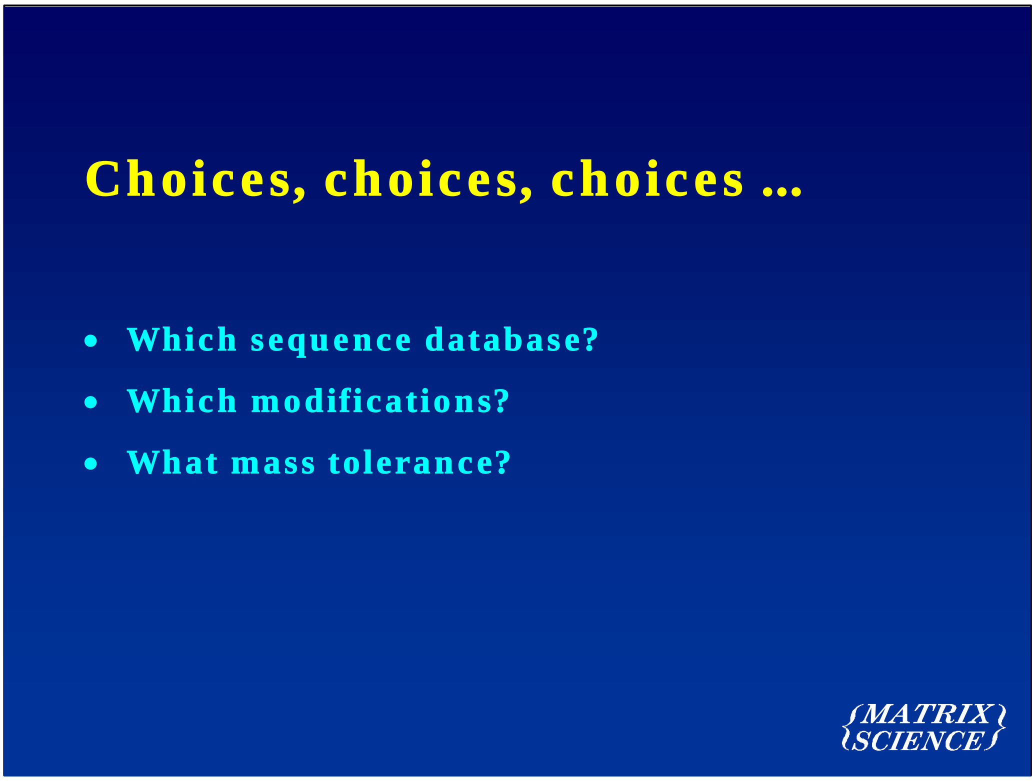
Optimization

{MATRIX}
{SCIENCE}

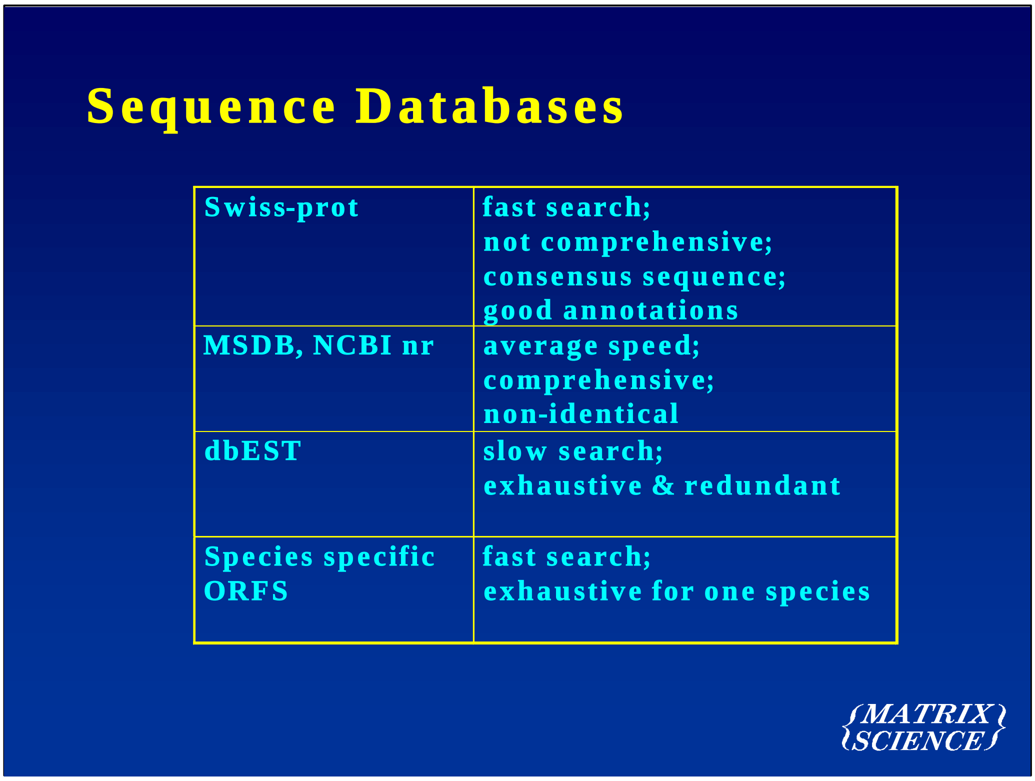


Choices, choices, choices …

- Which sequence database?
- Which modifications?
- What mass tolerance?

{MATRIX}
{SCIENCE}

Where to begin?



Sequence Databases

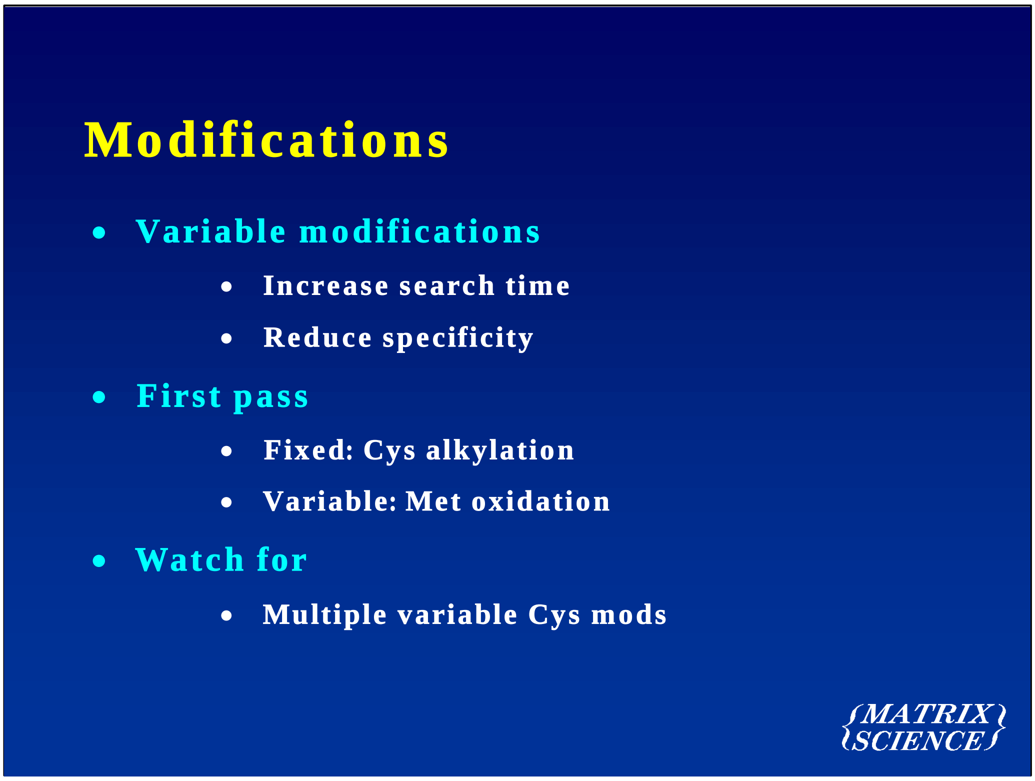
Swiss-prot	fast search;
	not comprehensive;
	consensus sequence;
	good annotations

MSDB, NCBI nr	average speed;
	comprehensive;
	non-identical

dbEST	slow search; exhaustive & redundant
Species specific ORFS	fast search; exhaustive for one species

MATRIX SCIENCE

Swiss-prot is the highest quality database, but many entries are consensus sequences, with variants described in the annotations. Mascot searches only the FASTA sequence, so these variants are missed. Better to use a database where variant sequences are included as separate entries.

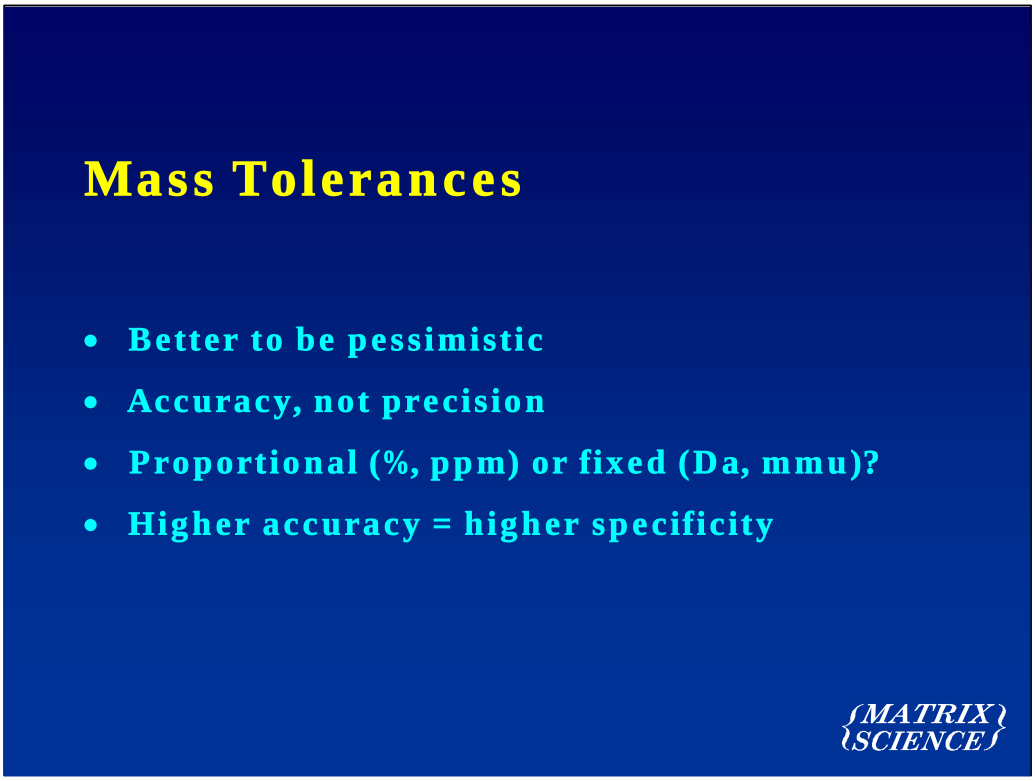


Modifications

- Variable modifications
  - Increase search time
  - Reduce specificity
- First pass
  - Fixed: Cys alkylation
  - Variable: Met oxidation
- Watch for
  - Multiple variable Cys mods

MATRIX SCIENCE

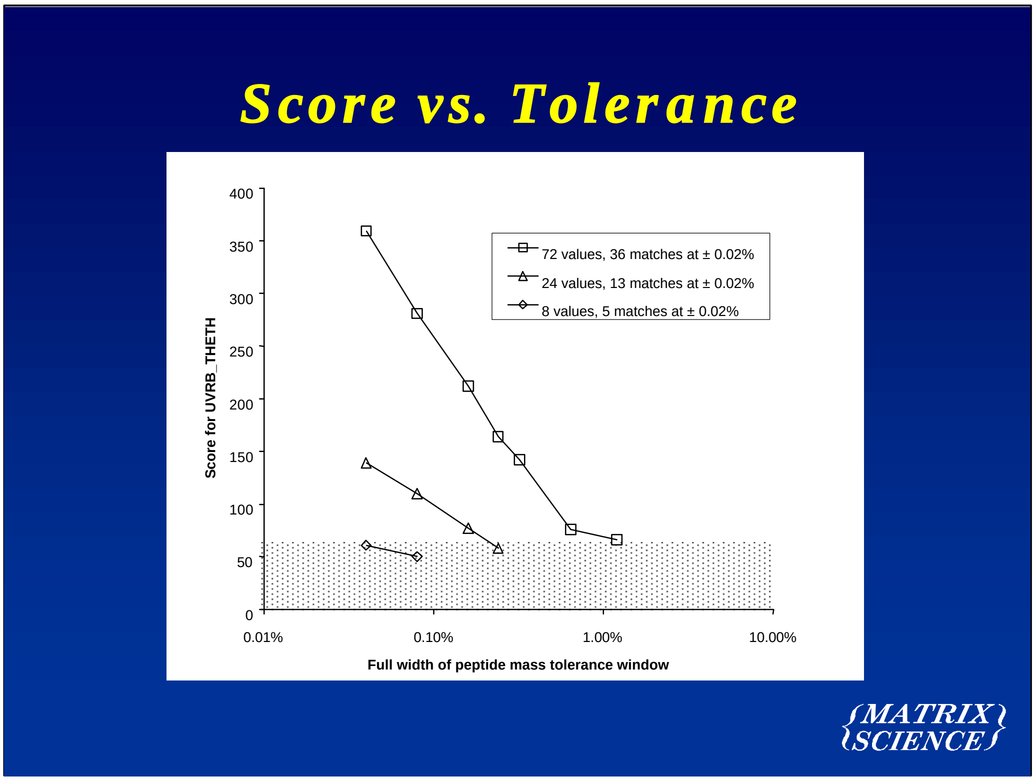
Modfications should be used sparingly in a first pass search.



Mass Tolerances

- Better to be pessimistic
- Accuracy, not precision
- Proportional (%, ppm) or fixed (Da, mmu)?
- Higher accuracy = higher specificity

{MATRIX}
{SCIENCE}



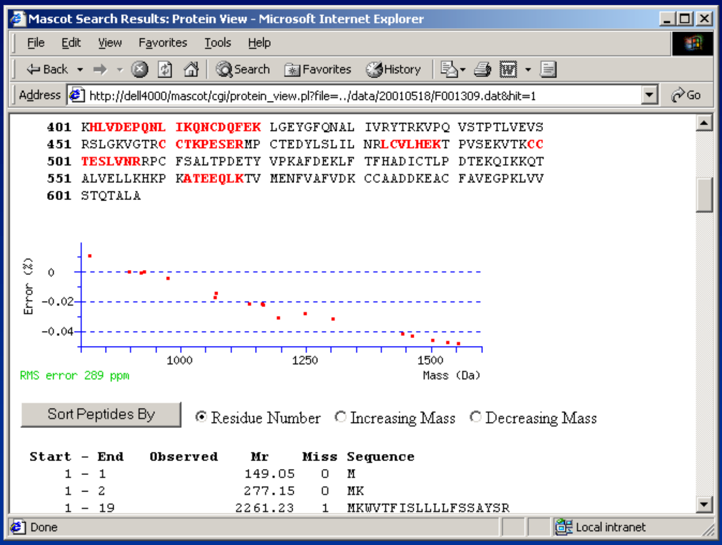
Score vs. Tolerance

{MATRIX}
{SCIENCE}

For peptide mass fingerprinting, high mass accuracy is most important when there are only a few mass values. As the data set becomes larger, high accuracy becomes less critical.

For a data set with 36 matches from72 mass values, a significant match can be obtained even when the mass tolerance approaches 1%. With a smaller data set, 13 matches from 24, a significant match requires a mass tolerance of better than 0.2%. If the data set is only 5 matches from 8, the match is never significant.

Error distribution graph


Mascot Search Results: Protein View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell4000/mascot/cgi/protein_view.pl?file=../data/20010518/F001309.dat&hit=1
Go
Error (%)
0
-0.02
-0.04
1000
1250
1500
Mass (Da)
Done
Local intranet

```
401 KHLVDEPQNL IKQNCDQFEK LGEYGFQNAL IVRYTRKVPQ VSTPTLVEVS
451 RSLGKVGTRC CTKPESERMP CTEDYLSLIL NRLCVLHEKT PVSEKVTKCC
501 TESLVNRRPC FSALTPDETY VPKAFDEKLF TFHADICTLP DTEKQIKKQT
551 ALVELLKHKP KATEEQLKTV MENFVAFVDK CCAADDKEAC FAVEGPKLVV
601 STQTALA
```

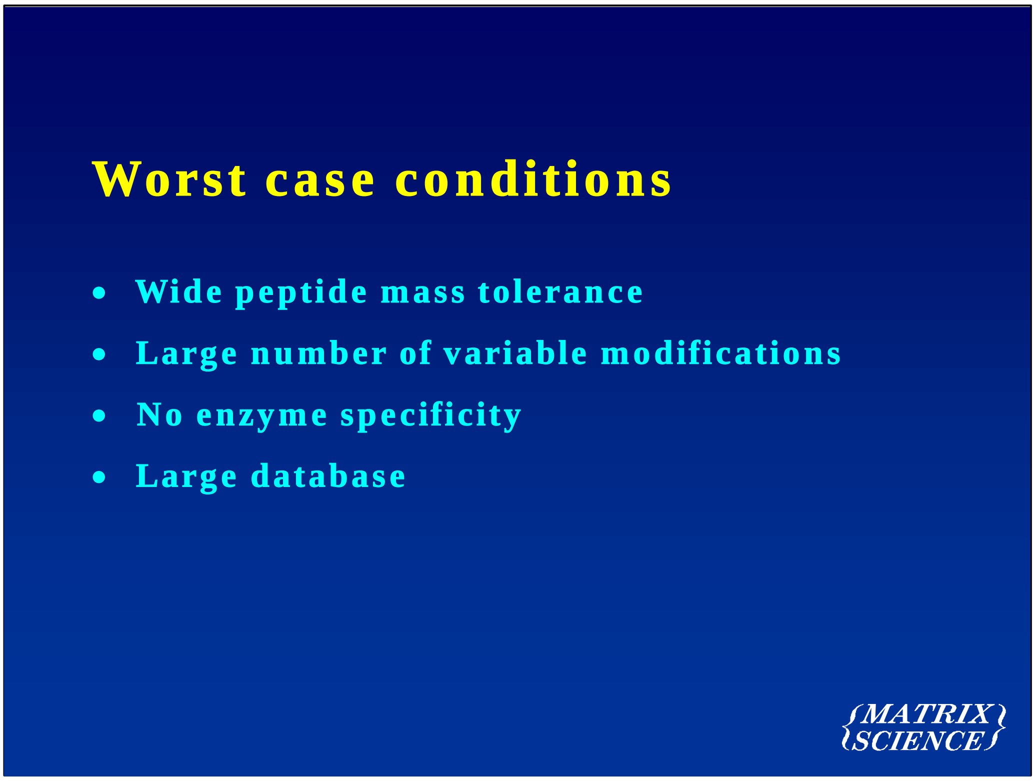
RMS error 289 ppm

Sort Peptides By ● Residue Number ○ Increasing Mass ○ Decreasing Mass

```
Start - End   Observed     Mr     Miss Sequence
    1 - 1               149.05    0   M
    1 - 2               277.15    0   MK
    1 - 19             2261.23    1   MKWVTFISLLLLFSSAYSR
```

The best way to decide on the mass tolerance setting is to look at the error graphs. For peptide mass values, the error graph is on the protein view report. For fragment ion mass values, the error graph is on the peptide view report.

The graphs will also give an indication of whether a constant (Da, mmu) or a fractional (%, ppm) error window is most appropriate.



Worst case conditions

- Wide peptide mass tolerance
- Large number of variable modifications
- No enzyme specificity
- Large database

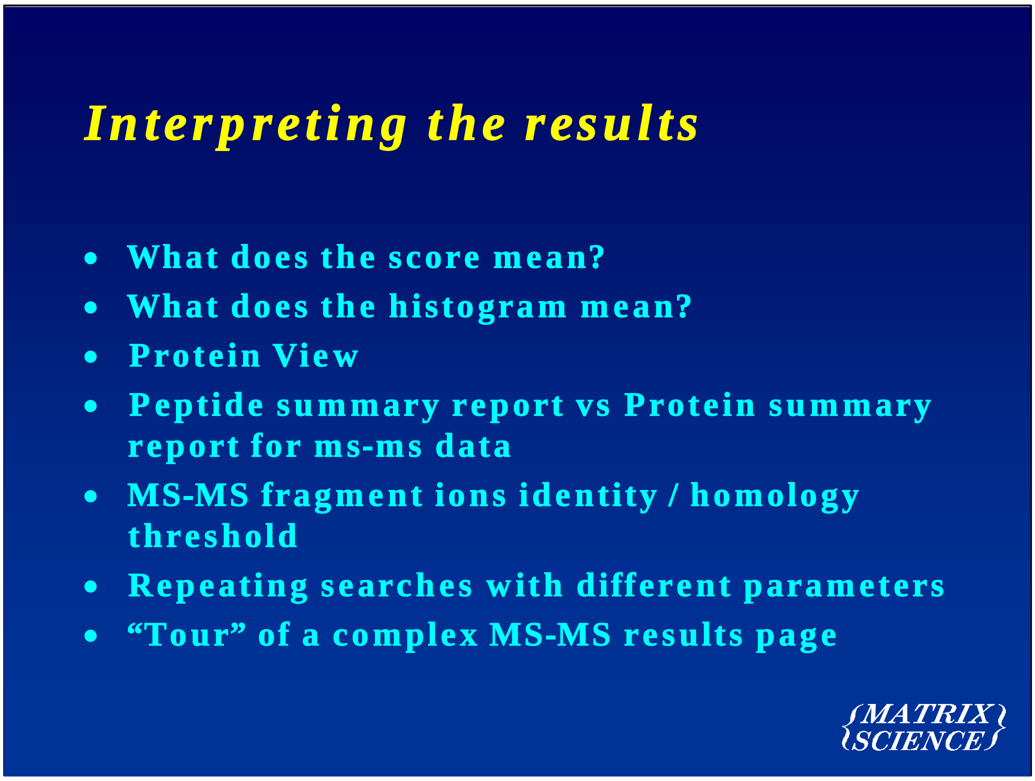
MATRIX SCIENCE

Search time and search specificity are inversely related.

Search time increases proportionately to peptide mass tolerance and database size.

Search time increases geometrically with the number of variable modifications.

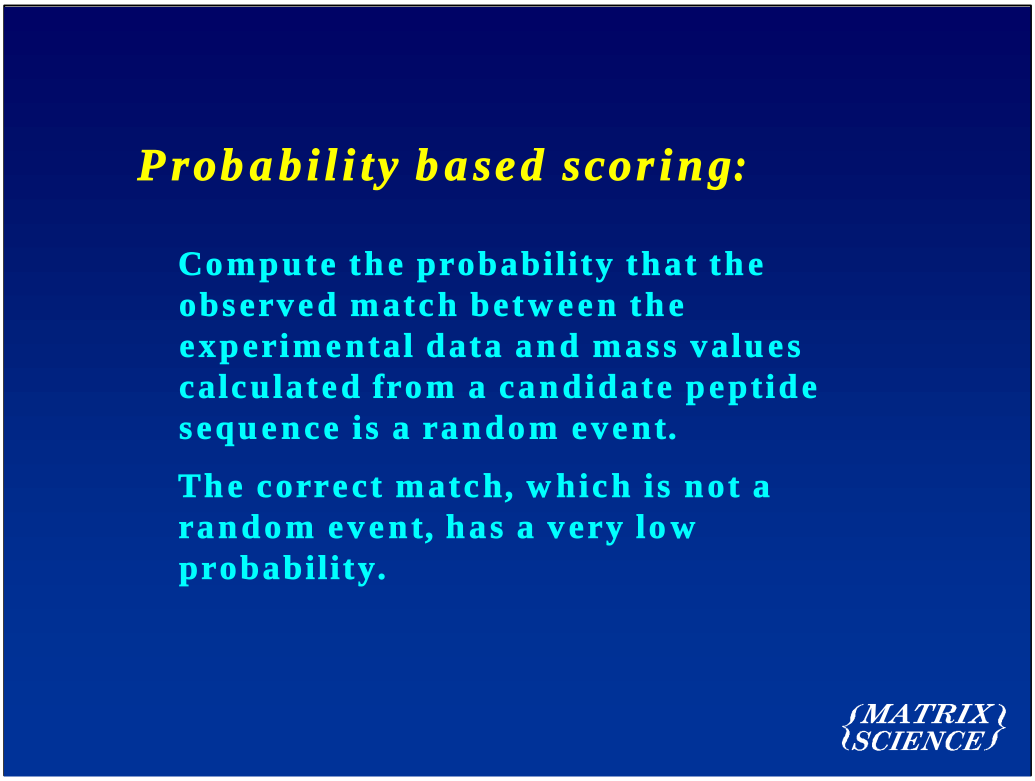
Going from tryptic specificity to no-enzyme will typically increase the search time by a factor between 100 and 1000



Interpreting the results

- What does the score mean?
- What does the histogram mean?
- Protein View
- Peptide summary report vs Protein summary report for ms-ms data
- MS-MS fragment ions identity / homology threshold
- Repeating searches with different parameters
- “Tour” of a complex MS-MS results page

{MATRIX}
{SCIENCE}

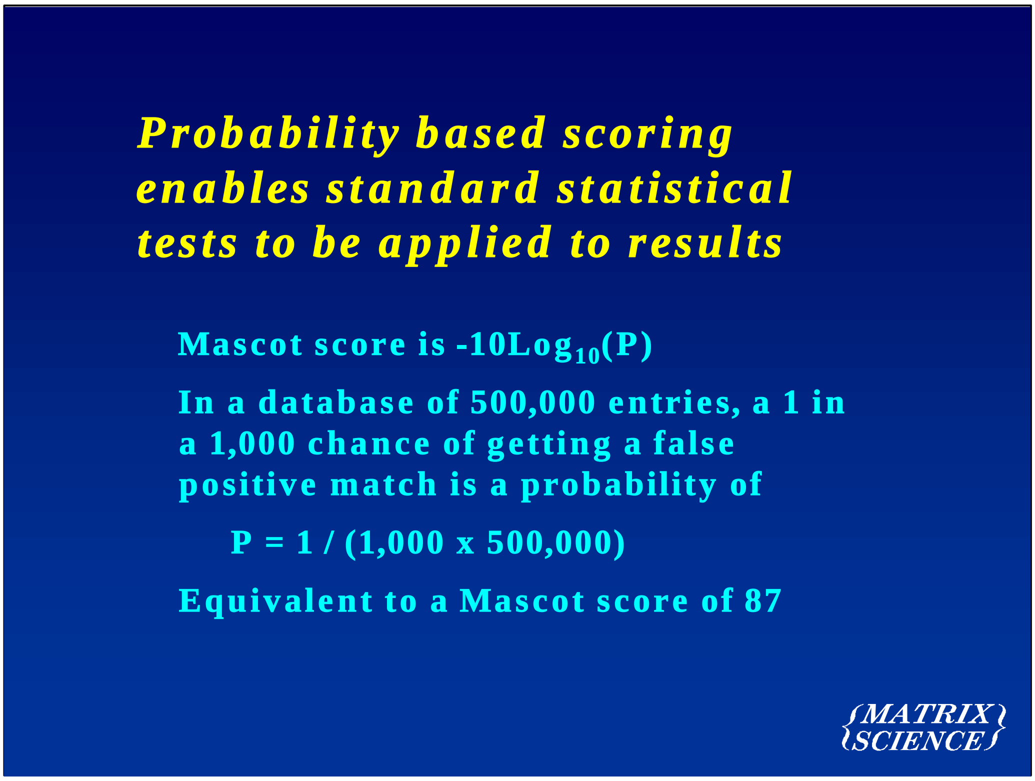


Probability based scoring:

Compute the probability that the observed match between the experimental data and mass values calculated from a candidate peptide sequence is a random event.

The correct match, which is not a random event, has a very low probability.

{MATRIX
SCIENCE}



Probability based scoring enables standard statistical tests to be applied to results

Mascot score is $-10\mathrm{Log}_{10}(P)$

In a database of 500,000 entries, a 1 in a 1,000 chance of getting a false positive match is a probability of

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

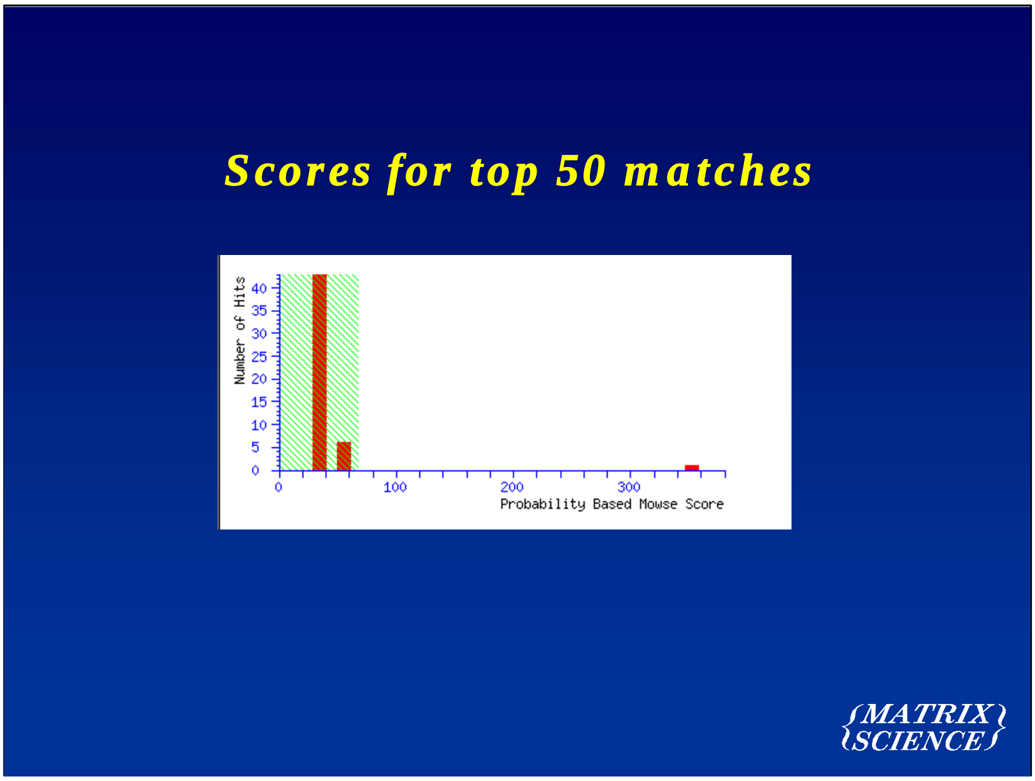
Equivalent to a Mascot score of 87

The most important advantage of probability based scoring is that we can use standard statistical tests to determine significance. That is, we have an objective means of determining whether a match is strong or weak ... or a false positive.

Assigning a significance threshold or confidence level to a match is extremely simple. Assume we are running a fully automated system and prefer to repeat an experiment rather than get a false positive. We might choose a significance threshold of 1 in 1,000. That is, we are only interested in results which have less than a 1 in 1,000 chance of being random events.

If the database being searched has 500,000 protein entries, a 1 in 1,000 chance of finding a match is simply 1 over 1,000 times 500,000. Which converts into a Mascot score of 87.

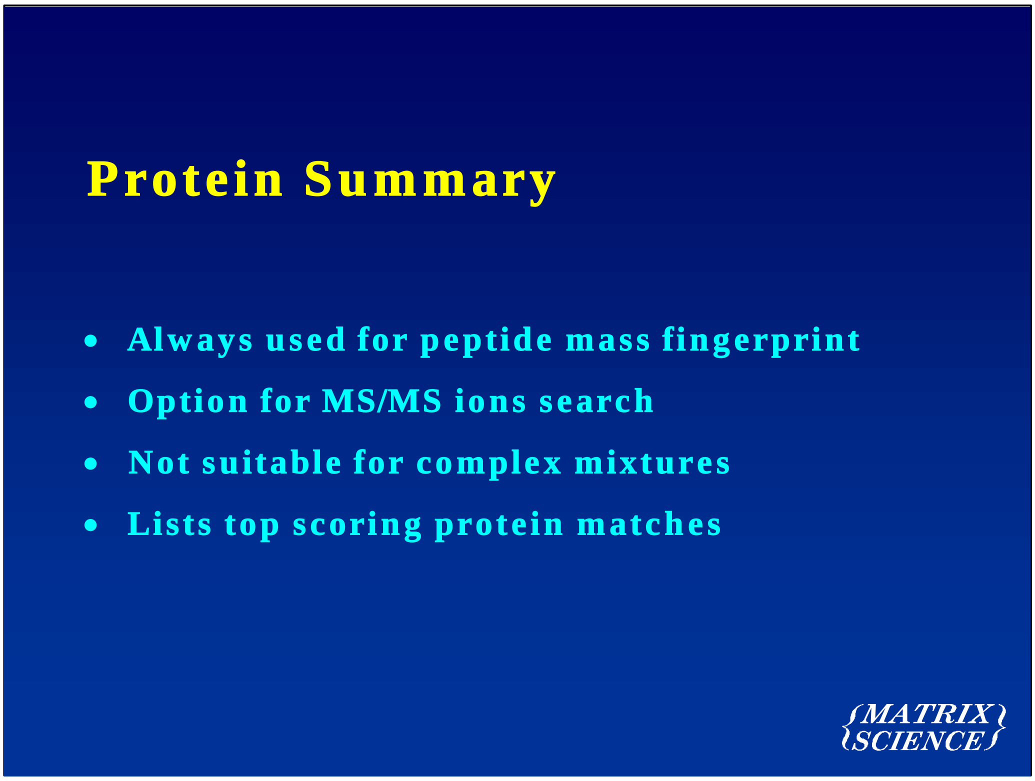
So, we can have a simple rule in software which looks for matches with scores greater than 87.



Scores for top 50 matches

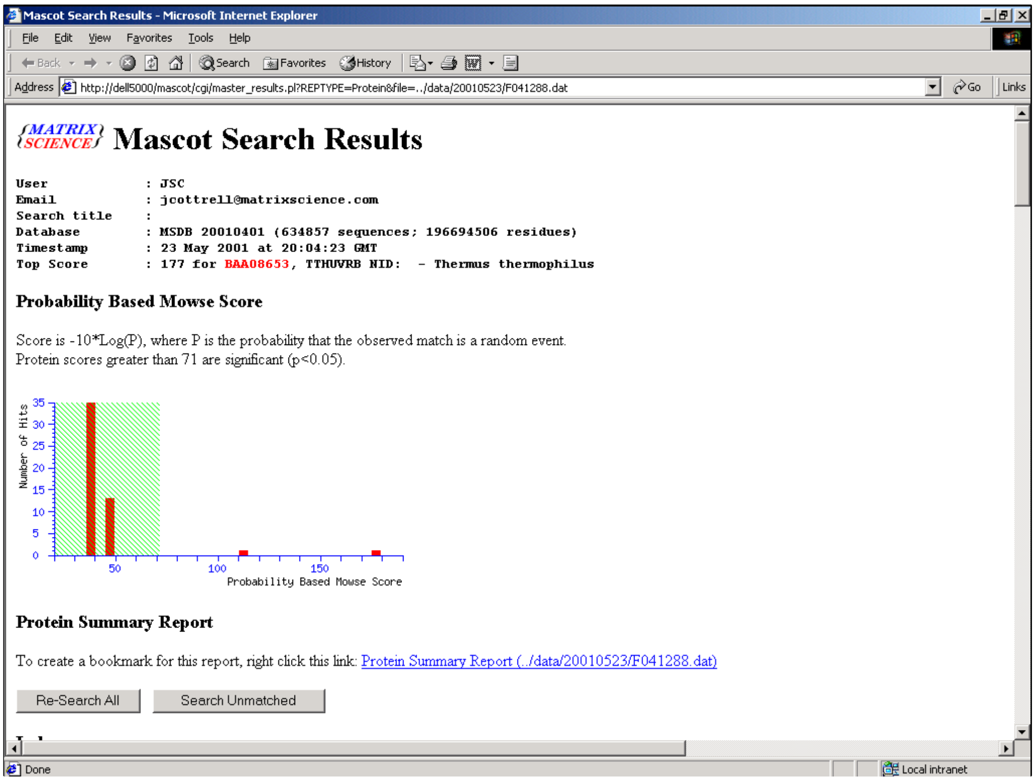
{MATRIX}
{SCIENCE}

At the top of each report, there is a histogram of the score distribution for the top 50 matches. Here, out of the top 50 protein hits, 49 have scores which are below the 5% significance threshold of 67. The area below the significance threshold is shaded green. One hit has a much higher score, 352. Very much higher when you appreciate that this is a logarithmic scale.



Protein Summary

- Always used for peptide mass fingerprint
- Option for MS/MS ions search
- Not suitable for complex mixtures
- Lists top scoring protein matches


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/master_results.pl?REPTYPE=Protein&file=../data/20010523/F041288.dat
Go Links
{MATRIX SCIENCE}
Number of Hits
35
30
25
20
15
10
5
0
50
100
150
Probability Based Mowse Score
Re-Search All
Search Unmatched
Done
Local intranet

Mascot Search Results

```
User            : JSC
Email           : jcottrell@matrixscience.com
Search title    :
Database        : MSDB 20010401 (634857 sequences; 196694506 residues)
Timestamp       : 23 May 2001 at 20:04:23 GMT
Top Score       : 177 for BAA08653, TTHUVRB NID:  - Thermus thermophilus
```

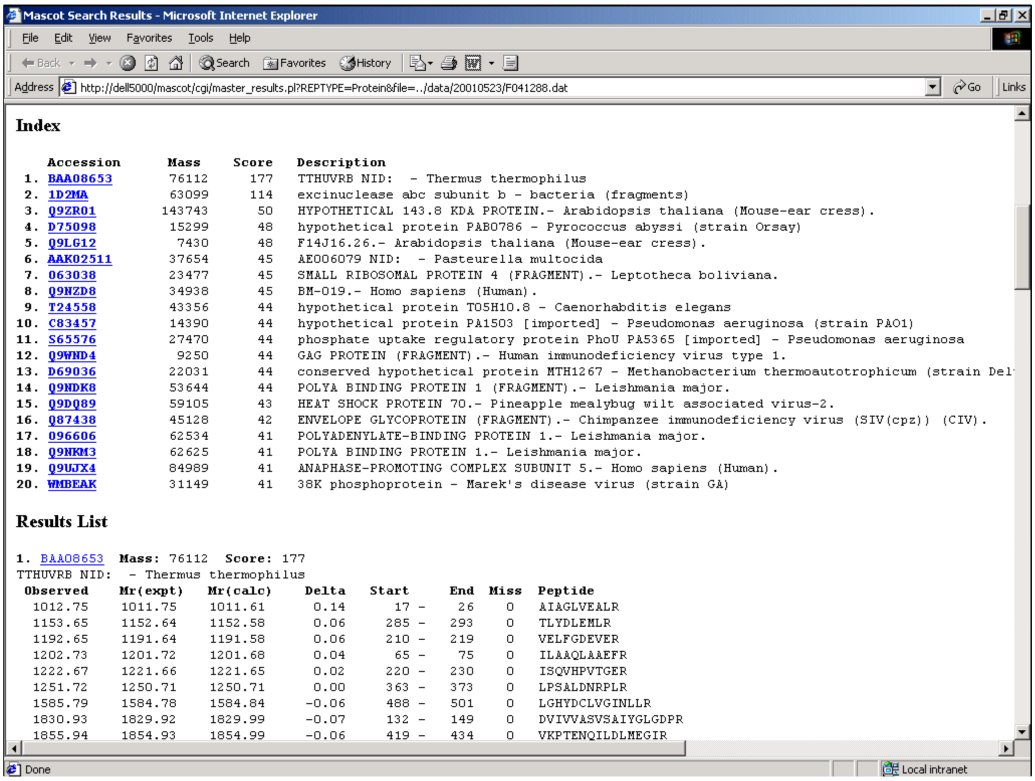
Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 71 are significant (p<0.05).

Protein Summary Report

To create a bookmark for this report, right click this link: Protein Summary Report (../data/20010523/F041288.dat)

The top of a protein summary report


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
History
Address http://dell5000/mascot/cgi/master_results.pl?REPTYPE=Protein&file=../data/20010523/F041288.dat
Go
Links
Done
Local intranet

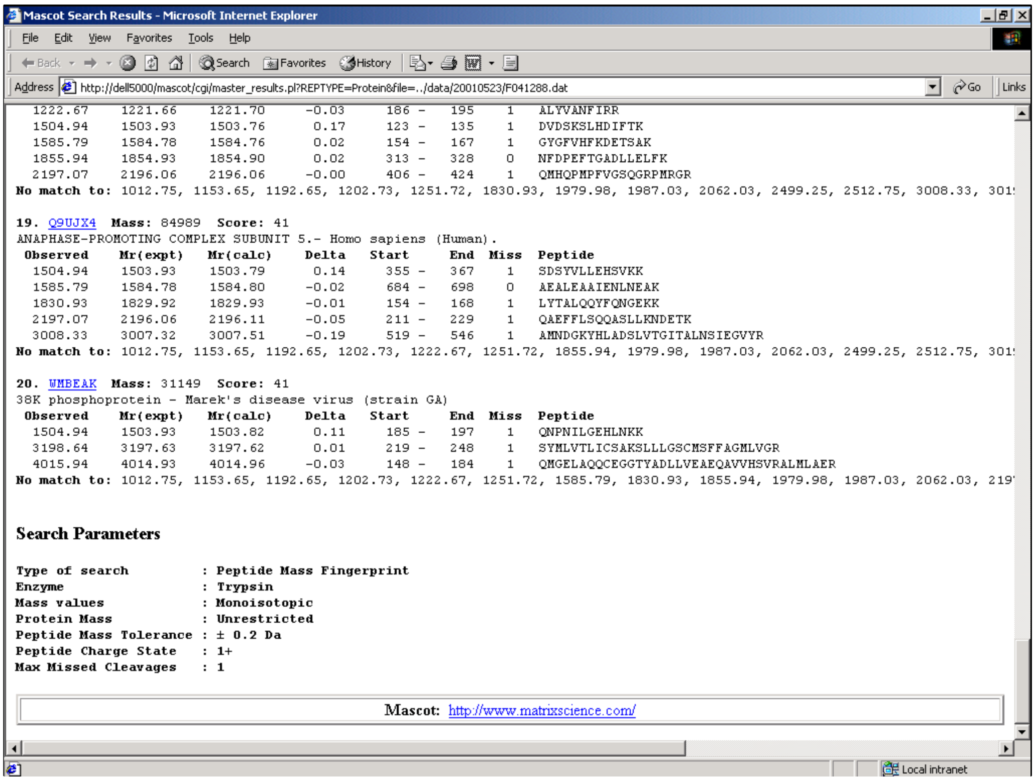
Index

```
     Accession       Mass    Score   Description
 1.  BAA08653       76112     177   TTHUVRB NID:  - Thermus thermophilus
 2.  1D2MA          63099     114   excinuclease abc subunit b - bacteria (fragments)
 3.  Q9ZR01        143743      50   HYPOTHETICAL 143.8 KDA PROTEIN.- Arabidopsis thaliana (Mouse-ear cress).
 4.  D75098         15299      48   hypothetical protein PAB0786 - Pyrococcus abyssi (strain Orsay)
 5.  Q9LG12          7430      48   F14J16.26.- Arabidopsis thaliana (Mouse-ear cress).
 6.  AAK02511       37654      45   AE006079 NID:  - Pasteurella multocida
 7.  063038         23477      45   SMALL RIBOSOMAL PROTEIN 4 (FRAGMENT).- Leptotheca boliviana.
 8.  Q9NZD8         34938      45   BM-019.- Homo sapiens (Human).
 9.  T24558         43356      44   hypothetical protein T05H10.8 - Caenorhabditis elegans
10.  C83457         14390      44   hypothetical protein PA1503 [imported] - Pseudomonas aeruginosa (strain PAO1)
11.  S65576         27470      44   phosphate uptake regulatory protein PhoU PA5365 [imported] - Pseudomonas aeruginosa
12.  Q9WND4          9250      44   GAG PROTEIN (FRAGMENT).- Human immunodeficiency virus type 1.
13.  D69036         22031      44   conserved hypothetical protein MTH1267 - Methanobacterium thermoautotrophicum (strain Del
14.  Q9NDK8         53644      44   POLYA BINDING PROTEIN 1 (FRAGMENT).- Leishmania major.
15.  Q9DQ89         59105      43   HEAT SHOCK PROTEIN 70.- Pineapple mealybug wilt associated virus-2.
16.  Q87438         45128      42   ENVELOPE GLYCOPROTEIN (FRAGMENT).- Chimpanzee immunodeficiency virus (SIV(cpz)) (CIV).
17.  096606         62534      41   POLYADENYLATE-BINDING PROTEIN 1.- Leishmania major.
18.  Q9NKM3         62625      41   POLYA BINDING PROTEIN 1.- Leishmania major.
19.  Q9UJX4         84989      41   ANAPHASE-PROMOTING COMPLEX SUBUNIT 5.- Homo sapiens (Human).
20.  WMBEAK         31149      41   38K phosphoprotein - Marek's disease virus (strain GA)
```

Results List

```
1. BAA08653  Mass: 76112  Score: 177
TTHUVRB NID:  - Thermus thermophilus
 Observed    Mr(expt)   Mr(calc)    Delta    Start     End  Miss  Peptide
  1012.75    1011.75    1011.61     0.14       17 -     26    0   AIAGLVEALR
  1153.65    1152.64    1152.58     0.06      285 -    293    0   TLYDLEMLR
  1192.65    1191.64    1191.58     0.06      210 -    219    0   VELFGDEVER
  1202.73    1201.72    1201.68     0.04       65 -     75    0   ILAAQLAAEFR
  1222.67    1221.66    1221.65     0.02      220 -    230    0   ISQVHPVTGER
  1251.72    1250.71    1250.71     0.00      363 -    373    0   LPSALDNRPLR
  1585.79    1584.78    1584.84    -0.06      488 -    501    0   LGHYDCLVGINLLR
  1830.93    1829.92    1829.99    -0.07      132 -    149    0   DVIVVASVSAIYGLGDPR
  1855.94    1854.93    1854.99    -0.06      419 -    434    0   VKPTENQILDLMEGIR
```

The hit list for a protein summary report



http://dell5000/mascot/cgi/master_results.pl?REPTYPE=Protein&file=../data/20010523/F041288.dat

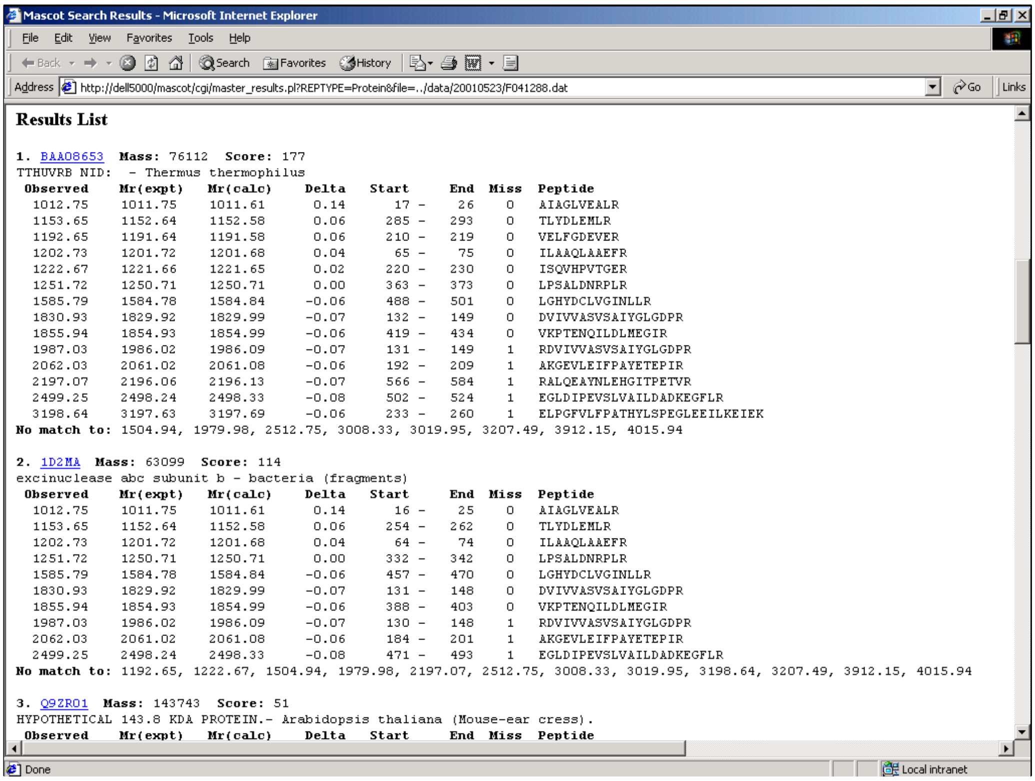
```
  1222.67    1221.66    1221.70     -0.03     186 -    195     1   ALYVANFIRR
  1504.94    1503.93    1503.76      0.17     123 -    135     1   DVDSKSLHDIFTK
  1585.79    1584.78    1584.76      0.02     154 -    167     1   GYGFVHFKDETSAK
  1855.94    1854.93    1854.90      0.02     313 -    328     0   NFDPEFTGADLLELFK
  2197.07    2196.06    2196.06     -0.00     406 -    424     1   QMHQPMPFVGSQGRPMRGR
No match to: 1012.75, 1153.65, 1192.65, 1202.73, 1251.72, 1830.93, 1979.98, 1987.03, 2062.03, 2499.25, 2512.75, 3008.33, 301

19. Q9UJX4  Mass: 84989  Score: 41
ANAPHASE-PROMOTING COMPLEX SUBUNIT 5.- Homo sapiens (Human).
 Observed    Mr(expt)   Mr(calc)    Delta   Start     End  Miss  Peptide
  1504.94    1503.93    1503.79      0.14     355 -    367     1   SDSYVLLEHSVKK
  1585.79    1584.78    1584.80     -0.02     684 -    698     0   AEALEAAIENLNEAK
  1830.93    1829.92    1829.93     -0.01     154 -    168     1   LYTALQQYFQNGEKK
  2197.07    2196.06    2196.11     -0.05     211 -    229     1   QAEFFLSQQASLLKNDETK
  3008.33    3007.32    3007.51     -0.19     519 -    546     1   AMNDGKYHLADSLVTGITALNSIEGVYR
No match to: 1012.75, 1153.65, 1192.65, 1202.73, 1222.67, 1251.72, 1855.94, 1979.98, 1987.03, 2062.03, 2499.25, 2512.75, 301

20. WMBEAK  Mass: 31149  Score: 41
38K phosphoprotein - Marek's disease virus (strain GA)
 Observed    Mr(expt)   Mr(calc)    Delta   Start     End  Miss  Peptide
  1504.94    1503.93    1503.82      0.11     185 -    197     1   QNPNILGEHLNKK
  3198.64    3197.63    3197.62      0.01     219 -    248     1   SYMLVTLICSAKSLLLGSCMSFFAGMLVGR
  4015.94    4014.93    4014.96     -0.03     148 -    184     1   QMGELAQQCEGGTYADLLVEAEQAVVHSVRALMLAER
No match to: 1012.75, 1153.65, 1192.65, 1202.73, 1222.67, 1251.72, 1585.79, 1830.93, 1855.94, 1979.98, 1987.03, 2062.03, 219
```

Search Parameters

```
Type of search         : Peptide Mass Fingerprint
Enzyme                 : Trypsin
Mass values            : Monoisotopic
Protein Mass           : Unrestricted
Peptide Mass Tolerance : ± 0.2 Da
Peptide Charge State   : 1+
Max Missed Cleavages   : 1
```

Mascot: http://www.matrixscience.com/

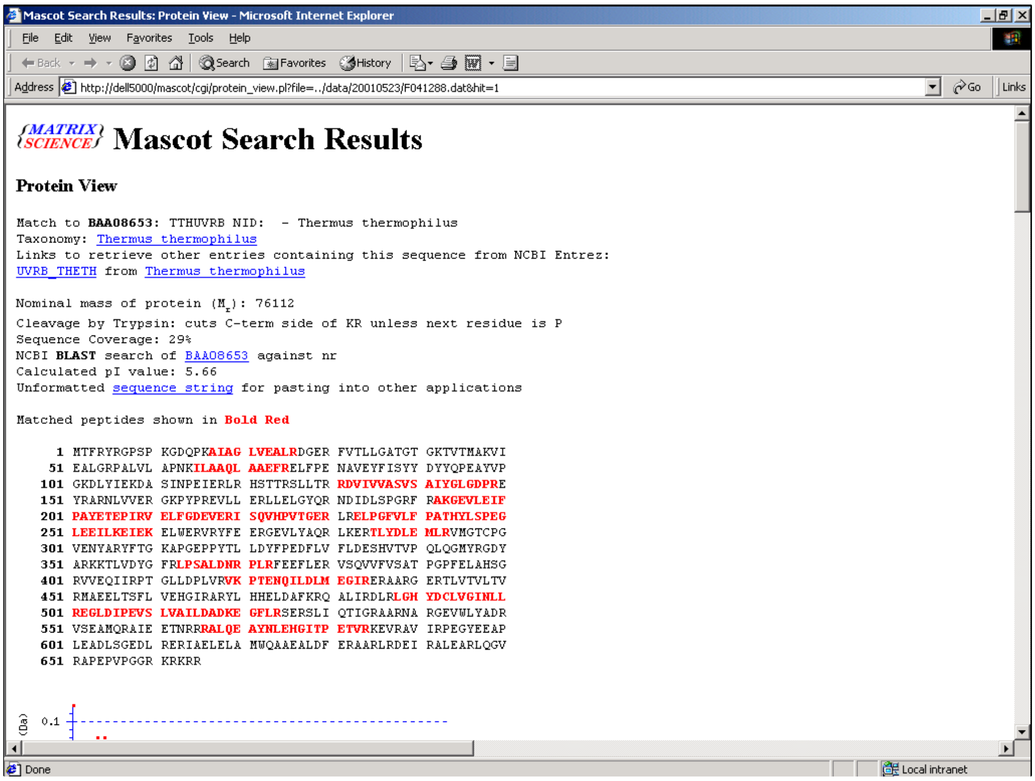
The bottom of the report showing the search parameter summary



Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/master_results.pl?REPTYPE=Protein&file=../data/20010523/F041288.dat
Go Links
Done
Local intranet

```
Results List

1. BAA08653  Mass: 76112  Score: 177
TTHUVRB NID:  - Thermus thermophilus
 Observed     Mr(expt)    Mr(calc)     Delta    Start      End  Miss  Peptide
  1012.75     1011.75     1011.61       0.14       17 -     26     0   AIAGLVEALR
  1153.65     1152.64     1152.58       0.06      285 -    293     0   TLYDLEMLR
  1192.65     1191.64     1191.58       0.06      210 -    219     0   VELFGDEVER
  1202.73     1201.72     1201.68       0.04       65 -     75     0   ILAAQLAAEFR
  1222.67     1221.66     1221.65       0.02      220 -    230     0   ISQVHPVTGER
  1251.72     1250.71     1250.71       0.00      363 -    373     0   LPSALDNRPLR
  1585.79     1584.78     1584.84      -0.06      488 -    501     0   LGHYDCLVGINLLR
  1830.93     1829.92     1829.99      -0.07      132 -    149     0   DVIVVASVSAIYGLGDPR
  1855.94     1854.93     1854.99      -0.06      419 -    434     0   VKPTENQILDLMEGIR
  1987.03     1986.02     1986.09      -0.07      131 -    149     1   RDVIVVASVSAIYGLGDPR
  2062.03     2061.02     2061.08      -0.06      192 -    209     1   AKGEVLEIFPAYETEPIR
  2197.07     2196.06     2196.13      -0.07      566 -    584     1   RALQEAYNLEHGITPETVR
  2499.25     2498.24     2498.33      -0.08      502 -    524     1   EGLDIPEVSLVAILDADKEGFLR
  3198.64     3197.63     3197.69      -0.06      233 -    260     1   ELPGFVLFPATHYLSPEGLEEILKEIEK
No match to: 1504.94, 1979.98, 2512.75, 3008.33, 3019.95, 3207.49, 3912.15, 4015.94

2. 1D2MA  Mass: 63099  Score: 114
excinuclease abc subunit b - bacteria (fragments)
 Observed     Mr(expt)    Mr(calc)     Delta    Start      End  Miss  Peptide
  1012.75     1011.75     1011.61       0.14       16 -     25     0   AIAGLVEALR
  1153.65     1152.64     1152.58       0.06      254 -    262     0   TLYDLEMLR
  1202.73     1201.72     1201.68       0.04       64 -     74     0   ILAAQLAAEFR
  1251.72     1250.71     1250.71       0.00      332 -    342     0   LPSALDNRPLR
  1585.79     1584.78     1584.84      -0.06      457 -    470     0   LGHYDCLVGINLLR
  1830.93     1829.92     1829.99      -0.07      131 -    148     0   DVIVVASVSAIYGLGDPR
  1855.94     1854.93     1854.99      -0.06      388 -    403     0   VKPTENQILDLMEGIR
  1987.03     1986.02     1986.09      -0.07      130 -    148     1   RDVIVVASVSAIYGLGDPR
  2062.03     2061.02     2061.08      -0.06      184 -    201     1   AKGEVLEIFPAYETEPIR
  2499.25     2498.24     2498.33      -0.08      471 -    493     1   EGLDIPEVSLVAILDADKEGFLR
No match to: 1192.65, 1222.67, 1504.94, 1979.98, 2197.07, 2512.75, 3008.33, 3019.95, 3198.64, 3207.49, 3912.15, 4015.94

3. Q9ZR01  Mass: 143743  Score: 51
HYPOTHETICAL 143.8 KDA PROTEIN.- Arabidopsis thaliana (Mouse-ear cress).
 Observed     Mr(expt)    Mr(calc)     Delta    Start      End  Miss  Peptide
```

The protein summary report tabulates details of the matches for the top hits. Here, we can see that hit 2 is not a different protein, it is just a fragment of hit 1.

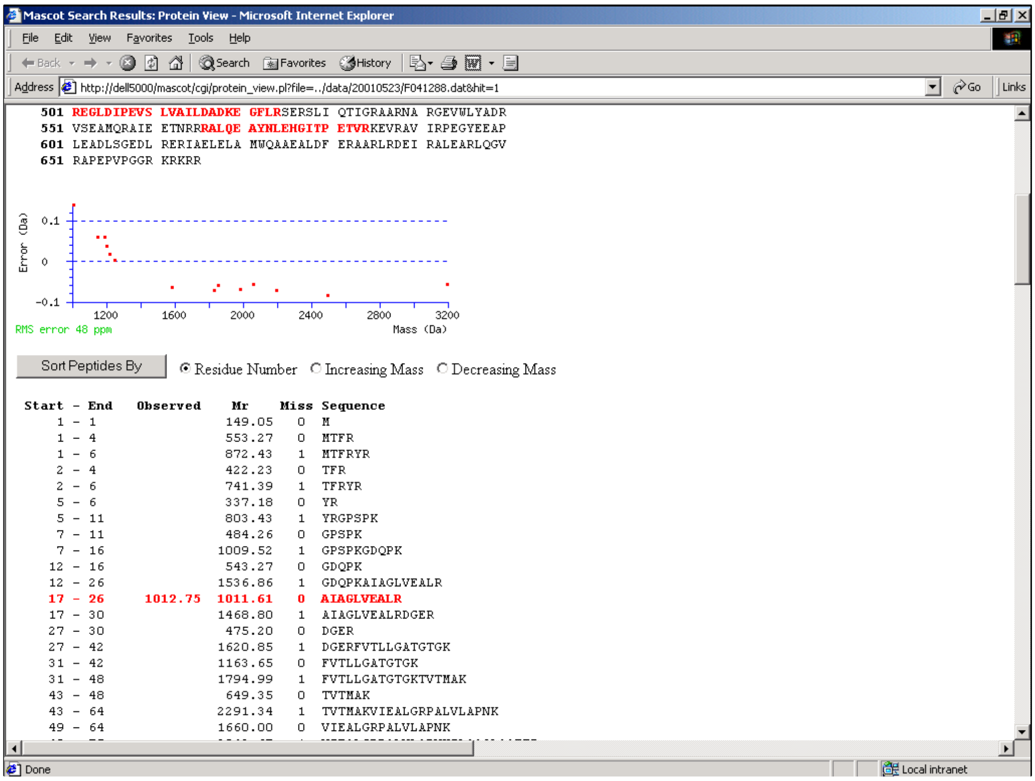

Mascot Search Results: Protein View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/protein_view.pl?file=../data/20010523/F041288.dat&hit=1
Go Links
MATRIX SCIENCE
(Da)
0.1
Done
Local intranet

Mascot Search Results

Protein View

```
Match to BAA08653: TTHUVRB NID:  - Thermus thermophilus
Taxonomy: Thermus thermophilus
Links to retrieve other entries containing this sequence from NCBI Entrez:
UVRB_THETH from Thermus thermophilus

Nominal mass of protein (Mr): 76112
Cleavage by Trypsin: cuts C-term side of KR unless next residue is P
Sequence Coverage: 29%
NCBI BLAST search of BAA08653 against nr
Calculated pI value: 5.66
Unformatted sequence string for pasting into other applications

Matched peptides shown in Bold Red

    1 MTFRYRGPSP KGDQPKAIAG LVEALRDGER FVTLLGATGT GKTVTMAKVI
   51 EALGRPALVL APNKILAAQL AAEFRELFPE NAVEYFISYY DYYQPEAYVP
  101 GKDLYIEKDA SINPEIERLR HSTTRSLLTR RDVIVVASVS AIYGLGDPRE
  151 YRARNLVVER GKPYPREVLL ERLLELGYQR NDIDLSPGRF RAKGEVLEIF
  201 PAYETEPIRV ELFGDEVERI SQVHPVTGER LRELPGFVLF PATHYLSPEG
  251 LEEILKEIEK ELWERVRYFE ERGEVLYAQR LKERTLYDLE MLRVMGTCPG
  301 VENYARYFTG KAPGEPPYTL LDYFPEDFLV FLDESHVTVP QLQGMYRGDY
  351 ARKKTLVDYG FRLPSALDNR PLRFEEFLER VSQVVFVSAT PGPFELAHSG
  401 RVVEQIIRPT GLLDPLVRVK PTENQILDLM EGIRERAARG ERTLVTVLTV
  451 RMAEELTSFL VEHGIRARYL HHELDAFKRQ ALIRDLRLGH YDCLVGINLL
  501 REGLDIPEVS LVAILDADKE GFLRSERSLI QTIGRAARNA RGEVWLYADR
  551 VSEAMQRAIE ETNRRRALQE AYNLEHGITP ETVRKEVRAV IRPEGYEEAP
  601 LEADLSGEDL RERIAELELA MWQAAEALDF ERAARLRDEI RALEARLQGV
  651 RAPEPVPGGR KRKRR
```

Clicking on the accession number link leads to the protein view report.



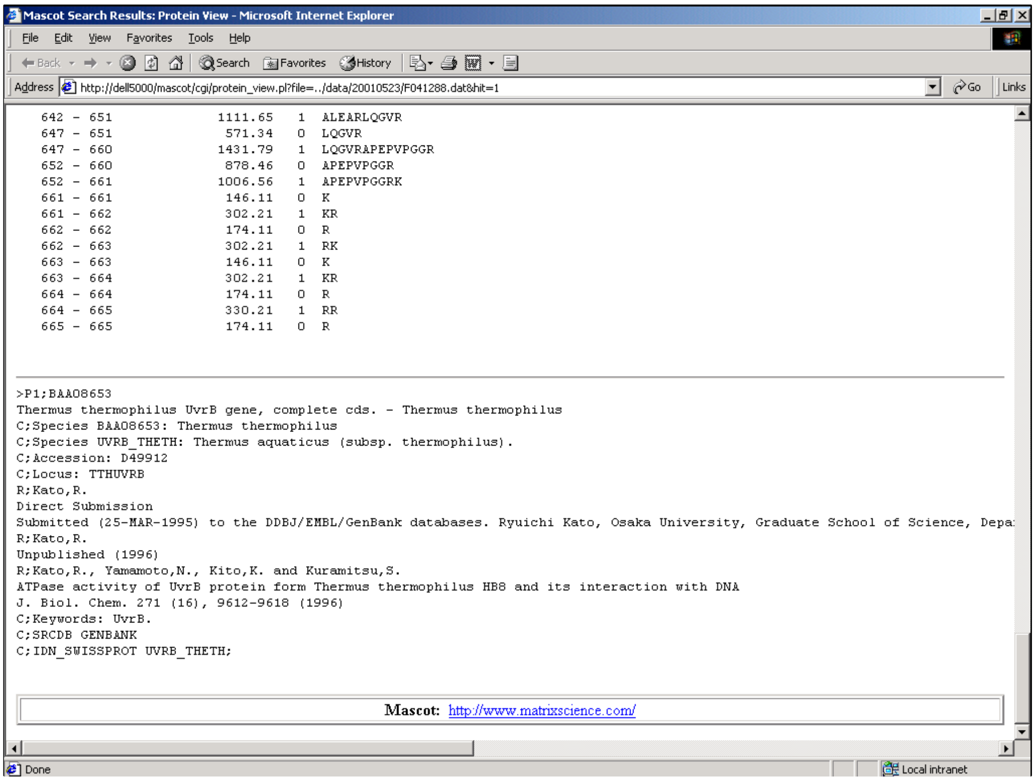
Mascot Search Results: Protein View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/protein_view.pl?file=../data/20010523/F041288.dat&hit=1
Go Links
Error (Da)
0.1
0
-0.1
1200 1600 2000 2400 2800 3200
Mass (Da)
RMS error 48 ppm
Sort Peptides By
Done
Local intranet

```
501 REGLDIPEVS LVAILDADKE GFLRSERSLI QTIGRAARNA RGEVWLYADR
551 VSEAMQRAIE ETNRRRALQE AYNLEHGITP ETVRKEVRAV IRPEGYEEAP
601 LEADLSGEDL RERIAELELA MWQAAEALDF ERAARLRDEI RALEARLQGV
651 RAPEPVPGGR KRKRR
```

● Residue Number ○ Increasing Mass ○ Decreasing Mass

```
Start - End   Observed    Mr     Miss Sequence
    1 - 1                149.05    0  M
    1 - 4                553.27    0  MTFR
    1 - 6                872.43    1  MTFRYR
    2 - 4                422.23    0  TFR
    2 - 6                741.39    1  TFRYR
    5 - 6                337.18    0  YR
    5 - 11               803.43    1  YRGPSPK
    7 - 11               484.26    0  GPSPK
    7 - 16              1009.52    1  GPSPKGDQPK
   12 - 16               543.27    0  GDQPK
   12 - 26              1536.86    1  GDQPKAIAGLVEALR
   17 - 26    1012.75  1011.61    0  AIAGLVEALR
   17 - 30              1468.80    1  AIAGLVEALRDGER
   27 - 30               475.20    0  DGER
   27 - 42              1620.85    1  DGERFVTLLGATGTGK
   31 - 42              1163.65    0  FVTLLGATGTGK
   31 - 48              1794.99    1  FVTLLGATGTGKTVTMAK
   43 - 48               649.35    0  TVTMAK
   43 - 64              2291.34    1  TVTMAKVIEALGRPALVLAPNK
   49 - 64              1660.00    0  VIEALGRPALVLAPNK
```

Besides the error graph mentioned earlier, the protein view also shows the hits highlighted on the protein sequence and a table of all the peptides from the *in silico* digest.

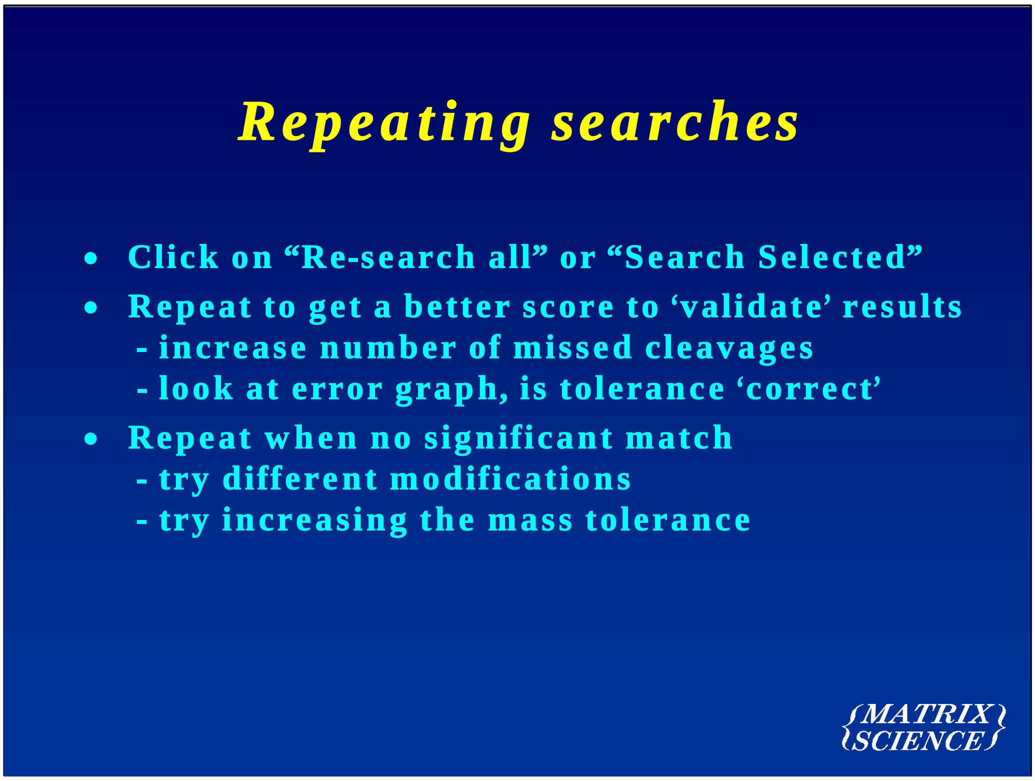

Mascot Search Results: Protein View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/protein_view.pl?file=../data/20010523/F041288.dat&hit=1
Go Links

```
   642 - 651             1111.65   1  ALEARLQGVR
   647 - 651              571.34   0  LQGVR
   647 - 660             1431.79   1  LQGVRAPEPVPGGR
   652 - 660              878.46   0  APEPVPGGR
   652 - 661             1006.56   1  APEPVPGGRK
   661 - 661              146.11   0  K
   661 - 662              302.21   1  KR
   662 - 662              174.11   0  R
   662 - 663              302.21   1  RK
   663 - 663              146.11   0  K
   663 - 664              302.21   1  KR
   664 - 664              174.11   0  R
   664 - 665              330.21   1  RR
   665 - 665              174.11   0  R
```

```
>P1;BAA08653
Thermus thermophilus UvrB gene, complete cds. - Thermus thermophilus
C;Species BAA08653: Thermus thermophilus
C;Species UVRB_THETH: Thermus aquaticus (subsp. thermophilus).
C;Accession: D49912
C;Locus: TTHUVRB
R;Kato,R.
Direct Submission
Submitted (25-MAR-1995) to the DDBJ/EMBL/GenBank databases. Ryuichi Kato, Osaka University, Graduate School of Science, Depa
R;Kato,R.
Unpublished (1996)
R;Kato,R., Yamamoto,N., Kito,K. and Kuramitsu,S.
ATPase activity of UvrB protein form Thermus thermophilus HB8 and its interaction with DNA
J. Biol. Chem. 271 (16), 9612-9618 (1996)
C;Keywords: UvrB.
C;SRCDB GENBANK
C;IDN_SWISSPROT UVRB_THETH;
```

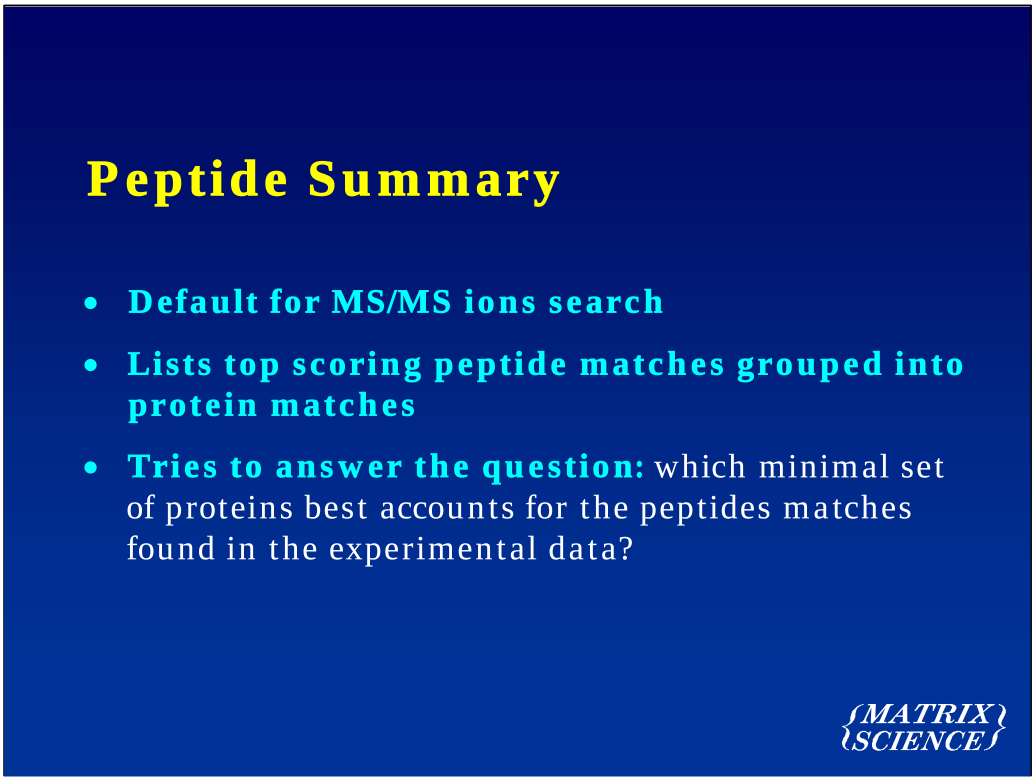
Mascot: http://www.matrixscience.com/

If available, the full annotation text is displayed at the bottom of the protein view.



Repeating searches

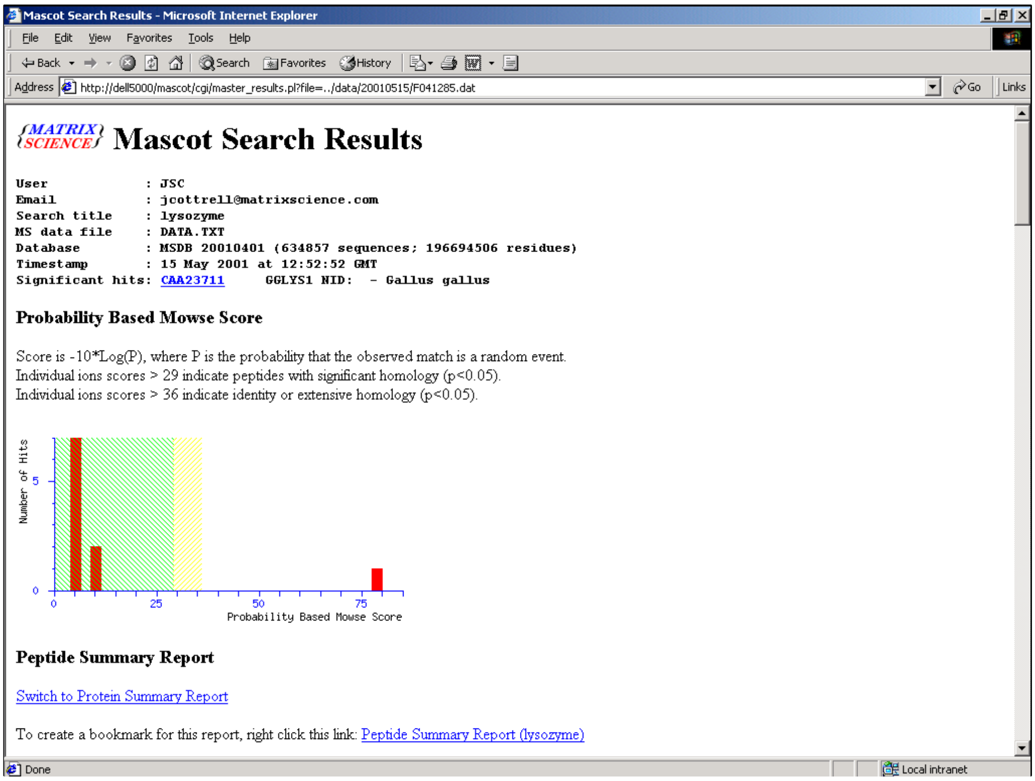
- Click on “Re-search all” or “Search Selected”
- Repeat to get a better score to ‘validate’ results
  - increase number of missed cleavages
  - look at error graph, is tolerance ‘correct’
- Repeat when no significant match
  - try different modifications
  - try increasing the mass tolerance



Peptide Summary

- **Default for MS/MS ions search**
- **Lists top scoring peptide matches grouped into protein matches**
- **Tries to answer the question:** which minimal set of proteins best accounts for the peptides matches found in the experimental data?

{MATRIX}
{SCIENCE}


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20010515/F041285.dat
Go Links
MATRIX SCIENCE
Done
Local intranet

Mascot Search Results

```
User            : JSC
Email           : jcottrell@matrixscience.com
Search title    : lysozyme
MS data file    : DATA.TXT
Database        : MSDB 20010401 (634857 sequences; 196694506 residues)
Timestamp       : 15 May 2001 at 12:52:52 GMT
Significant hits: CAA23711      GGLYS1 NID:  - Gallus gallus
```

Probability Based Mowse Score

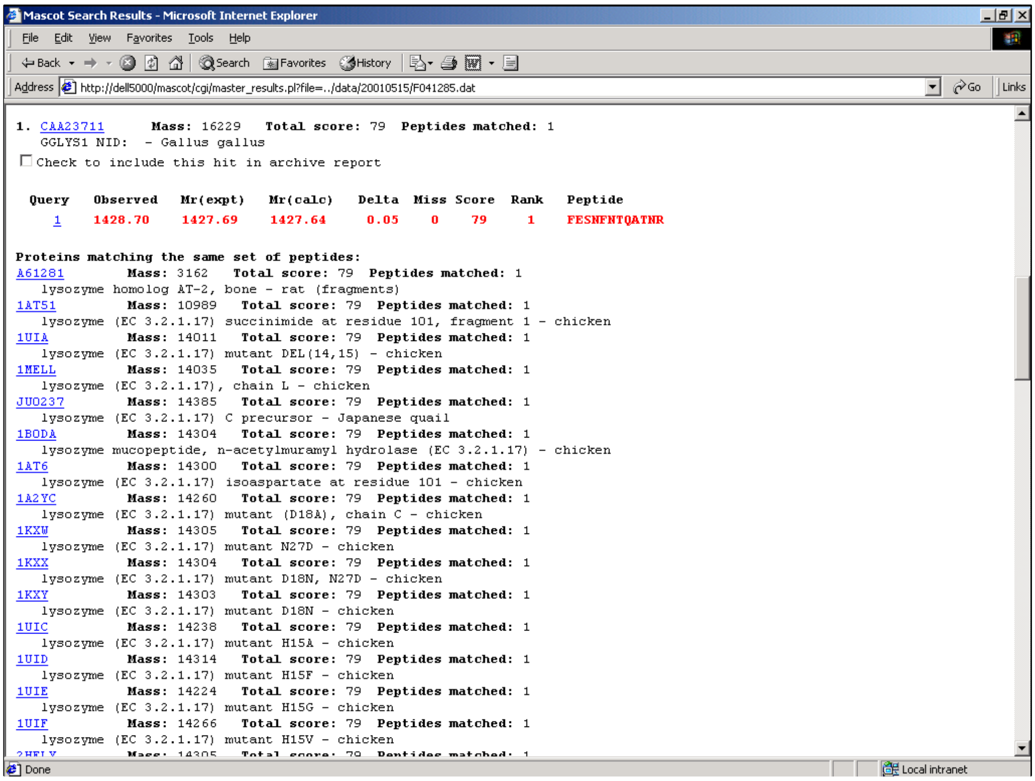
Score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 29 indicate peptides with significant homology (p<0.05).
Individual ions scores > 36 indicate identity or extensive homology (p<0.05).

Peptide Summary Report

Switch to Protein Summary Report

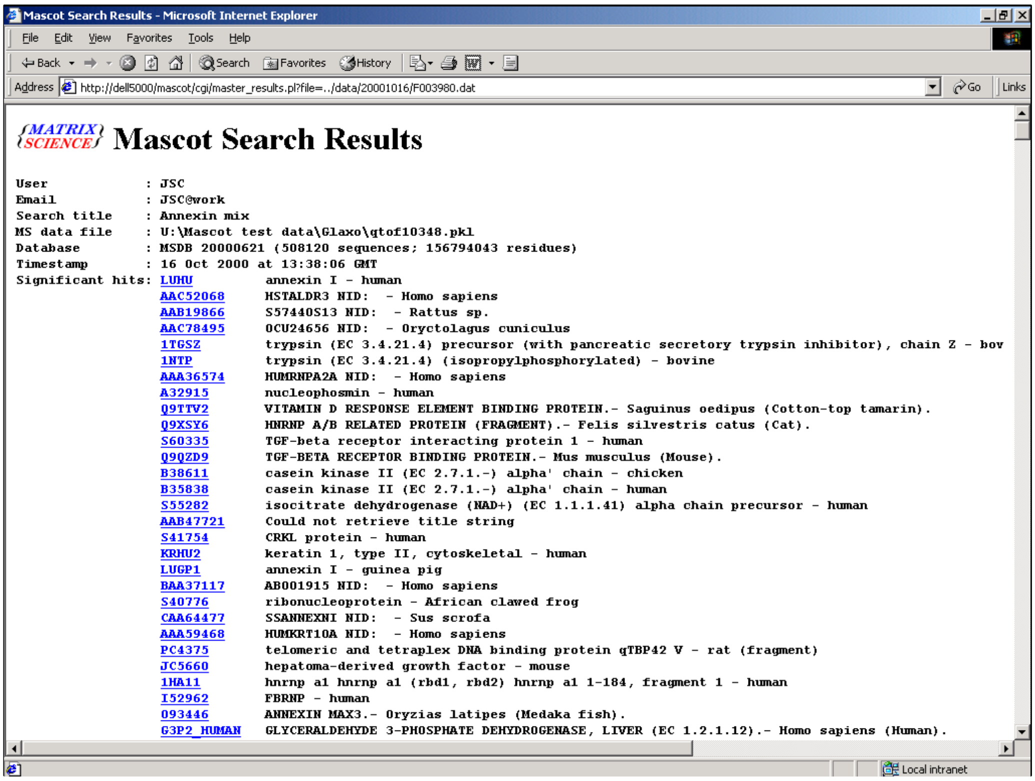
To create a bookmark for this report, right click this link: Peptide Summary Report (lysozyme)

When we have just a single MS/MS spectrum, life is simple. Either we get a peptide match, or we don't.


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20010515/F041285.dat
Go Links
Done
Local intranet

```
1. CAA23711      Mass: 16229   Total score: 79  Peptides matched: 1
   GGLYS1 NID:  - Gallus gallus
□ Check to include this hit in archive report

 Query   Observed   Mr(expt)   Mr(calc)   Delta  Miss Score  Rank   Peptide
    1     1428.70    1427.69    1427.64    0.05    0    79     1    FESNFNTQATNR

Proteins matching the same set of peptides:
A61281        Mass: 3162   Total score: 79  Peptides matched: 1
   lysozyme homolog AT-2, bone - rat (fragments)
1AT51         Mass: 10989   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) succinimide at residue 101, fragment 1 - chicken
1UIA          Mass: 14011   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant DEL(14,15) - chicken
1MELL         Mass: 14035   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17), chain L - chicken
JU0237        Mass: 14385   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) C precursor - Japanese quail
1BODA         Mass: 14304   Total score: 79  Peptides matched: 1
   lysozyme mucopeptide, n-acetylmuramyl hydrolase (EC 3.2.1.17) - chicken
1AT6          Mass: 14300   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) isoaspartate at residue 101 - chicken
1A2YC         Mass: 14260   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant (D18A), chain C - chicken
1KXW          Mass: 14305   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant N27D - chicken
1KXX          Mass: 14304   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant D18N, N27D - chicken
1KXY          Mass: 14303   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant D18N - chicken
1UIC          Mass: 14238   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant H15A - chicken
1UID          Mass: 14314   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant H15F - chicken
1UIE          Mass: 14224   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant H15G - chicken
1UIF          Mass: 14266   Total score: 79  Peptides matched: 1
   lysozyme (EC 3.2.1.17) mutant H15V - chicken
2HFLY         Mass: 14305   Total score: 79  Peptides matched: 1
```

If we get a match,and the peptide is unique to one protein family, we have a protein match

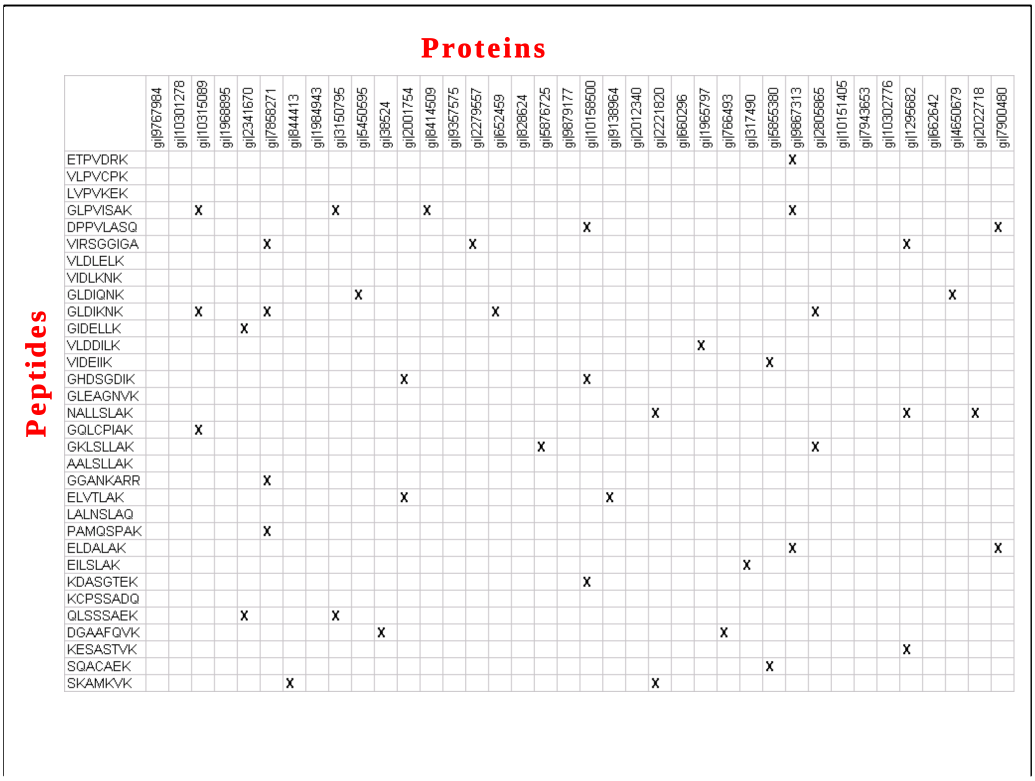


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F003980.dat
Go Links
MATRIX SCIENCE
Local intranet

Mascot Search Results

```
User             : JSC
Email            : JSC@work
Search title     : Annexin mix
MS data file     : U:\Mascot test data\Glaxo\qtof10348.pkl
Database         : MSDB 20000621 (508120 sequences; 156794043 residues)
Timestamp        : 16 Oct 2000 at 13:38:06 GMT
Significant hits: LUHU         annexin I - human
                  AAC52068     HSTALDR3 NID:  - Homo sapiens
                  AAB19866     S57440S13 NID:  - Rattus sp.
                  AAC78495     OCU24656 NID:  - Oryctolagus cuniculus
                  1TGSZ        trypsin (EC 3.4.21.4) precursor (with pancreatic secretory trypsin inhibitor), chain Z - bov
                  1NTP         trypsin (EC 3.4.21.4) (isopropylphosphorylated) - bovine
                  AAA36574     HUMRNPA2A NID:  - Homo sapiens
                  A32915       nucleophosmin - human
                  Q9TTV2       VITAMIN D RESPONSE ELEMENT BINDING PROTEIN.- Saguinus oedipus (Cotton-top tamarin).
                  Q9XSY6       HNRNP A/B RELATED PROTEIN (FRAGMENT).- Felis silvestris catus (Cat).
                  S60335       TGF-beta receptor interacting protein 1 - human
                  Q9QZD9       TGF-BETA RECEPTOR BINDING PROTEIN.- Mus musculus (Mouse).
                  B38611       casein kinase II (EC 2.7.1.-) alpha' chain - chicken
                  B35838       casein kinase II (EC 2.7.1.-) alpha' chain - human
                  S55282       isocitrate dehydrogenase (NAD+) (EC 1.1.1.41) alpha chain precursor - human
                  AAB47721     Could not retrieve title string
                  S41754       CRKL protein - human
                  KRHU2        keratin 1, type II, cytoskeletal - human
                  LUGP1        annexin I - guinea pig
                  BAA37117     AB001915 NID:  - Homo sapiens
                  S40776       ribonucleoprotein - African clawed frog
                  CAA64477     SSANNEXNI NID:  - Sus scrofa
                  AAA59468     HUMKRT10A NID:  - Homo sapiens
                  PC4375       telomeric and tetraplex DNA binding protein qTBP42 V - rat (fragment)
                  JC5660       hepatoma-derived growth factor - mouse
                  1HA11        hnrnp a1 hnrnp a1 (rbd1, rbd2) hnrnp a1 1-184, fragment 1 - human
                  I52962       FBRNP - human
                  O93446       ANNEXIN MAX3.- Oryzias latipes (Medaka fish).
                  G3P2_HUMAN   GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (EC 1.2.1.12).- Homo sapiens (Human).
```

However, if we have a complex data set, containing many MS/MS spectra which match to peptides from a number of different proteins, trying to report which proteins have been identified becomes more subjective.



Proteins

Peptides

	gi|9767984	gi|10301278	gi|10315089	gi|1968895	gi|2341670	gi|7858271	gi|844413	gi|1984943	gi|3150795	gi|5450595	gi|38524	gi|2001754	gi|8414509	gi|9357575	gi|2279557	gi|652459	gi|828624	gi|5876725	gi|9879177	gi|10158500	gi|9138964	gi|2012340	gi|2221820	gi|660296	gi|1965797	gi|766493	gi|317490	gi|5855380	gi|9867313	gi|2805865	gi|10151405	gi|7943653	gi|10302776	gi|1295682	gi|662642	gi|4650679	gi|2022718	gi|7900480
ETPVDRK																													X									
VLPVCPK																																						
LVPVKEK																																						
GLPVISAK			X						X				X																X									
DPPVLASQ																				X																		X
VIRSGGIGA						X									X																			X				
VLDLELK																																						
VIDLKNK																																						
GLDIQNK										X																										X		
GLDIKNK			X			X										X														X								
GIDELLK					X																																	
VLDDILK																									X													
VIDEIIK																												X										
GHDSGDIK												X								X																		
GLEAGNVK																																						
NALLSLAK																							X											X			X	
GQLCPIAK			X																																			
GKLSLLAK																		X												X								
AALSLLAK																																						
GGANKARR						X																																
ELVTLAK												X									X																	
LALNSLAQ																																						
PAMQSPAK						X																																
ELDALAK																													X									X
EILSLAK																											X											
KDASGTEK																				X																		
KCPSSADQ																																						
QLSSSAEK					X				X																													
DGAAFQVK											X															X												
KESASTVK																																		X				
SQACAEK																												X										
SKAMKVK							X																X															

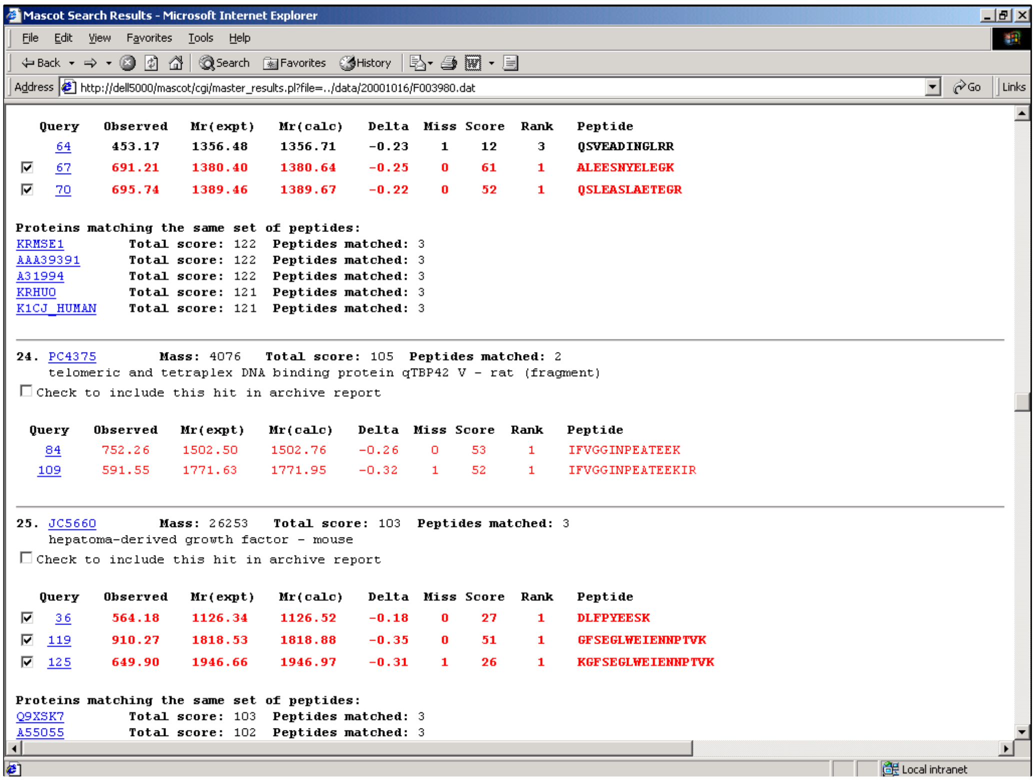
When Mascot searches MS/MS data, it is getting peptide matches.

Looking at the peptide matches and trying to determine which proteins were present is a secondary process, which is actually done by the report script.

We can think of the results from a Mascot search of an LC-MS/MS search as a huge matrix. The columns are proteins and the rows are peptides.

This isn’t a diagonal matrix, with just one cross in each row or column. In most cases, a peptide match can be found in several proteins. And, very often, a protein will contain several peptide matches.

To produce a simple, linear list of protein matches, we take the column with the highest score, and call that protein hit number 1. Any other proteins which match the same set of peptides, or a subset, are considered to be equivalent, but inferior matches, and collapsed into the same hit. These proteins are removed from the matrix, and we then look for the next highest scoring column … and so on.

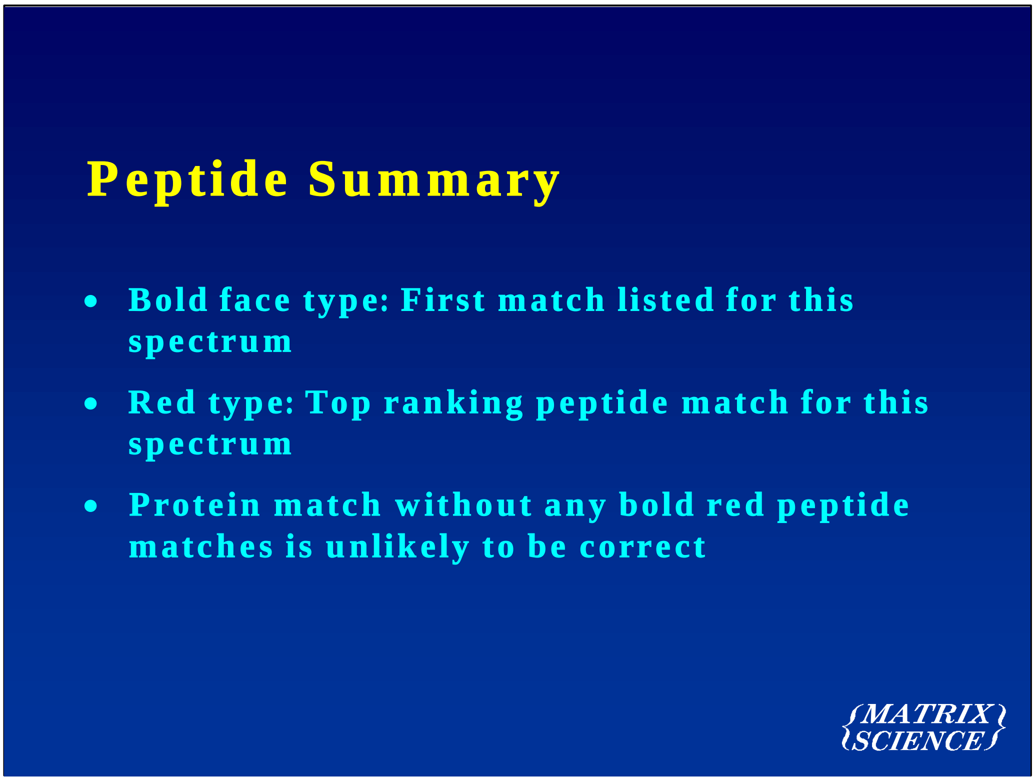


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F003980.dat
Go Links
Local intranet

```
    Query   Observed   Mr(expt)   Mr(calc)    Delta  Miss Score  Rank    Peptide
     64      453.17    1356.48    1356.71    -0.23    1    12     3     QSVEADINGLRR
☑    67      691.21    1380.40    1380.64    -0.25    0    61     1     ALEESNYELEGK
☑    70      695.74    1389.46    1389.67    -0.22    0    52     1     QSLEASLAETEGR

Proteins matching the same set of peptides:
KRMSE1         Total score: 122  Peptides matched: 3
AAA39391       Total score: 122  Peptides matched: 3
A31994         Total score: 122  Peptides matched: 3
KRHUO          Total score: 121  Peptides matched: 3
K1CJ_HUMAN     Total score: 121  Peptides matched: 3

24. PC4375         Mass: 4076   Total score: 105  Peptides matched: 2
    telomeric and tetraplex DNA binding protein qTBP42 V - rat (fragment)
□ Check to include this hit in archive report

   Query   Observed   Mr(expt)   Mr(calc)    Delta  Miss Score  Rank    Peptide
    84      752.26    1502.50    1502.76    -0.26    0    53     1     IFVGGINPEATEEK
   109      591.55    1771.63    1771.95    -0.32    1    52     1     IFVGGINPEATEEKIR

25. JC5660         Mass: 26253   Total score: 103  Peptides matched: 3
    hepatoma-derived growth factor - mouse
□ Check to include this hit in archive report

    Query   Observed   Mr(expt)   Mr(calc)    Delta  Miss Score  Rank    Peptide
☑    36      564.18    1126.34    1126.52    -0.18    0    27     1     DLFPYEESK
☑   119      910.27    1818.53    1818.88    -0.35    0    51     1     GFSEGLWEIENNPTVK
☑   125      649.90    1946.66    1946.97    -0.31    1    26     1     KGFSEGLWEIENNPTVK

Proteins matching the same set of peptides:
Q9XSK7         Total score: 103  Peptides matched: 3
A55055         Total score: 102  Peptides matched: 3
```

In the reports, we try to provide clues as to the most likely assignments. We use red to indicate that a peptide match is the top ranking match. We use bold type to indicate that this is the first time in the report that we have listed a match to a particular spectrum.

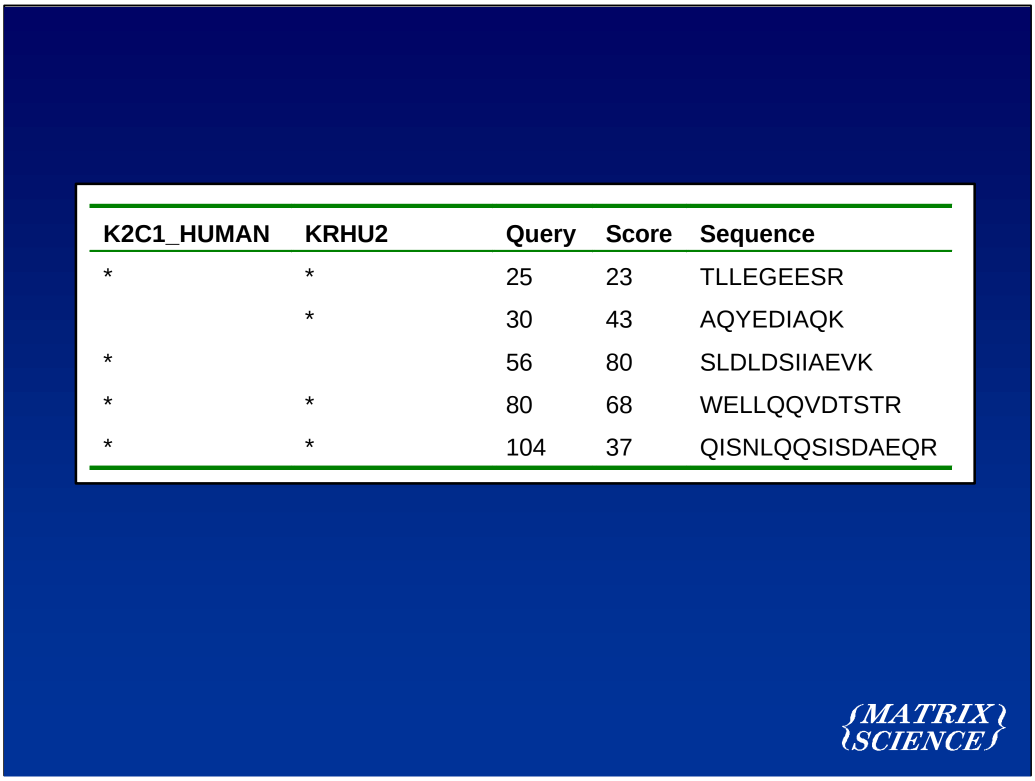
So, Hit 24 has two nice, top-ranking matches, but they are not in bold face type. This indicates that we have already seen matches to these spectra in earlier, i.e. higher scoring, proteins, which probably means that this protein match is spurious … but one can't be sure.



Peptide Summary

- Bold face type: First match listed for this spectrum
- Red type: Top ranking peptide match for this spectrum
- Protein match without any bold red peptide matches is unlikely to be correct

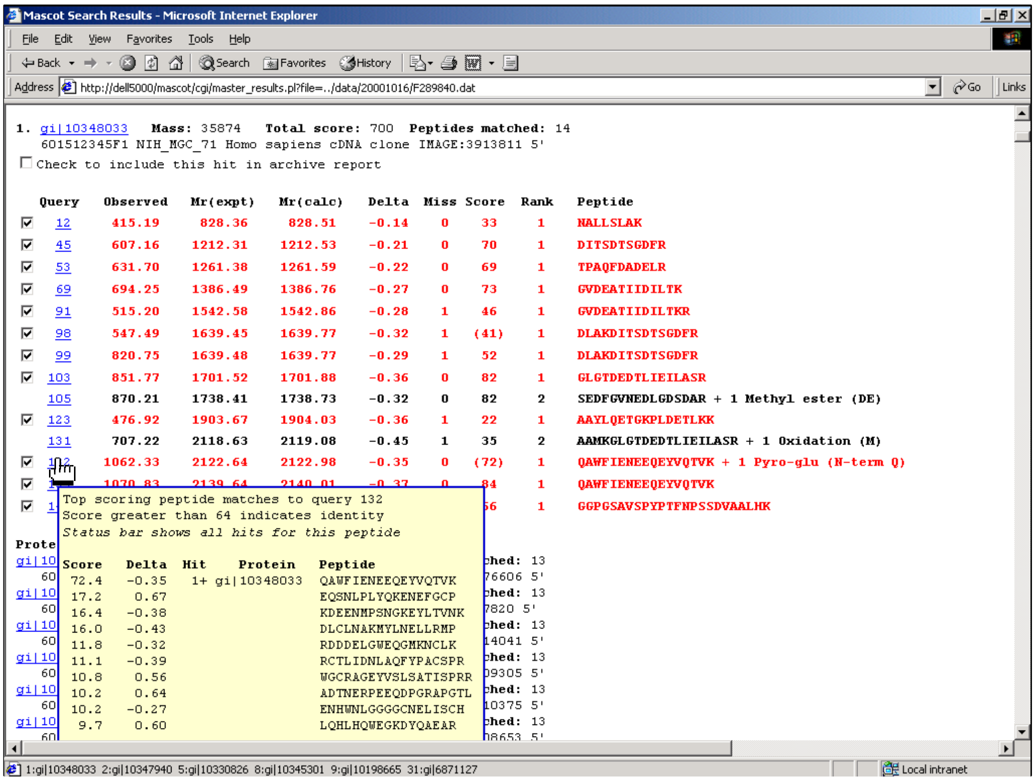
{MATRIX}
{SCIENCE}



K2C1_HUMAN	KRHU2	Query	Score	Sequence
*	*	25	23	TLLEGEESR
	*	30	43	AQYEDIAQK
*		56	80	SLDLDSIIAEVK
*	*	80	68	WELLQQVDTSTR
*	*	104	37	QISNLQQSISDAEQR

The peptide summary report represents one reasonable interpretation of the results. Sometimes, there are alternatives which cannot be resolved. For example, we might have this situation, where there are four matches to one keratin and four matches to another keratin.

It could be that only the left hand keratin was actually present in the sample, and the match to AQYEDIAQK is unreliable, or belongs to a different protein. Or, it could be that the keratin in the sample was a variant, not present in the database, which contains all five peptide matches. There are several other possible interpretations, and we cannot be certain which is correct.


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
History
Address
Go
Links

http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F289840.dat

1. gi|10348033 **Mass:** 35874 **Total score:** 700 **Peptides 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	Rank	Peptide
☑	12	415.19	828.36	828.51	-0.14	0	33	1	NALLSLAK
☑	45	607.16	1212.31	1212.53	-0.21	0	70	1	DITSDTSGDFR
☑	53	631.70	1261.38	1261.59	-0.22	0	69	1	TPAQFDADELR
☑	69	694.25	1386.49	1386.76	-0.27	0	73	1	GVDEATIIDILTK
☑	91	515.20	1542.58	1542.86	-0.28	1	46	1	GVDEATIIDILTKR
☑	98	547.49	1639.45	1639.77	-0.32	1	(41)	1	DLAKDITSDTSGDFR
☑	99	820.75	1639.48	1639.77	-0.29	1	52	1	DLAKDITSDTSGDFR
☑	103	851.77	1701.52	1701.88	-0.36	0	82	1	GLGTDEDTLIEILASR
	105	870.21	1738.41	1738.73	-0.32	0	82	2	SEDFGVNEDLGDSDAR + 1 Methyl ester (DE)
☑	123	476.92	1903.67	1904.03	-0.36	1	22	1	AAYLQETGKPLDETLKK
	131	707.22	2118.63	2119.08	-0.45	1	35	2	AAMKGLGTDEDTLIEILASR + 1 Oxidation (M)
☑	1[illegible]2	1062.33	2122.64	2122.98	-0.35	0	(72)	1	QAWFIENEEQEYVQTVK + 1 Pyro-glu (N-term Q)
☑	1	1070.83	2139.64	2140.01	-0.37	0	84	1	QAWFIENEEQEYVQTVK
☑	1						56	1	GGPGSAVSPYPTFNPSSDVAALHK

Top scoring peptide matches to query 132
Score greater than 64 indicates identity
Status bar shows all hits for this peptide

Score	Delta	Hit	Protein	Peptide
72.4	-0.35	1+	gi|10348033	QAWFIENEEQEYVQTVK
17.2	0.67			EQSNLPLYQKENEFGCP
16.4	-0.38			KDEENMPSNGKEYLTVNK
16.0	-0.43			DLCLNAKMYLNELLRMP
11.8	-0.32			RDDDELGWEQGMKNCLK
11.1	-0.39			RCTLIDNLAQFYPACSPR
10.8	0.56			WGCRAGEYVSLSATISPRR
10.2	0.64			ADTNERPEEQDPGRAPGTL
10.2	-0.27			ENHWNLGGGGCNELISCH
9.7	0.60			LQHLHQWEGKDYQAEAR

```
Prote
gi|10        hed: 13
   60        76606 5'
gi|10        hed: 13
   60        7820 5'
gi|10        hed: 13
   60        14041 5'
gi|10        hed: 13
   60        09305 5'
gi|10        hed: 13
   60        10375 5'
gi|10        hed: 13
   60        08653 5'
```

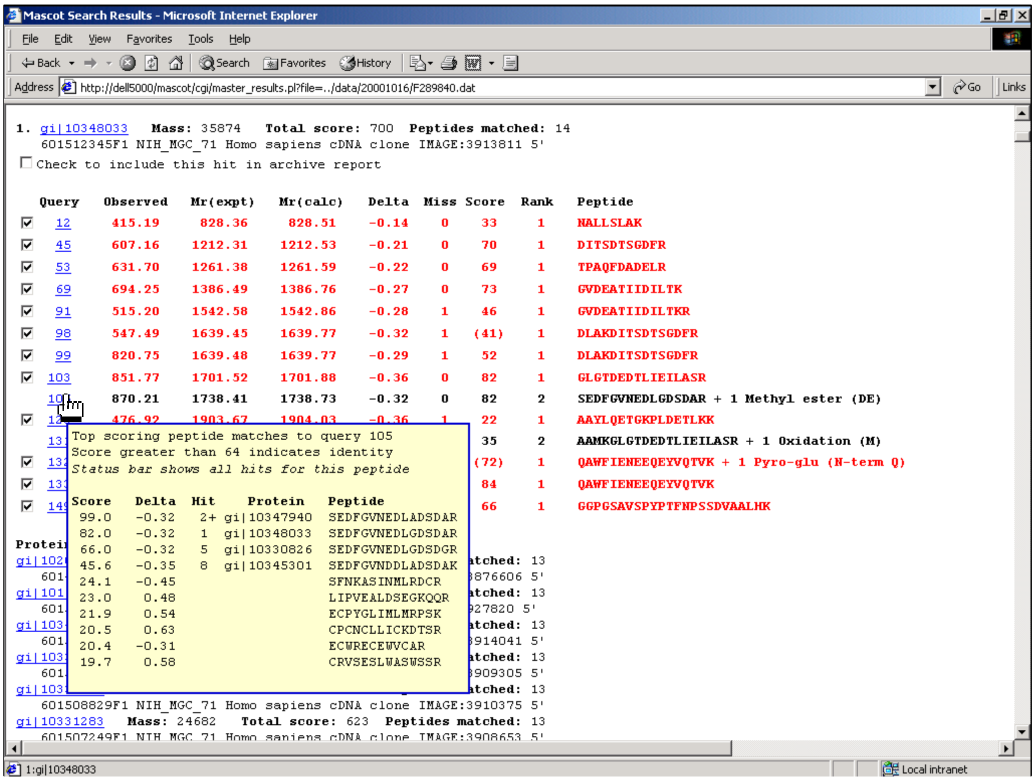
1:gi|10348033 2:gi|10347940 5:gi|10330826 8:gi|10345301 9:gi|10198665 31:gi|6871127

A search of a complete LC-MS/MS run generates a wealth of data, and presenting these results in a complex and intuitive fashion is not trivial.

Here, we have part of the Mascot report for such a search. A number of peptide matches have been assigned to a particular database entry.

For each peptide match listed in the main table, there may be better or worse matches to peptides from other entries in the database. These are shown in a pop-up window when the mouse cursor is held over a query number link.

In this example, we have one match with a high, and significant score. The remaining matches are random matches with random scores.

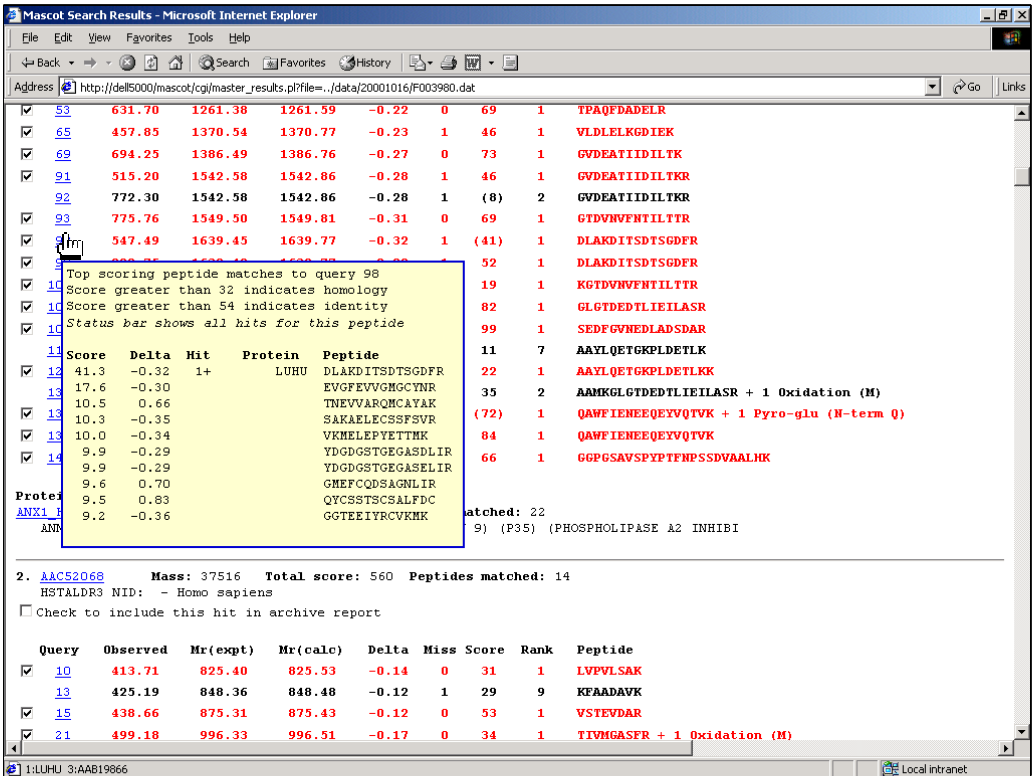


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F289840.dat
Go Links
1:gi|10348033
Local intranet

```
1. gi|10348033    Mass: 35874    Total score: 700   Peptides 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  Rank   Peptide
☑   12     415.19     828.36     828.51    -0.14    0    33      1    NALLSLAK
☑   45     607.16    1212.31    1212.53    -0.21    0    70      1    DITSDTSGDFR
☑   53     631.70    1261.38    1261.59    -0.22    0    69      1    TPAQFDADELR
☑   69     694.25    1386.49    1386.76    -0.27    0    73      1    GVDEATIIDILTK
☑   91     515.20    1542.58    1542.86    -0.28    1    46      1    GVDEATIIDILTKR
☑   98     547.49    1639.45    1639.77    -0.32    1   (41)     1    DLAKDITSDTSGDFR
☑   99     820.75    1639.48    1639.77    -0.29    1    52      1    DLAKDITSDTSGDFR
☑  103     851.77    1701.52    1701.88    -0.36    0    82      1    GLGTDEDTLIEILASR
   10      870.21    1738.41    1738.73    -0.32    0    82      2    SEDFGVNEDLGDSDAR + 1 Methyl ester (DE)
☑  12      476.92    1903.67    1904.03    -0.36    1    22      1    AAYLQETGKPLDETLKK
   13  Top scoring peptide matches to query 105   35      2    AAMKGLGTDEDTLIEILASR + 1 Oxidation (M)
☑  13  Score greater than 64 indicates identity  (72)     1    QAWFIENEEQEYVQTVK + 1 Pyro-glu (N-term Q)
☑  13  Status bar shows all hits for this peptide  84      1    QAWFIENEEQEYVQTVK
☑  14                                              66      1    GGPGSAVSPYPTFNPSSDVAALHK
       Score   Delta   Hit   Protein      Peptide
Protei  99.0   -0.32   2+ gi|10347940  SEDFGVNEDLADSDAR
gi|102  82.0   -0.32   1  gi|10348033  SEDFGVNEDLGDSDAR   atched: 13
   601  66.0   -0.32   5  gi|10330826  SEDFGVNEDLGDSDGR   876606 5'
gi|101  45.6   -0.35   8  gi|10345301  SEDFGVNDDLADSDAK   atched: 13
   601  24.1   -0.45                    SFNKASINMLRDCR     27820 5'
gi|103  23.0    0.48                    LIPVEALDSEGKQQR    atched: 13
   601  21.9    0.54                    ECPYGLIMLMRPSK     914041 5'
gi|103  20.5    0.63                    CPCNCLLICKDTSR     atched: 13
   601  20.4   -0.31                    ECWRECEWVCAR       909305 5'
gi|103  19.7    0.58                    CRVSESLWASWSSR     atched: 13
   601508829F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3910375 5'
gi|10331283    Mass: 24682    Total score: 623   Peptides matched: 13
   601507249F1 NIH_MGC_71 Homo sapiens cDNA clone IMAGE:3908653 5'
```

In contrast, here we see several non-random, significant matches, because there are four peptides in the database which are almost, but not quite, identical.

The peptide match to this protein has a very high score, but there is another sequence with a slightly higher score. Since this protein has several other excellent matches, we are faced with a question: which of the top two peptide matches do we believe? Does the analyte have a variant sequence from that in the database, and the top match is correct? Or, is the spectrum ambiguous, and there is insufficient information to differentiate the top two matches with confidence? Either is perfectly possible.



Mascot Search Results - Microsoft Internet Explorer

http://dell5000/mascot/cgi/master_results.pl?file=../data/20001016/F003980.dat

```
☑  53   631.70  1261.38  1261.59  -0.22  0   69    1   TPAQFDADELR
☑  65   457.85  1370.54  1370.77  -0.23  1   46    1   VLDLELKGDIEK
☑  69   694.25  1386.49  1386.76  -0.27  0   73    1   GVDEATIIDILTK
☑  91   515.20  1542.58  1542.86  -0.28  1   46    1   GVDEATIIDILTKR
   92   772.30  1542.58  1542.86  -0.28  1   (8)   2   GVDEATIIDILTKR
☑  93   775.76  1549.50  1549.81  -0.31  0   69    1   GTDVNVFNTILTTR
☑  9    547.49  1639.45  1639.77  -0.32  1   (41)  1   DLAKDITSDTSGDFR
☑  9                                         52    1   DLAKDITSDTSGDFR
☑  10                                        19    1   KGTDVNVFNTILTTR
☑  10                                        82    1   GLGTDEDTLIEILASR
☑  10                                        99    1   SEDFGVNEDLADSDAR
   11                                        11    7   AAYLQETGKPLDETLK
☑  12                                        22    1   AAYLQETGKPLDETLKK
   13                                        35    2   AAMKGLGTDEDTLIEILASR + 1 Oxidation (M)
☑  13                                        (72)  1   QAWFIENEEQEYVQTVK + 1 Pyro-glu (N-term Q)
☑  13                                        84    1   QAWFIENEEQEYVQTVK
☑  14                                        66    1   GGPGSAVSPYPTFNPSSDVAALHK

Protei
ANX1_H                                       atched: 22
  ANN                                         9) (P35) (PHOSPHOLIPASE A2 INHIBI
```

```
Top scoring peptide matches to query 98
Score greater than 32 indicates homology
Score greater than 54 indicates identity
Status bar shows all hits for this peptide

Score   Delta  Hit    Protein  Peptide
 41.3   -0.32   1+       LUHU  DLAKDITSDTSGDFR
 17.6   -0.30                  EVGFEVVGMGCYNR
 10.5    0.66                  TNEVVARQMCAYAK
 10.3   -0.35                  SAKAELECSSFSVR
 10.0   -0.34                  VKMELEPYETTMK
  9.9   -0.29                  YDGDGSTGEGASDLIR
  9.9   -0.29                  YDGDGSTGEGASELIR
  9.6    0.70                  GMEFCQDSAGNLIR
  9.5    0.83                  QYCSSTSCSALFDC
  9.2   -0.36                  GGTEEIYRCVKMK
```

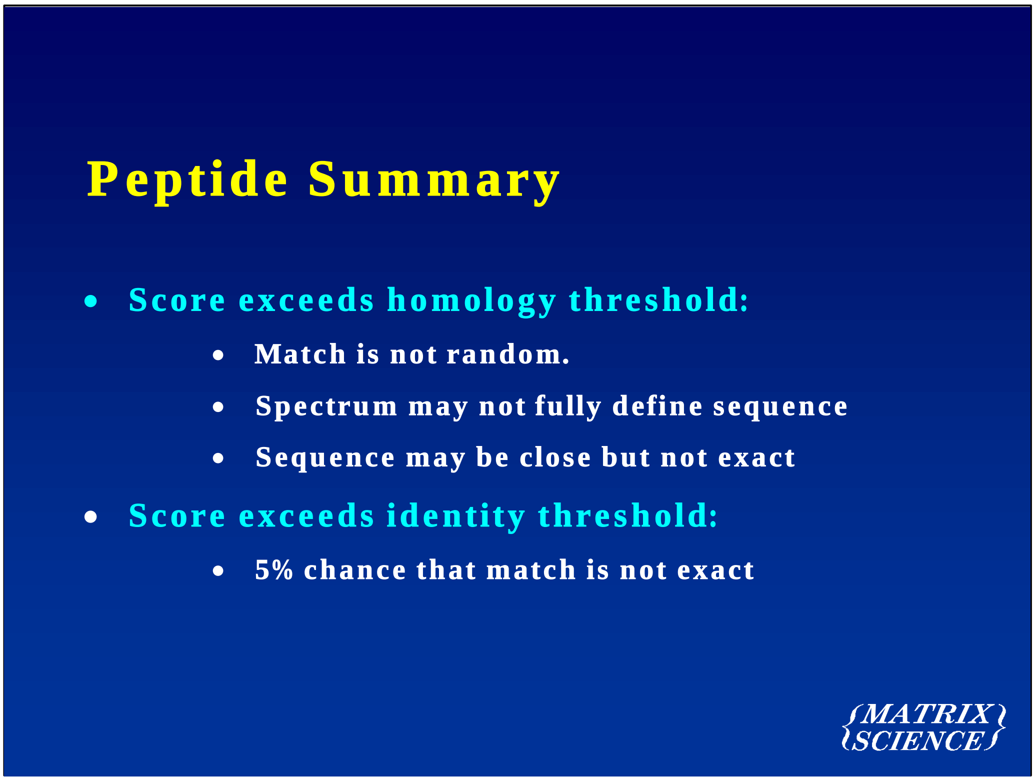
```
2. AAC52068      Mass: 37516   Total score: 560   Peptides matched: 14
   HSTALDR3 NID:  - Homo sapiens
□ Check to include this hit in archive report

   Query  Observed  Mr(expt)  Mr(calc)  Delta  Miss Score  Rank  Peptide
☑  10     413.71    825.40    825.53    -0.14  0     31     1    LVPVLSAK
   13     425.19    848.36    848.48    -0.12  1     29     9    KFAADAVK
☑  15     438.66    875.31    875.43    -0.12  0     53     1    VSTEVDAR
☑  21     499.18    996.33    996.51    -0.17  0     34     1    TIVMGASFR + 1 Oxidation (M)
```

1:LUHU 3:AAB19866

This third example shows a weak match.

Very often, this is because the quality of the MS/MS spectrum is poor. If the signal to noise ratio is low, a match to the "correct" sequence might not exceed the absolute significance threshold. Even so, the match to the correct sequence could have a relatively high score, well differentiated from the quasi-normal distribution of random scores. In other words, the score is an outlier.

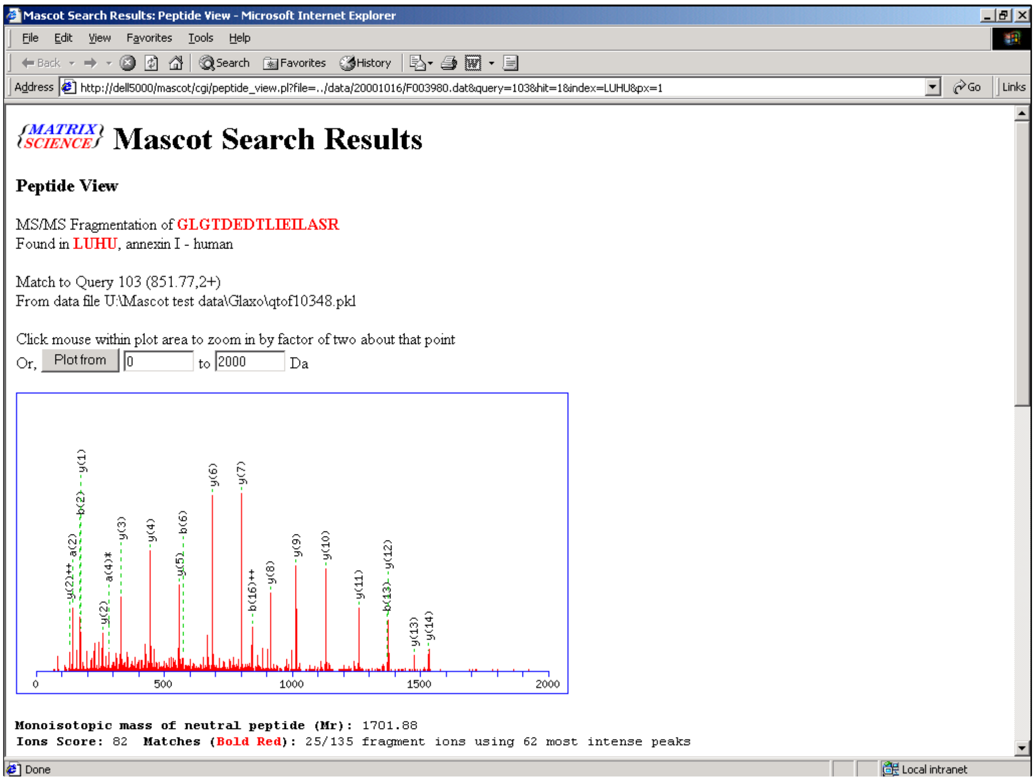
This would indicate that the match was not a random event and, on inspection, such matches are often found to be either the correct match or a match to a close homolog. For this reason, Mascot also attempts to characterise the distribution of random scores, and provide a second, lower threshold to highlight the presence of any outlier. The lower, relative threshold is reported as the "homology" threshold while the higher, absolute threshold is reported as the "identity" threshold.



Peptide Summary

- Score exceeds homology threshold:
  - Match is not random.
  - Spectrum may not fully define sequence
  - Sequence may be close but not exact
- Score exceeds identity threshold:
  - 5% chance that match is not exact

{MATRIX}
{SCIENCE}



Mascot Search Results: Peptide View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/peptide_view.pl?file=../data/20001016/F003980.dat&query=103&hit=1&index=LUHU&px=1
Go Links
MATRIX SCIENCE
Plot from
0
2000
y(2)++ a(2)
b(2)
y(1)
y(2)
a(4)*
y(3)
y(4)
y(5)
b(6)
y(6)
y(7)
b(16)++
y(8)
y(9)
y(10)
y(11)
b(13)
y(12)
y(13)
y(14)
0
500
1000
1500
2000
Done
Local intranet

Mascot Search Results

Peptide View

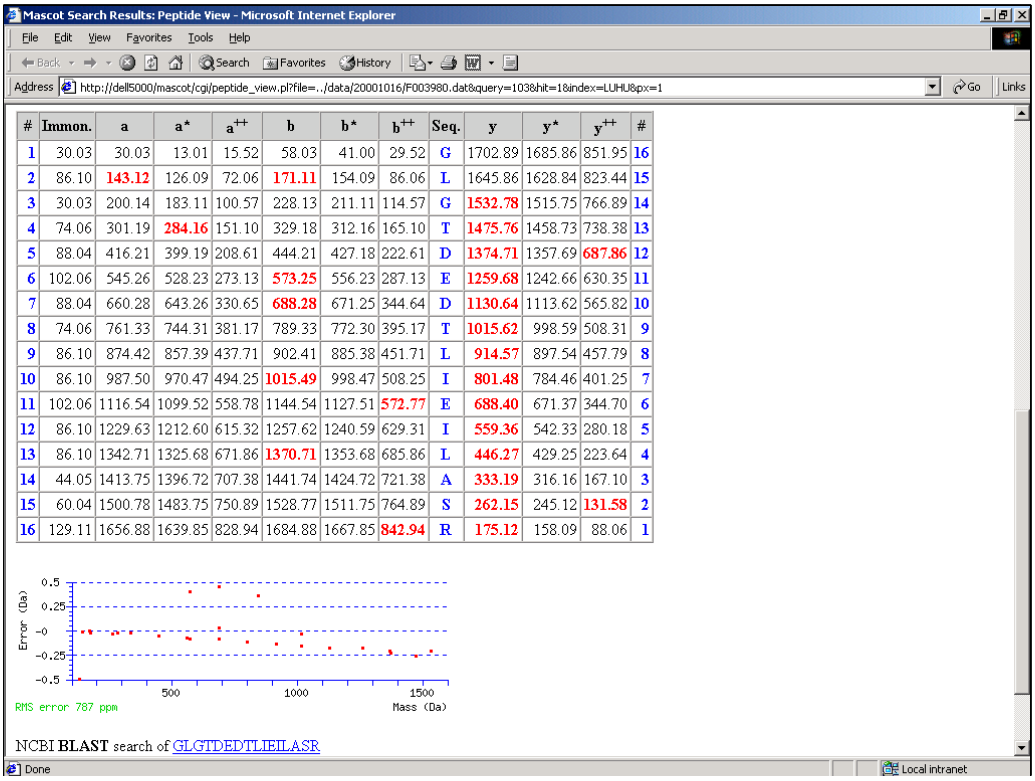
MS/MS Fragmentation of GLGTDEDTLIEILASR
Found in LUHU, annexin I - human

Match to Query 103 (851.77,2+)
From data file U:\Mascot test data\Glaxo\qtof10348.pkl

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 0 to 2000 Da

Monoisotopic mass of neutral peptide (Mr): 1701.88
Ions Score: 82 **Matches (Bold Red):** 25/135 fragment ions using 62 most intense peaks

Clicking on a query number link in the summary report loads the peptide view report. This illustrates the fragment ion matches highlighted on the MS/MS spectrum. Here we have a strong match with an almost complete series of y ions



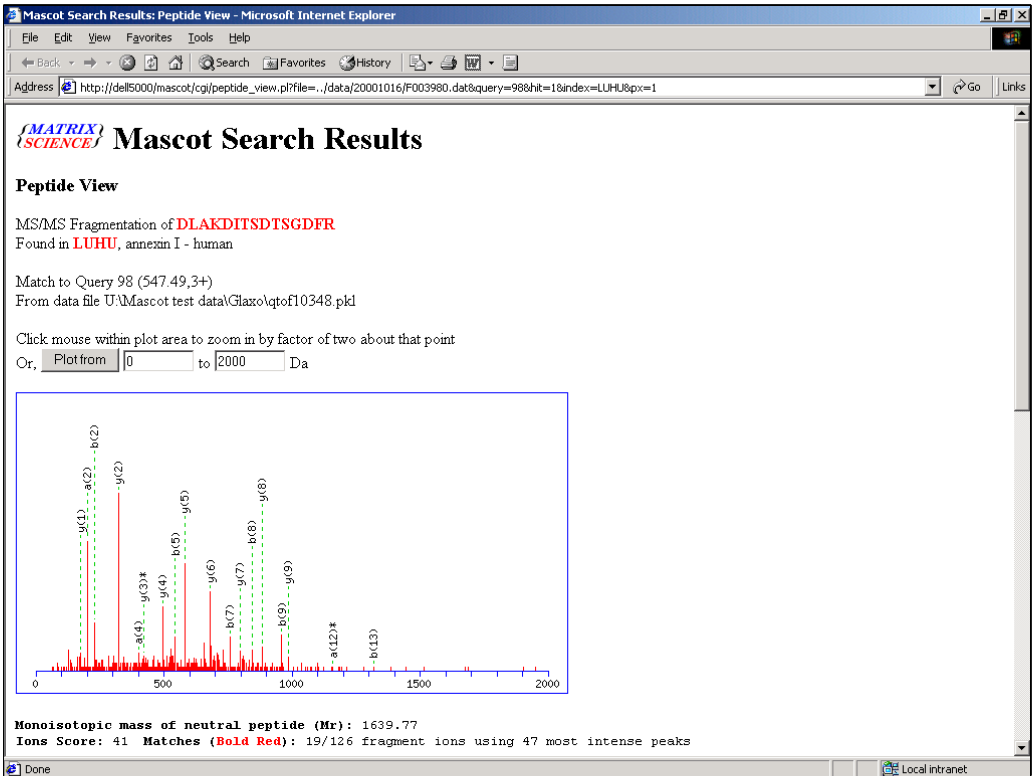
Mascot Search Results: Peptide View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/peptide_view.pl?file=../data/20001016/F003980.dat&query=103&hit=1&index=LUHU&px=1
Go Links

#	Immon.	a	a*	a++	b	b*	b++	Seq.	y	y*	y++	#
1	30.03	30.03	13.01	15.52	58.03	41.00	29.52	G	1702.89	1685.86	851.95	16
2	86.10	**143.12**	126.09	72.06	**171.11**	154.09	86.06	L	1645.86	1628.84	823.44	15
3	30.03	200.14	183.11	100.57	228.13	211.11	114.57	G	**1532.78**	1515.75	766.89	14
4	74.06	301.19	**284.16**	151.10	329.18	312.16	165.10	T	**1475.76**	1458.73	738.38	13
5	88.04	416.21	399.19	208.61	444.21	427.18	222.61	D	**1374.71**	1357.69	**687.86**	12
6	102.06	545.26	528.23	273.13	**573.25**	556.23	287.13	E	**1259.68**	1242.66	630.35	11
7	88.04	660.28	643.26	330.65	**688.28**	671.25	344.64	D	**1130.64**	1113.62	565.82	10
8	74.06	761.33	744.31	381.17	789.33	772.30	395.17	T	**1015.62**	998.59	508.31	9
9	86.10	874.42	857.39	437.71	902.41	885.38	451.71	L	**914.57**	897.54	457.79	8
10	86.10	987.50	970.47	494.25	**1015.49**	998.47	508.25	I	**801.48**	784.46	401.25	7
11	102.06	1116.54	1099.52	558.78	1144.54	1127.51	**572.77**	E	**688.40**	671.37	344.70	6
12	86.10	1229.63	1212.60	615.32	1257.62	1240.59	629.31	I	**559.36**	542.33	280.18	5
13	86.10	1342.71	1325.68	671.86	**1370.71**	1353.68	685.86	L	**446.27**	429.25	223.64	4
14	44.05	1413.75	1396.72	707.38	1441.74	1424.72	721.38	A	**333.19**	316.16	167.10	3
15	60.04	1500.78	1483.75	750.89	1528.77	1511.75	764.89	S	**262.15**	245.12	**131.58**	2
16	129.11	1656.88	1639.85	828.94	1684.88	1667.85	**842.94**	R	**175.12**	158.09	88.06	1

RMS error 787 ppm

NCBI **BLAST** search of GLGTDEDTLIEILASR

Further down, the matched peaks are highlighted in a table of calculated fragment ion masses. The peptide view is also where you can find the new graph of the error distribution for fragment ion masses.


Mascot Search Results: Peptide View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/peptide_view.pl?file=../data/20001016/F003980.dat&query=98&hit=1&index=LUHU&px=1
Go Links
MATRIX SCIENCE
y(1)
a(2)
b(2)
y(2)
a(4)
y(3)*
y(4)
b(5)
y(5)
y(6)
b(7)
y(7)
b(8)
y(8)
b(9)
y(9)
a(12)*
b(13)
0
500
1000
1500
2000
Done
Local intranet

Mascot Search Results

Peptide View

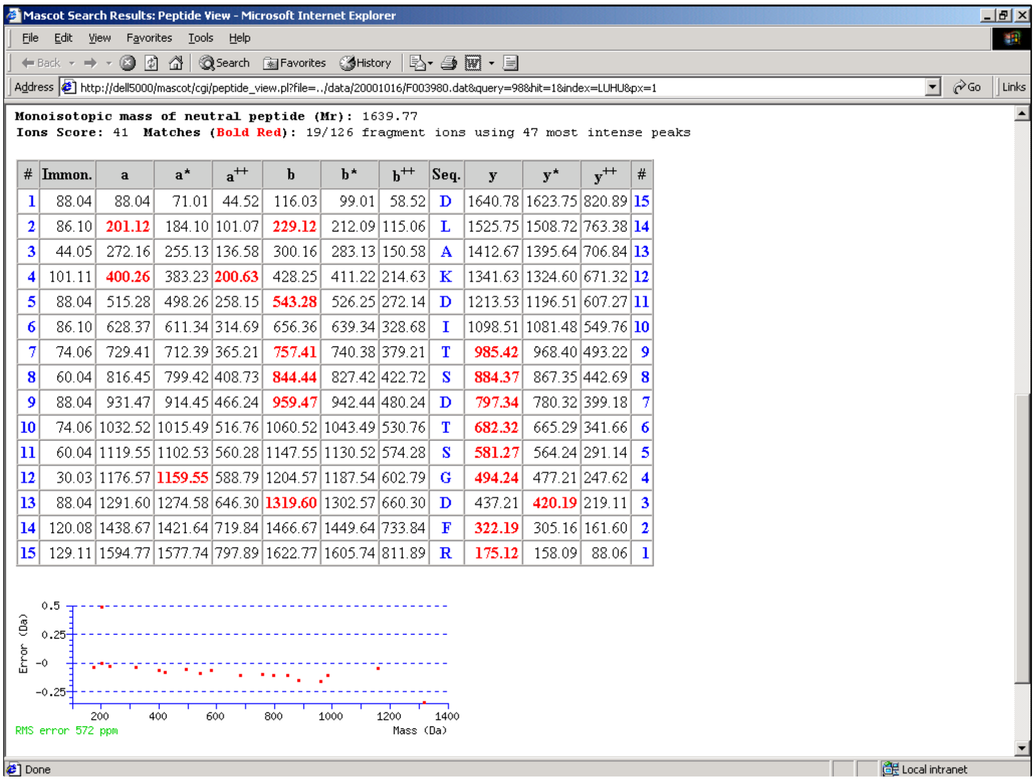
MS/MS Fragmentation of **DLAKDITSDTSGDFR**
Found in **LUHU**, annexin I - human

Match to Query 98 (547.49,3+)
From data file U:\Mascot test data\Glaxo\qtof10348.pkl

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 0 to 2000 Da

Monoisotopic mass of neutral peptide (Mr): 1639.77
Ions Score: 41 **Matches (Bold Red):** 19/126 fragment ions using 47 most intense peaks

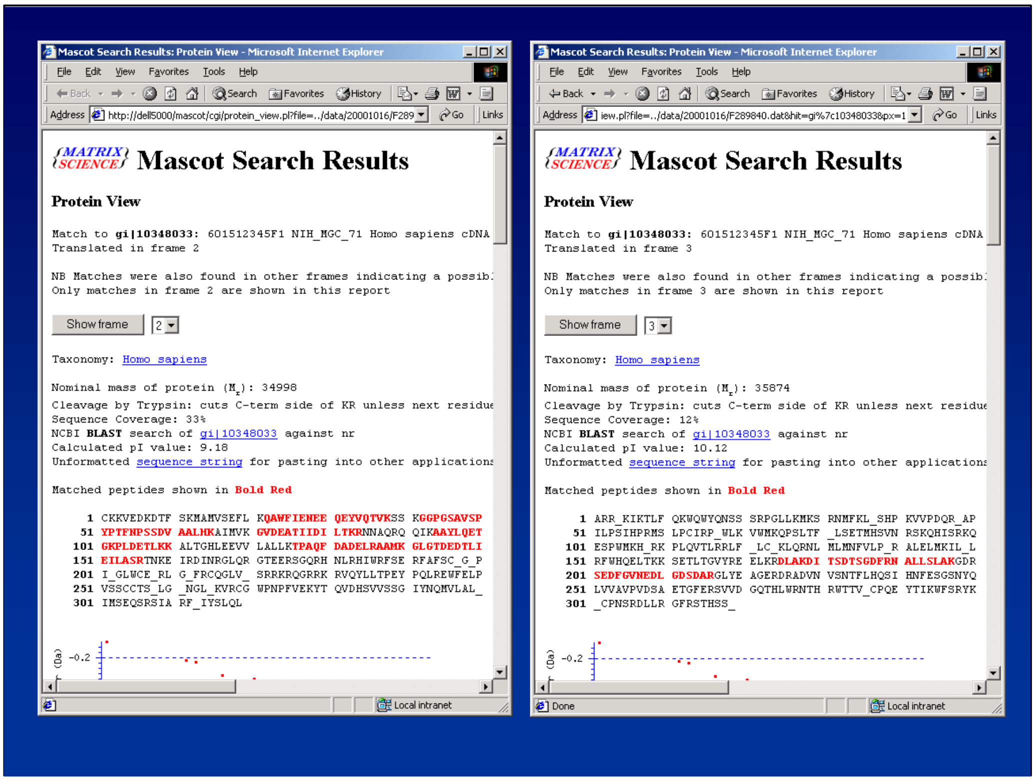
This is the peptide view for the weak match shown earlier. It can be seen that there is very little information above the precursor, and the signal to noise is not great


Mascot Search Results: Peptide View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/peptide_view.pl?file=../data/20001016/F003980.dat&query=98&hit=1&index=LUHU&px=1
Go Links

```
Monoisotopic mass of neutral peptide (Mr): 1639.77
Ions Score: 41  Matches (Bold Red): 19/126 fragment ions using 47 most intense peaks
```

#	Immon.	a	a*	a^{++}	b	b*	b^{++}	Seq.	y	y*	y^{++}	#
1	88.04	88.04	71.01	44.52	116.03	99.01	58.52	D	1640.78	1623.75	820.89	15
2	86.10	**201.12**	184.10	101.07	**229.12**	212.09	115.06	L	1525.75	1508.72	763.38	14
3	44.05	272.16	255.13	136.58	300.16	283.13	150.58	A	1412.67	1395.64	706.84	13
4	101.11	**400.26**	383.23	**200.63**	428.25	411.22	214.63	K	1341.63	1324.60	671.32	12
5	88.04	515.28	498.26	258.15	**543.28**	526.25	272.14	D	1213.53	1196.51	607.27	11
6	86.10	628.37	611.34	314.69	656.36	639.34	328.68	I	1098.51	1081.48	549.76	10
7	74.06	729.41	712.39	365.21	**757.41**	740.38	379.21	T	**985.42**	968.40	493.22	9
8	60.04	816.45	799.42	408.73	**844.44**	827.42	422.72	S	**884.37**	867.35	442.69	8
9	88.04	931.47	914.45	466.24	**959.47**	942.44	480.24	D	**797.34**	780.32	399.18	7
10	74.06	1032.52	1015.49	516.76	1060.52	1043.49	530.76	T	**682.32**	665.29	341.66	6
11	60.04	1119.55	1102.53	560.28	1147.55	1130.52	574.28	S	**581.27**	564.24	291.14	5
12	30.03	1176.57	**1159.55**	588.79	1204.57	1187.54	602.79	G	**494.24**	477.21	247.62	4
13	88.04	1291.60	1274.58	646.30	**1319.60**	1302.57	660.30	D	437.21	**420.19**	219.11	3
14	120.08	1438.67	1421.64	719.84	1466.67	1449.64	733.84	F	**322.19**	305.16	161.60	2
15	129.11	1594.77	1577.74	797.89	1622.77	1605.74	811.89	R	**175.12**	158.09	88.06	1

The N terminal end of the sequence is pretty much undefined. This is a good example of a spectrum which might get a match above the homology threshold, but lacks the information required to exceed the identity threshold



Mascot Search Results: Protein View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dell5000/mascot/cgi/protein_view.pl?file=../data/20001016/F289
Go Links
MATRIX SCIENCE

Mascot Search Results

Protein View

```
Match to gi|10348033: 601512345F1 NIH_MGC_71 Homo sapiens cDNA
Translated in frame 2

NB Matches were also found in other frames indicating a possibl
Only matches in frame 2 are shown in this report
```

```
Taxonomy: Homo sapiens

Nominal mass of protein (Mr): 34998
Cleavage by Trypsin: cuts C-term side of KR unless next residue
Sequence Coverage: 33%
NCBI BLAST search of gi|10348033 against nr
Calculated pI value: 9.18
Unformatted sequence string for pasting into other applications

Matched peptides shown in Bold Red

  1 CKKVEDKDTF SKMAMVSEFL KQAWFIENEE QEYVQTVKSS KGGPGSAVSP
 51 YPTFNPSSDV AALHKAIMVK GVDEATIIDI LTKRNNAQRQ QIKAAYLQET
101 GKPLDETLKK ALTGHLEEVV LALLKTPAQF DADELRAAMK GLGTDEDTLI
151 EILASRTNKE IRDINRGLQR GTEERSGQRH NLRHIWRFSE RFAFSC_G_P
201 I_GLWCE_RL G_FRCQGLV_ SRRKRQGRRK RVQYLLTPEY PQLREWFELP
251 VSSCCTS_LG _NGL_KVRCG WPNPFVEKYT QVDHSVVSSG IYNQMVLAL_
301 IMSEQSRSIA RF_IYSLQL
```

Mascot Search Results: Protein View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address iew.pl?file=../data/20001016/F289840.dat&hit=gi%7c10348033&px=1
Go Links
MATRIX SCIENCE

Mascot Search Results

Protein View

```
Match to gi|10348033: 601512345F1 NIH_MGC_71 Homo sapiens cDNA
Translated in frame 3

NB Matches were also found in other frames indicating a possibl
Only matches in frame 3 are shown in this report
```

```
Taxonomy: Homo sapiens

Nominal mass of protein (Mr): 35874
Cleavage by Trypsin: cuts C-term side of KR unless next residue
Sequence Coverage: 12%
NCBI BLAST search of gi|10348033 against nr
Calculated pI value: 10.12
Unformatted sequence string for pasting into other applications

Matched peptides shown in Bold Red

  1 ARR_KIKTLF QKWQWYQNSS SRPGLLKMKS RNMFKL_SHP KVVPDQR_AP
 51 ILPSIHPRMS LPCIRP_WLK VWMKQPSLTF _LSETMHSVN RSKQHISRKQ
101 ESPWMKH_RK PLQVTLRRLF _LC_KLQRNL MLMNFVLP_R ALELMKIL_L
151 RFWHQELTKK SETLTGVYRE ELKRDLAKDI TSDTSGDFRN ALLSLAKGDR
201 SEDFGVNEDL GDSDARGLYE AGERDRADVN VSNTFLHQSI HNFESGSNYQ
251 LVVAVPVDSA ETGFERSVVD GQTHLWRNTH RWTTV_CPQE YTIKWFSRYK
301 _CPNSRDLLR GFRSTHSS_
```

Finally, a major difference between reports from searching a protein database and those from searching a nucleic acid database is the possibility of frame shifts within the entry.

Thus, in the protein view report, there is a drop down list for the different translation frames. For this particular entry, most of the matches have been found in reading frame 2. But, as so often happens, there is a frame shift in this entry, and there are additional matches in frame 3.