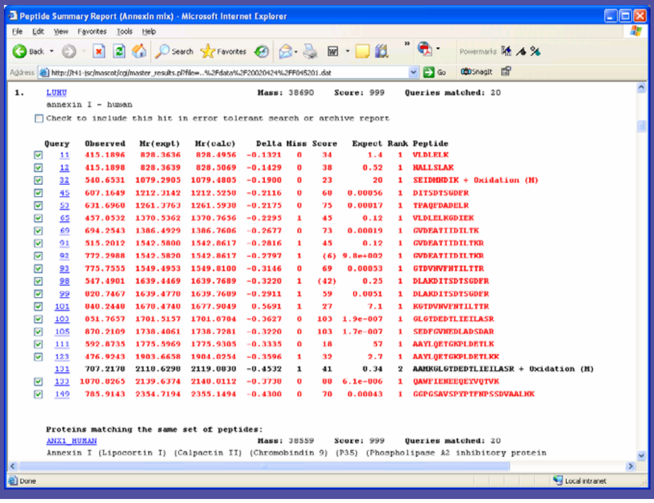
ASMS 2005

MATRIX SCIENCE

Unfortunately, mass 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 this statistical 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, 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 additional true positive matches without exceeding the specified false positive rate.

What is an expectation value?


Peptide Summary Report (Annexin mix) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Powermarks
Address http://t41-jsc/mascot/cgi/master_results.pl?file=../%2Fdata%2F20020424%2FF045201.dat
Go
SnagIt
Done
Local intranet

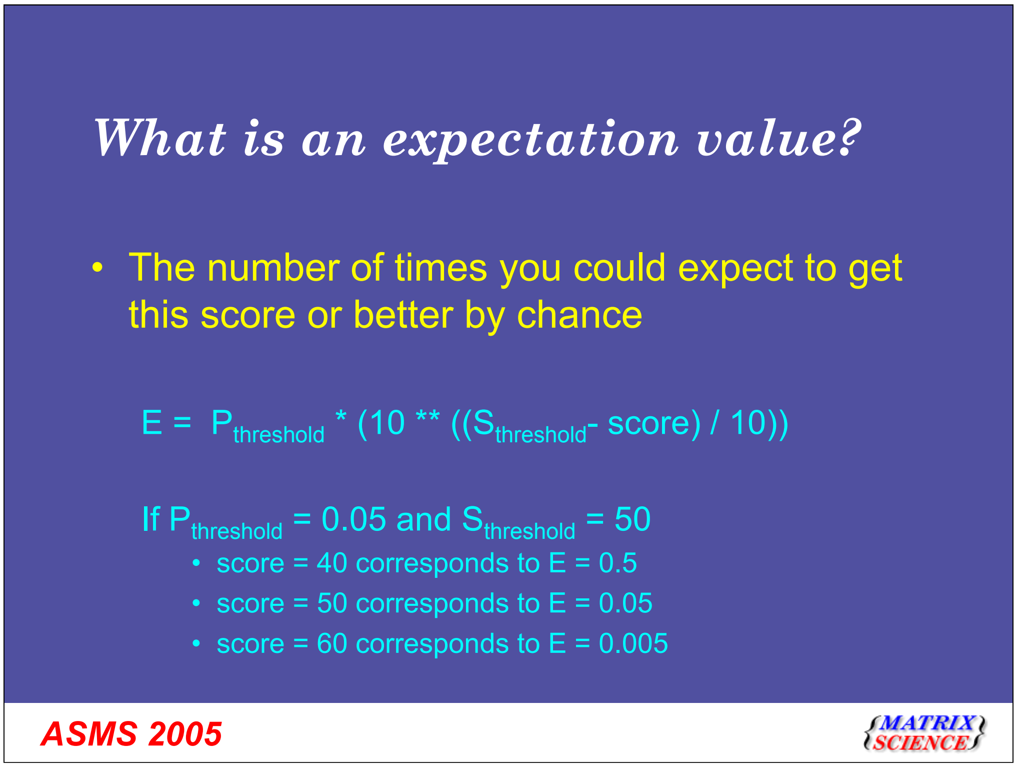
1. LUHU **Mass:** 38690 **Score:** 999 **Queries matched:** 20
annexin I - human

□ Check to include this hit in error tolerant search or archive report

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑	11	415.1896	828.3636	828.4956	-0.1321	0	34	1.4	1	VLDLELK
☑	12	415.1898	828.3639	828.5069	-0.1429	0	38	0.52	1	NALLSLAK
☑	32	540.6531	1079.2905	1079.4805	-0.1900	0	23	20	1	SEIDMNDIK + Oxidation (M)
☑	45	607.1649	1212.3142	1212.5258	-0.2116	0	68	0.00056	1	DITSDTSGDFR
☑	53	631.6960	1261.3763	1261.5938	-0.2175	0	75	0.00017	1	TPAQFDADELR
☑	65	457.0532	1370.5362	1370.7656	-0.2295	1	45	0.12	1	VLDLELKGDIEK
☑	69	694.2543	1386.4929	1386.7606	-0.2677	0	73	0.00019	1	GVDEATIIDILTK
☑	91	515.2012	1542.5800	1542.8617	-0.2816	1	45	0.12	1	GVDEATIIDILTKR
☑	92	772.2988	1542.5820	1542.8617	-0.2797	1	(6)	9.8e+002	1	GVDEATIIDILTKR
☑	93	775.7555	1549.4953	1549.8100	-0.3146	0	69	0.00053	1	GTDVNVFNTILTTR
☑	98	547.4901	1639.4469	1639.7689	-0.3220	1	(42)	0.25	1	DLAKDITSDTSGDFR
☑	99	820.7467	1639.4778	1639.7689	-0.2911	1	59	0.0051	1	DLAKDITSDTSGDFR
☑	101	840.2440	1678.4740	1677.9049	0.5691	1	27	7.1	1	KGTDVNVFNTILTTR
☑	103	851.7657	1701.5157	1701.8784	-0.3627	0	103	1.9e-007	1	GLGTDEDTLIEILASR
☑	105	870.2109	1738.4061	1738.7281	-0.3220	0	103	1.7e-007	1	SEDFGVNEDLADSDAR
☑	111	592.8735	1775.5969	1775.9305	-0.3335	0	18	57	1	AAYLQETGKPLDETLK
☑	123	476.9243	1903.6658	1904.0254	-0.3596	1	32	2.7	1	AAYLQETGKPLDETLKK
	131	707.2178	2118.6298	2119.0830	-0.4532	1	41	0.34	2	AAMKGLGTDEDTLIEILASR + Oxidation (M)
☑	133	1070.8265	2139.6374	2140.0112	-0.3738	0	88	6.1e-006	1	QAWFIENEEQEYVQTVK
☑	149	785.9143	2354.7194	2355.1494	-0.4300	0	70	0.00043	1	GGPGSAVSPYPTFNPSSDVAALHK

Proteins matching the same set of peptides:

ANX1_HUMAN **Mass:** 38559 **Score:** 999 **Queries matched:** 20
Annexin I (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhibitory protein



What is an expectation value?

- 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

ASMS 2005

MATRIX SCIENCE

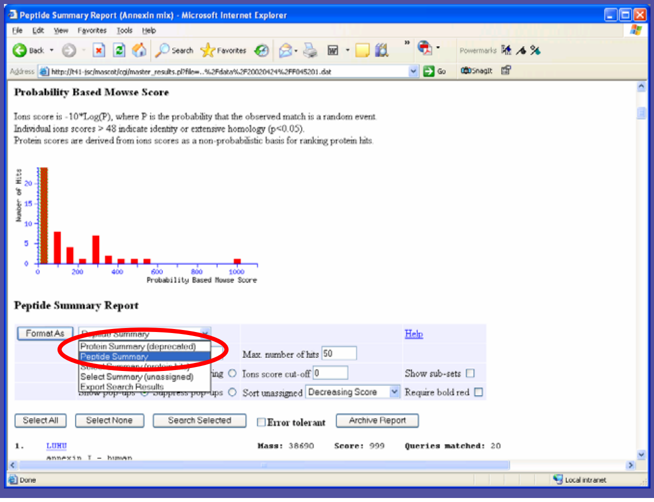
The expectation value does not contain new information. It can be derived directly from the score and the threshold. The advantage is that it tells you everything you need to know in a single number.

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

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

The better the match, the smaller the expectation value.

Why does it say that the Protein Summary is deprecated?


Peptide Summary Report (Annexin mix) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Powermarks
Address http://t41-jsc/mascot/cgi/master_results.pl?file=..%2Fdata%2F20020424%2FF045201.dat
Go
SnagIt
Number of Hits
0
5
10
15
20
0
200
400
600
800
1000
Probability Based Mowse Score
Format As
Peptide Summary
Protein Summary (deprecated)
Peptide Summary
Select Summary (unassigned)
Export Search Results
Help
Select All
Select None
Search Selected
Archive Report
Done
Local intranet

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 48 indicate identity or extensive homology (p<0.05).
Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.

Peptide Summary Report

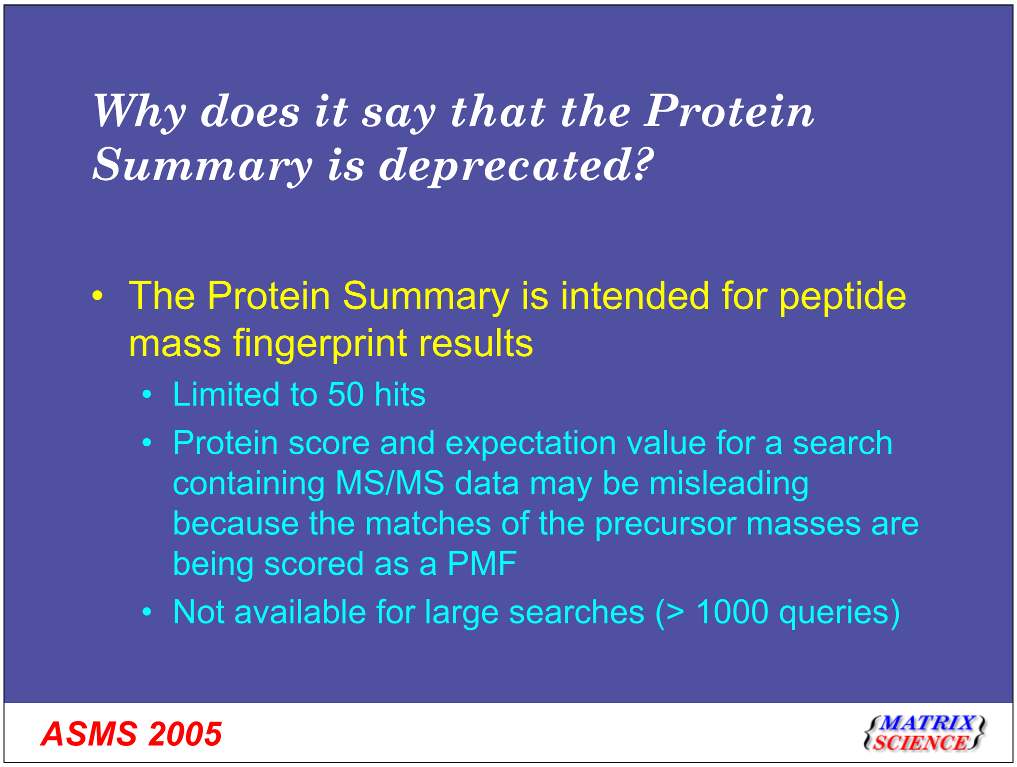
Max. number of hits 50

Ions score cut-off 0 Show sub-sets □

Sort unassigned Decreasing Score Require bold red □

□ Error tolerant

```
1.     LUHU                           Mass: 38690    Score: 999    Queries matched: 20
       annexin I - human
```



Why does it say that the Protein Summary is deprecated?

- The Protein Summary is intended for peptide mass fingerprint results
  - Limited to 50 hits
  - Protein score and expectation value for a search containing MS/MS data may be misleading because the matches of the precursor masses are being scored as a PMF
  - Not available for large searches (> 1000 queries)

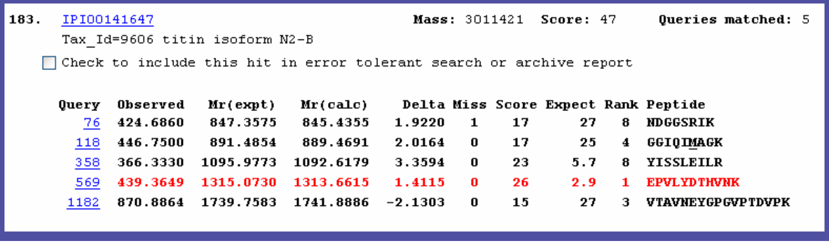
ASMS 2005

MATRIX SCIENCE

Peptide Summary is the default for any search containing MS/MS data. Unfortunately, some older, third party software specifies a Protein Summary report when submitting an MS/MS search. This means that you have to switch formats yourself. A lot of people didn’t realise this, so we have made the warnings more prominent.

What is MudPIT scoring?

- Standard protein score
  - the sum of the ions scores
  - excluding the scores for duplicate matches, which are shown in parentheses
  - correction to reduce the contribution of low-scoring random matches

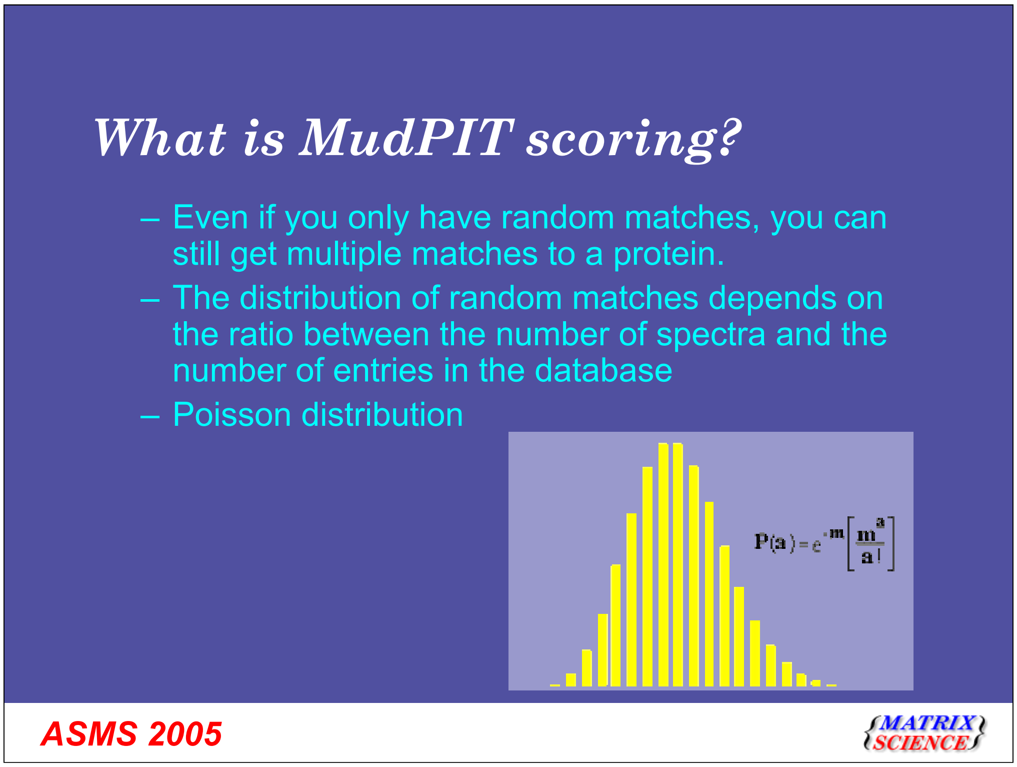


```
183.  IPI00141647                              Mass: 3011421  Score: 47     Queries matched: 5
      Tax_Id=9606 titin isoform N2-B
    □ Check to include this hit in error tolerant search or archive report

      Query  Observed   Mr(expt)   Mr(calc)    Delta Miss Score Expect Rank Peptide
         76  424.6860   847.3575   845.4355   1.9220    1    17     27    8 NDGGSRIK
        118  446.7500   891.4854   889.4691   2.0164    0    17     25    4 GGIQIMAGK
        358  366.3330  1095.9773  1092.6179   3.3594    0    23    5.7    8 YISSLEILR
        569  439.3649  1315.0730  1313.6615   1.4115    0    26    2.9    1 EPVLYDTHVNK
       1182  870.8864  1739.7583  1741.8886  -2.1303    0    15     27    3 VTAVNEYGPGVPTDVPK
```

ASMS 2005

With standard peptide summary scoring, the protein score is the sum of the ions scores of all the non-duplicate peptides. Where there are duplicate peptides, the highest scoring peptide is used.

This example shows how we can get a protein score of 47 even though none of the peptide matches are significant



What is MudPIT scoring?

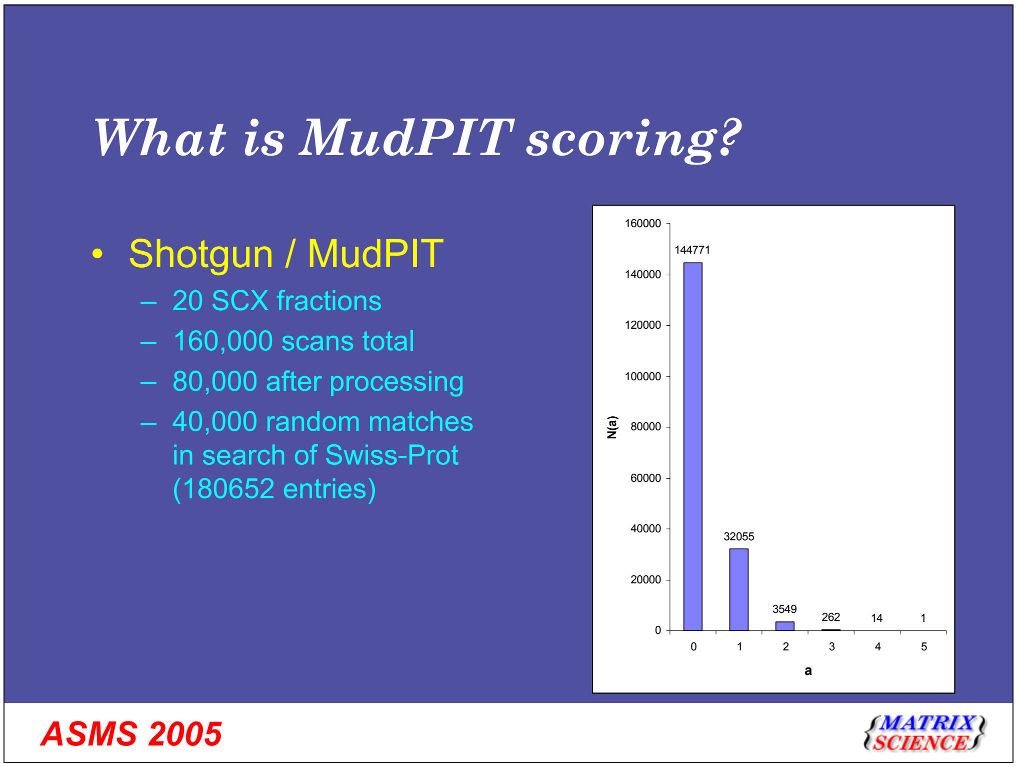
- Even if you only have random matches, you can still get multiple matches to a protein.
- The distribution of random matches depends on the ratio between the number of spectra and the number of entries in the database
- Poisson distribution

$$P(a) = e^{-m}\left[\frac{m^a}{a!}\right]$$

ASMS 2005

Even if peptide matches are random, you can still get multiple matches to a single protein. How likely this is depends on the ratio between the number of spectra and the number of entries in the database. We can predict whether this will be a serious problem or not using a function called a Poisson distribution.

If average number of events per interval is m, then the Poisson distribution gives us the probability of observing a events in a particular interval.



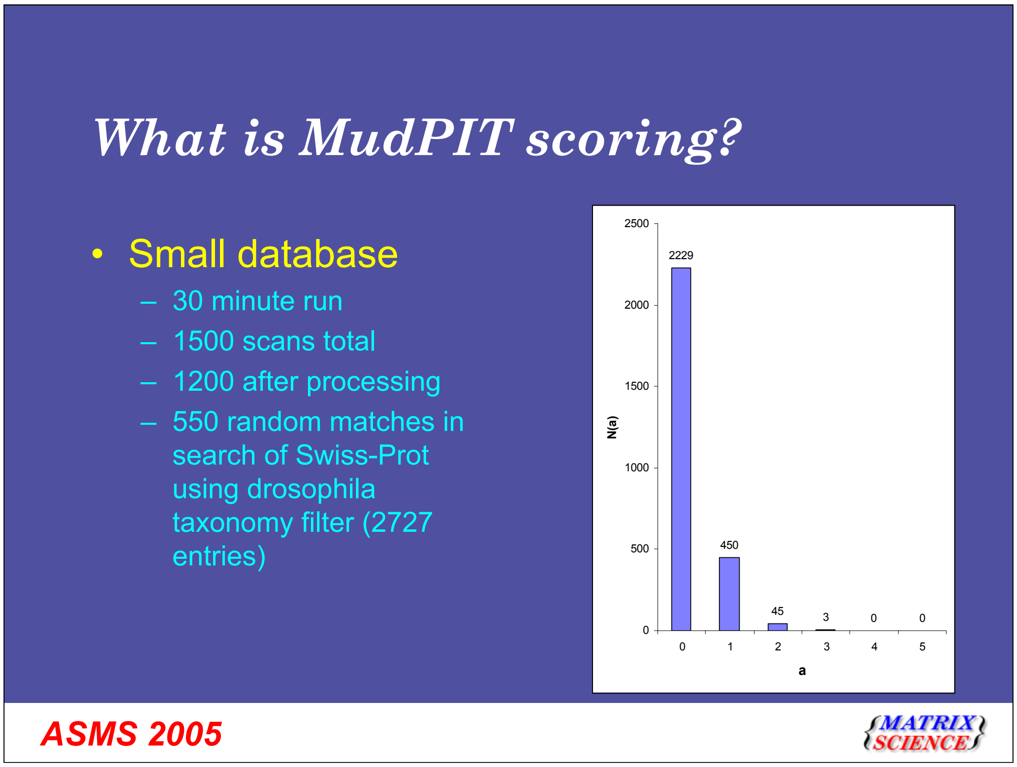
What is MudPIT scoring?

- Shotgun / MudPIT
  - 20 SCX fractions
  - 160,000 scans total
  - 80,000 after processing
  - 40,000 random matches in search of Swiss-Prot (180652 entries)

ASMS 2005

{MATRIX}
{SCIENCE}

For this MudPIT search, 262 proteins are expected to pick up 3 random matches by chance. 1 protein will pick up 5



What is MudPIT scoring?

- Small database
  - 30 minute run
  - 1500 scans total
  - 1200 after processing
  - 550 random matches in search of Swiss-Prot using drosophila taxonomy filter (2727 entries)

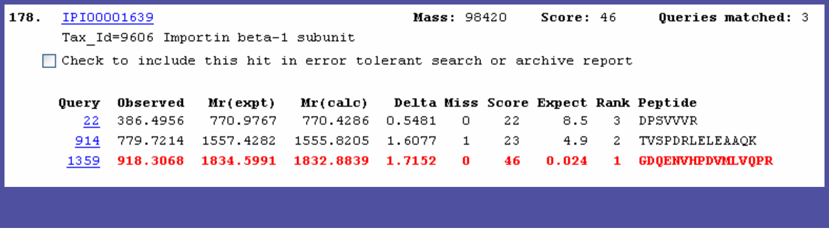
ASMS 2005

MATRIX SCIENCE

The problem isn’t limited to large searches. It is the ratio between the number of spectra and the number of entries in the database that matters. So, a small search against a small database can give similar numbers

What is MudPIT scoring?

- MudPIT protein score
  - The sum of the excess of the ions score over the identity or homology threshold for each query
  - Plus 1 x the average threshold

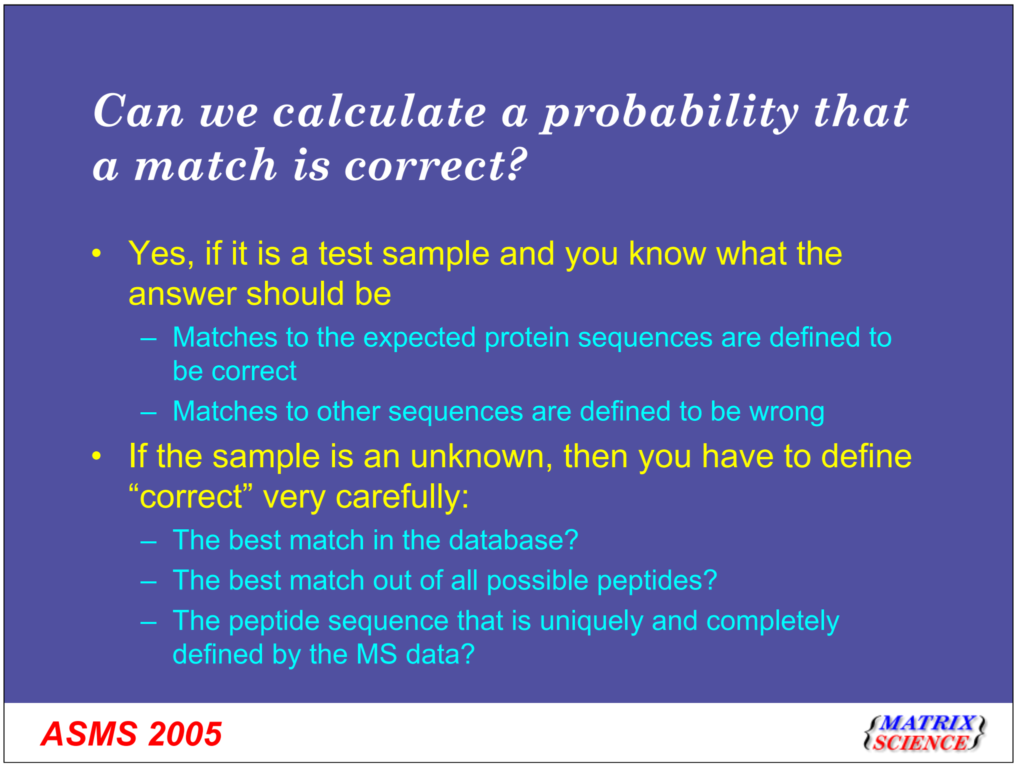


```
178.    IPI00001639                         Mass: 98420    Score: 46    Queries matched: 3
        Tax_Id=9606 Importin beta-1 subunit
      □ Check to include this hit in error tolerant search or archive report

       Query  Observed   Mr(expt)   Mr(calc)   Delta Miss Score Expect Rank Peptide
          22  386.4956   770.9767   770.4286  0.5481    0    22    8.5    3  DPSVVVR
         914  779.7214  1557.4282  1555.8205  1.6077    1    23    4.9    2  TVSPDRLELEAAQK
        1359  918.3068  1834.5991  1832.8839  1.7152    0    46  0.024    1  GDQENVHPDVMLVQPR
```

For MudPIT scoring, the score for each peptide is not its absolute score, but the amount that it is above the threshold. Therefore, peptides with a score below the threshold do not contribute to the score. Finally, the average of the thresholds used is added to the score. For each peptide, the "threshold" is the homology threshold if it exists, otherwise it is the identity threshold.

You shouldn't see proteins with a large number of weak peptide matches getting a good score. If there are no significant peptides, the protein score will be 0.

By default, MudPIT protein scoring is used for searches with more than 1000 spectra. You can also choose which scoring to use in the format controls at the top of the report



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:
  - The best match in the database?
  - The best match out of all possible peptides?
  - The peptide sequence that is uniquely and completely defined by the MS data?

ASMS 2005

Probability based scoring tells you the probability that the match is random. This is, the probability that the match is meaningless. Many people would prefer a probability that the match is correct. Is this possible?

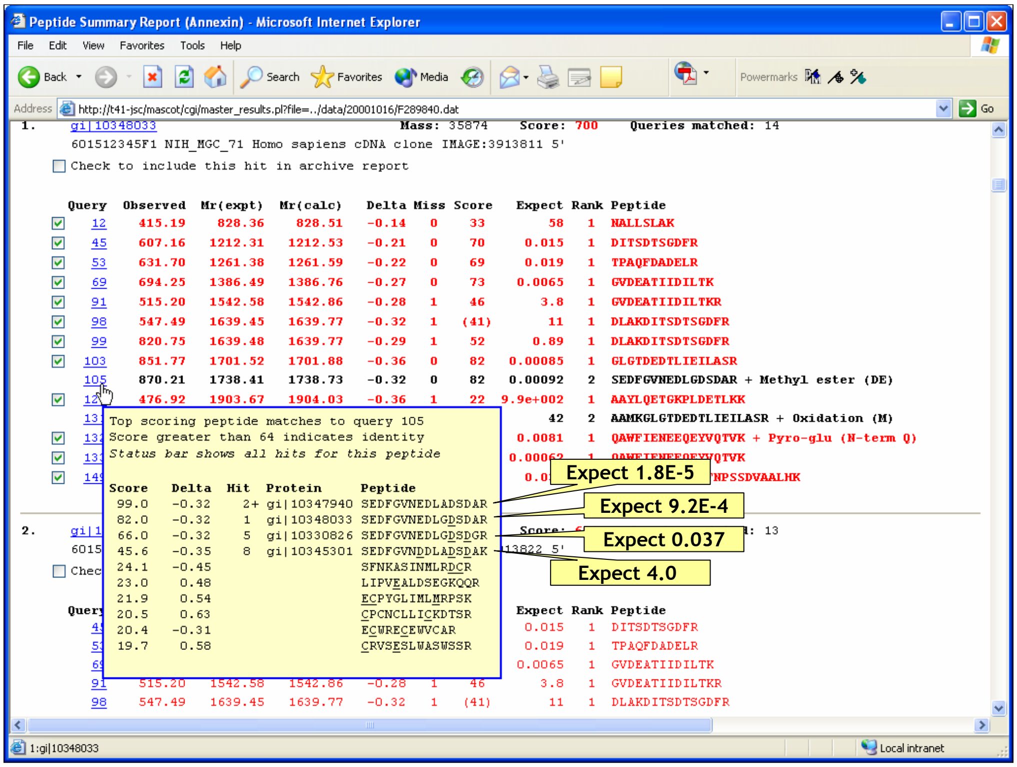
It is certainly possible if you are analysing a known protein or standard mixture of proteins. If you know what the sequences are, or think you know, then the matches to the known sequences are defined to be correct and those to any other sequence are defined to be wrong.

If the sample is an unknown, then it is difficult to define what is meant by a correct match.

Is the correct match the best match in the database? Certainly not ... this would be a false positive if the correct sequence was not in the database.

What about the best match out of all possible peptides. Yes, a reasonable definition, but not a very practical one. This is what we try to find in de novo sequencing. The reason for searching a database is that the data quality are not good enough for reliable de novo, so we reduce the size of the search space to the content of the chosen sequence database.

How about the peptide sequence that is uniquely and completely defined by the MS data? This is equally impractical. One rarely, if ever, sees a mass spectrum perfect enough to meet this criterion



Peptide Summary Report (Annexin) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media Powermarks
Address http://t41-jsc/mascot/cgi/master_results.pl?file=../data/20001016/F289840.dat
Go
Expect 1.8E-5
Expect 9.2E-4
Expect 0.037
Expect 4.0
1:gi|10348033
Local intranet

```
1.     gi|10348033                          Mass: 35874    Score: 700    Queries matched: 14
       601512345F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3913811 5'
     □ Check to include this hit in archive report

      Query  Observed  Mr(expt)  Mr(calc)  Delta Miss Score   Expect Rank Peptide
     ☑  12     415.19    828.36    828.51  -0.14   0    33        58   1  NALLSLAK
     ☑  45     607.16   1212.31   1212.53  -0.21   0    70     0.015   1  DITSDTSGDFR
     ☑  53     631.70   1261.38   1261.59  -0.22   0    69     0.019   1  TPAQFDADELR
     ☑  69     694.25   1386.49   1386.76  -0.27   0    73    0.0065   1  GVDEATIIDILTK
     ☑  91     515.20   1542.58   1542.86  -0.28   1    46       3.8   1  GVDEATIIDILTKR
     ☑  98     547.49   1639.45   1639.77  -0.32   1   (41)       11   1  DLAKDITSDTSGDFR
     ☑  99     820.75   1639.48   1639.77  -0.29   1    52      0.89   1  DLAKDITSDTSGDFR
     ☑ 103     851.77   1701.52   1701.88  -0.36   0    82   0.00085   1  GLGTDEDTLIEILASR
       105     870.21   1738.41   1738.73  -0.32   0    82   0.00092   2  SEDFGVNEDLGDSDAR + Methyl ester (DE)
     ☑ 12      476.92   1903.67   1904.03  -0.36   1    22  9.9e+002   1  AAYLQETGKPLDETLKK
       13                                                       42   2  AAMKGLGTDEDTLIEILASR + Oxidation (M)
     ☑ 13                                                   0.0081   1  QAWFIENEEQEYVQTVK + Pyro-glu (N-term Q)
     ☑ 13                                                  0.00062   1  QAWFIENEEQEYVQTVK
     ☑ 14                                                     0.01                NPSSDVAALHK

2.     gi|1                                       Score:                 : 13
       6015                                         13822 5'
     □ Chec

      Query                                                 Expect Rank Peptide
        4                                                    0.015   1  DITSDTSGDFR
        5                                                    0.019   1  TPAQFDADELR
        6                                                   0.0065   1  GVDEATIIDILTK
        91     515.20   1542.58   1542.86  -0.28   1    46       3.8   1  GVDEATIIDILTKR
        98     547.49   1639.45   1639.77  -0.32   1   (41)       11   1  DLAKDITSDTSGDFR
```

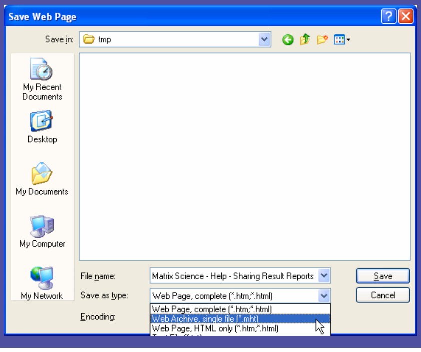
```
Top scoring peptide matches to query 105
Score greater than 64 indicates identity
Status bar shows all hits for this peptide

Score   Delta  Hit  Protein     Peptide
 99.0   -0.32   2+  gi|10347940 SEDFGVNEDLADSDAR
 82.0   -0.32   1   gi|10348033 SEDFGVNEDLGDSDAR
 66.0   -0.32   5   gi|10330826 SEDFGVNEDLGDSDGR
 45.6   -0.35   8   gi|10345301 SEDFGVNDDLADSDAK
 24.1   -0.45                   SFNKASINMLRDCR
 23.0    0.48                   LIPVEALDSEGKQQR
 21.9    0.54                   ECPYGLIMLMRPSK
 20.5    0.63                   CPCNCLLICKDTSR
 20.4   -0.31                   ECWRECEWVCAR
 19.7    0.58                   CRVSESLWASWSSR
```

This is a typical MS/MS search result, where we see a series of high scoring homologous peptides. The sequences of the top four matches are very similar, and their expectation values vary from random through to very unlikely to be random. The best match has an expectation value of 2E-5. However, we cannot be sure that this is an identity match to the analyte peptide. It is simply the best match we could find in the database. There is always the possibility that a better match exists, that is not in the database, so to call it the correct match can be misleading.

The important thing is that we can recognise and discard matches that are nothing more than random matches. I guess we aren’t even sure how to define correct, never mind calculate a probability for a particular match being correct

How can I send a result report to a colleague?

- Save a single report as web page complete or web archive



Save in: tmp

File name: Matrix Science - Help - Sharing Result Reports

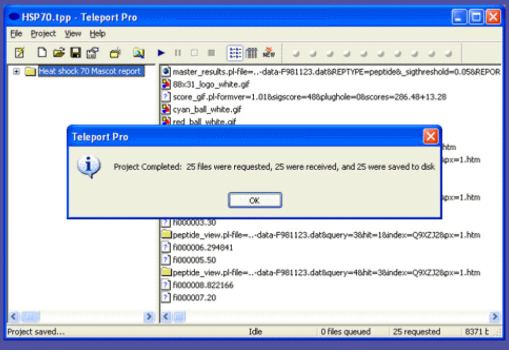
Save as type: Web Page, complete (*.htm;*.html)

Web Page, complete (*.htm;*.html)
Web Archive, single file (*.mht)
Web Page, HTML only (*.htm;*.html)

Saving as "Web page, HTML only" is no good because graphics like the score histogram will be missing.

How can I send a result report to a colleague?

- Print a single report to an Acrobat PDF file
- Capture a complete set of reports using an off-line browser utility



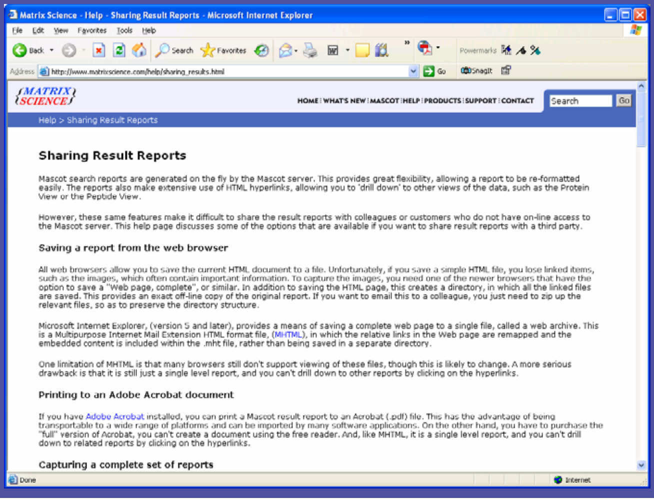
master_results.pl-file=..-data-F981123.dat&REPTYPE=peptide&_sigthreshold=0.05&REPOR
88x31_logo_white.gif
score_gif.pl-formver=1.01&sigscore=48&plughole=0&scores=286.48+13.28
cyan_ball_white.gif
red_ball_white.gif

Project Completed: 25 files were requested, 25 were received, and 25 were saved to disk

fi000003.30
peptide_view.pl-file=..-data-F981123.dat&query=3&hit=1&index=Q9XZJ2&px=1.htm
fi000006.294841
fi000005.50
peptide_view.pl-file=..-data-F981123.dat&query=4&hit=3&index=Q9XZJ2&px=1.htm
fi000008.822166
fi000007.20

Project saved... Idle 0 files queued 25 requested 8371 t

How can I send a result report to a colleague?



Matrix Science - Help - Sharing Result Reports - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Powermarks
Address http://www.matrixscience.com/help/sharing_results.html Go SnagIt
{MATRIX} {SCIENCE}
HOME | WHAT'S NEW | MASCOT | HELP | PRODUCTS | SUPPORT | CONTACT
Search Go
Help > Sharing Result Reports
Done
Internet

Sharing Result Reports

Mascot search reports are generated on the fly by the Mascot server. This provides great flexibility, allowing a report to be re-formatted easily. The reports also make extensive use of HTML hyperlinks, allowing you to 'drill down' to other views of the data, such as the Protein View or the Peptide View.

However, these same features make it difficult to share the result reports with colleagues or customers who do not have on-line access to the Mascot server. This help page discusses some of the options that are available if you want to share result reports with a third party.

Saving a report from the web browser

All web browsers allow you to save the current HTML document to a file. Unfortunately, if you save a simple HTML file, you lose linked items, such as the images, which often contain important information. To capture the images, you need one of the newer browsers that have the option to save a "Web page, complete", or similar. In addition to saving the HTML page, this creates a directory, in which all the linked files are saved. This provides an exact off-line copy of the original report. If you want to email this to a colleague, you just need to zip up the relevant files, so as to preserve the directory structure.

Microsoft Internet Explorer, (version 5 and later), provides a means of saving a complete web page to a single file, called a web archive. This is a Multipurpose Internet Mail Extension HTML format file, (MHTML), in which the relative links in the Web page are remapped and the embedded content is included within the .mht file, rather than being saved in a separate directory.

One limitation of MHTML is that many browsers still don't support viewing of these files, though this is likely to change. A more serious drawback is that it is still just a single level report, and you can't drill down to other reports by clicking on the hyperlinks.

Printing to an Adobe Acrobat document

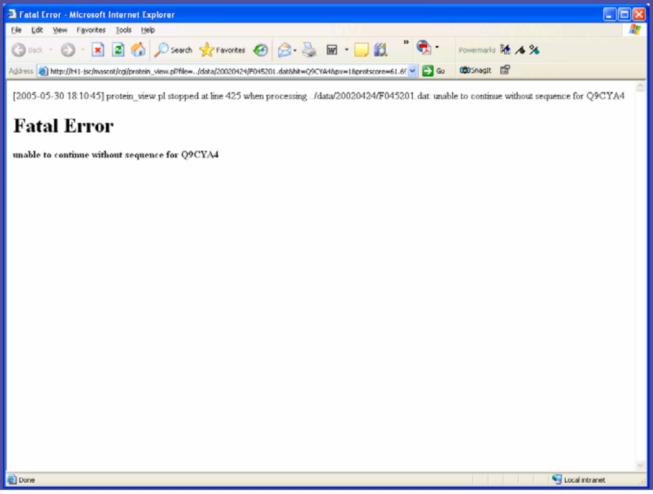
If you have Adobe Acrobat installed, you can print a Mascot result report to an Acrobat (.pdf) file. This has the advantage of being transportable to a wide range of platforms and can be imported by many software applications. On the other hand, you have to purchase the "full" version of Acrobat, you can't create a document using the free reader. And, like MHTML, it is a single level report, and you can't drill down to related reports by clicking on the hyperlinks.

Capturing a complete set of reports

ASMS 2005

A detailed answer to this question can be found at http://www.matrixscience.com/help/sharing_results.html

Why can’t I get a Protein View report for some hits from an old search?


Fatal Error - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Powermarks
Address http://t41-jsc/mascot/cgi/protein_view.pl?file=../data/20020424/F045201.dat&hit=Q9CYA4&px=1&protscore=61.6 Go SnagIt
Done
Local intranet

[2005-05-30 18:10:45] protein_view.pl stopped at line 425 when processing ../data/20020424/F045201.dat: unable to continue without sequence for Q9CYA4

Fatal Error

unable to continue without sequence for Q9CYA4

ASMS 2005

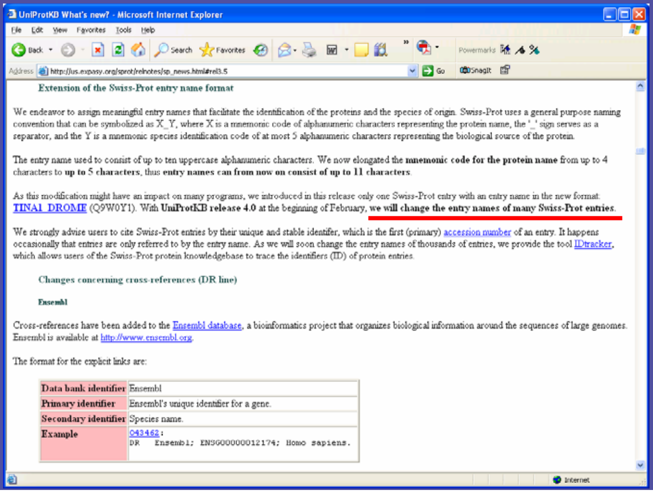
MATRIX SCIENCE

Mascot cannot save complete protein sequences to the result file. To do so would make each result file enormous. When you request a Protein View report, the script gets the protein sequence from the current Fasta file.

Unfortunately, database accessions are not nearly as stable as one might expect. A percentage disappear in each update because the entry is revised and gets a new accession. Even Swiss-Prot suffers from this problem

One fix would be to retain old database files on your Mascot server by creating a new database definition for each update, rather than just replacing the file. Alternatively, if the result is important, it may be easiest to repeat the search against the current Fasta file. I prefer the second route. You just have to choose "Search selected" from the Master results report

Why can't I get a Protein View report for some hits from an old search?



UniProtKB What's new? - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Powermarks
Address http://us.expasy.org/sprot/relnotes/sp_news.html#rel3.5 Go SnagIt
Internet

Extension of the Swiss-Prot entry name format

We endeavor to assign meaningful entry names that facilitate the identification of the proteins and the species of origin. Swiss-Prot uses a general purpose naming convention that can be symbolized as X_Y, where X is a mnemonic code of alphanumeric characters representing the protein name, the '_' sign serves as a separator, and the Y is a mnemonic species identification code of at most 5 alphanumeric characters representing the biological source of the protein.

The entry name used to consist of up to ten uppercase alphanumeric characters. We now elongated the **mnemonic code for the protein name** from up to 4 characters to **up to 5 characters**, thus **entry names can from now on consist of up to 11 characters**.

As this modification might have an impact on many programs, we introduced in this release only one Swiss-Prot entry with an entry name in the new format: **TINA1_DROME** (Q9W0Y1). With **UniProtKB release 4.0** at the beginning of February, **we will change the entry names of many Swiss-Prot entries**.

We strongly advise users to cite Swiss-Prot entries by their unique and stable identifer, which is the first (primary) accession number of an entry. It happens occasionally that entries are only referred to by the entry name. As we will soon change the entry names of thousands of entries, we provide the tool IDtracker, which allows users of the Swiss-Prot protein knowledgebase to trace the identifiers (ID) of protein entries.

Changes concerning cross-references (DR line)

Ensembl

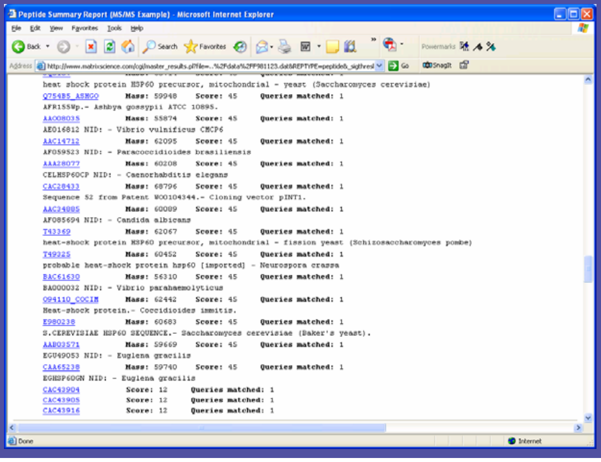
Cross-references have been added to the Ensembl database, a bioinformatics project that organizes biological information around the sequences of large genomes. Ensembl is available at http://www.ensembl.org.

The format for the explicit links are:

Data bank identifier	Ensembl
Primary identifier	Ensembl's unique identifier for a gene.
Secondary identifier	Species name.
Example	O43462: `DR   Ensembl; ENSG00000012174; Homo sapiens.`

ASMS 2005

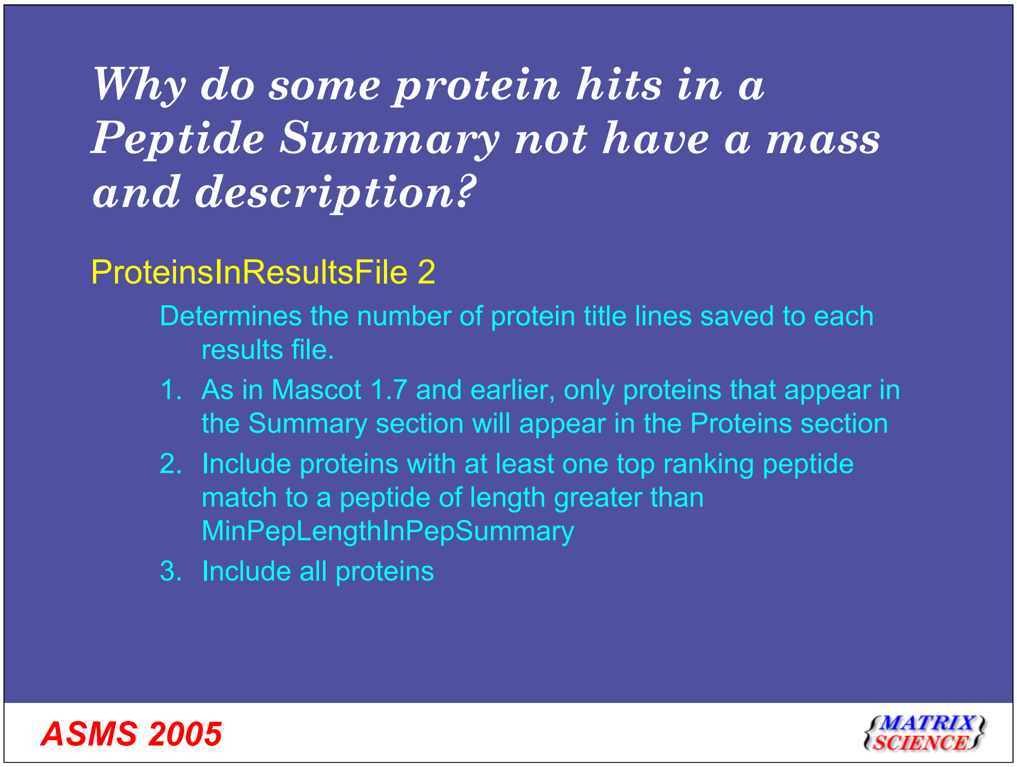
Why do some protein hits in a Peptide Summary not have a mass and description?



Peptide Summary Report (MS/MS Example) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Powermarks
Address http://www.matrixscience.com/cgi/master_results.pl?file=../%2Fdata%2FF981123.dat&REPTYPE=peptide&_sigthresl
Go
SnagIt
Done
Internet

```
heat shock protein HSP60 precursor, mitochondrial - yeast (Saccharomyces cerevisiae)
Q754B5_ASHGO      Mass: 59948   Score: 45    Queries matched: 1
AFR155Wp.- Ashbya gossypii ATCC 10895.
AAC08035          Mass: 55874   Score: 45    Queries matched: 1
AE016812 NID: - Vibrio vulnificus CMCP6
AAC14712          Mass: 62095   Score: 45    Queries matched: 1
AF059523 NID: - Paracoccidioides brasiliensis
AAA28077          Mass: 60208   Score: 45    Queries matched: 1
CELHSP60CP NID: - Caenorhabditis elegans
CAC28433          Mass: 68796   Score: 45    Queries matched: 1
Sequence 52 from Patent WO0104344.- Cloning vector pINT1.
AAC34005          Mass: 60009   Score: 45    Queries matched: 1
AF085694 NID: - Candida albicans
T43369            Mass: 62067   Score: 45    Queries matched: 1
heat-shock protein HSP60 precursor, mitochondrial - fission yeast (Schizosaccharomyces pombe)
T49325            Mass: 60452   Score: 45    Queries matched: 1
probable heat-shock protein hsp60 [imported] - Neurospora crassa
BAC61630          Mass: 56310   Score: 45    Queries matched: 1
BA000032 NID: - Vibrio parahaemolyticus
O94110_COCIM      Mass: 62442   Score: 45    Queries matched: 1
Heat-shock protein.- Coccidioides immitis.
E980238           Mass: 60683   Score: 45    Queries matched: 1
S.CEREVISIAE HSP60 SEQUENCE.- Saccharomyces cerevisiae (Baker's yeast).
AAB03571          Mass: 59669   Score: 45    Queries matched: 1
EGU49053 NID: - Euglena gracilis
CAA65238          Mass: 59740   Score: 45    Queries matched: 1
EGHSP60GN NID: - Euglena gracilis
CAC43904          Score: 12    Queries matched: 1
CAC43905          Score: 12    Queries matched: 1
CAC43916          Score: 12    Queries matched: 1
```

The Mascot result file includes title strings and mass values for all the proteins it "expects" to display in a standard report. However, in a large search, it may miss a few. If this information is missing for a primary hit, (the first one listed for a given hit number), the report uses a utility (ms-getseq.exe) to retrieve this information from the Fasta file. For non-primary hits, it does not do this because it would greatly increase the time taken to load the report.



Why do some protein hits in a Peptide Summary not have a mass and description?

ProteinsInResultsFile 2

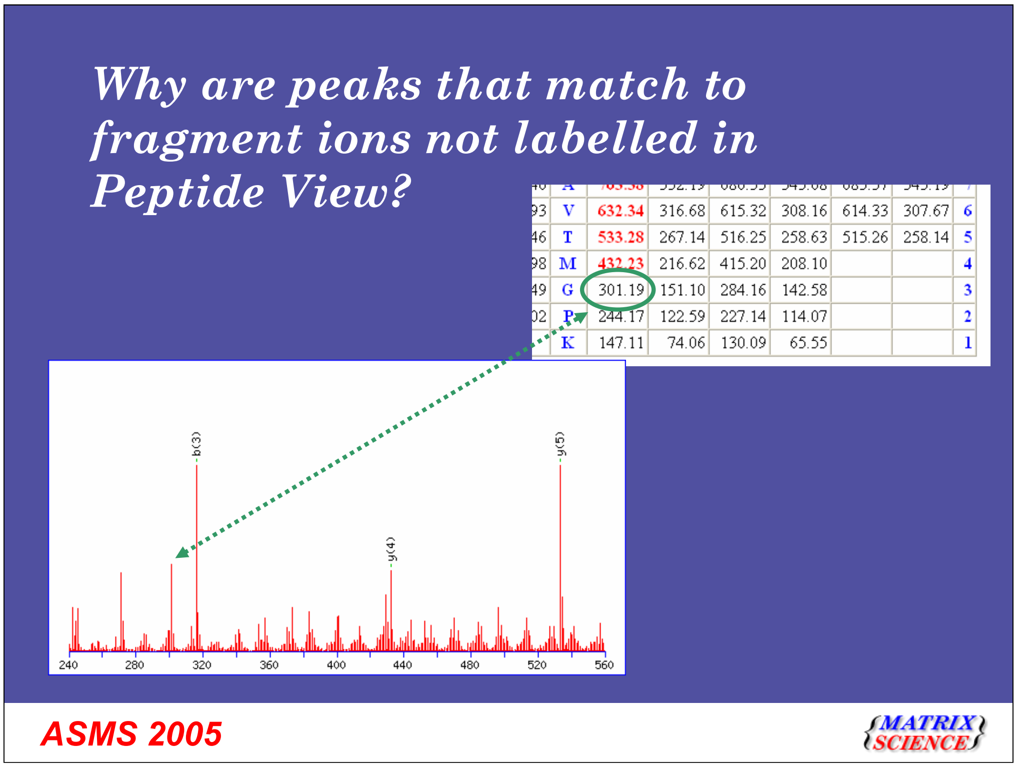
Determines the number of protein title lines saved to each results file.

1. As in Mascot 1.7 and earlier, only proteins that appear in the Summary section will appear in the Proteins section
2. Include proteins with at least one top ranking peptide match to a peptide of length greater than MinPepLengthInPepSummary
3. Include all proteins

ASMS 2005

This behaviour is controlled by a setting in the Options section of mascot.dat. You can force the title and mass for all proteins to be stored in the result file, so that they are never missing. The down side is that this will cause the size of the result files to increase.

The setting is ProteinsInResultsFile in the options section of mascot.dat (see chapter 6 of the Mascot installation and setup manual).



Why are peaks that match to fragment ions not labelled in Peptide View?

[illegible]	A	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7
93	V	**632.34**	316.68	615.32	308.16	614.33	307.67	6
46	T	**533.28**	267.14	516.25	258.63	515.26	258.14	5
98	M	**432.23**	216.62	415.20	208.10			4
49	G	301.19	151.10	284.16	142.58			3
02	P	244.17	122.59	227.14	114.07			2
	K	147.11	74.06	130.09	65.55			1

ASMS 2005

Mascot begins by selecting a small number of experimental peaks on the basis of normalised intensity. It calculates a probability based score according to the number of matches. It then increases the number of selected peaks and recalculates the score. It continues to iterate until it is clear that the score can only get worse. It then reports the best score it found, which should correspond to an optimum selection, taking mostly real peaks and leaving behind mostly noise

Mascot is not trying to find all possible matches in the spectrum. As in this example, many spectra have "peak at every mass" noise, and can match any fragment ion from any sequence if there is no intensity discrimination. So, you may look at a peptide view report and see obvious matches that are unlabelled. However, if the peak selection was to be extended to include these additional matches, it would also have to include a number of additional noise peaks, and the score would decrease.

There are no user parameters to influence this behaviour.