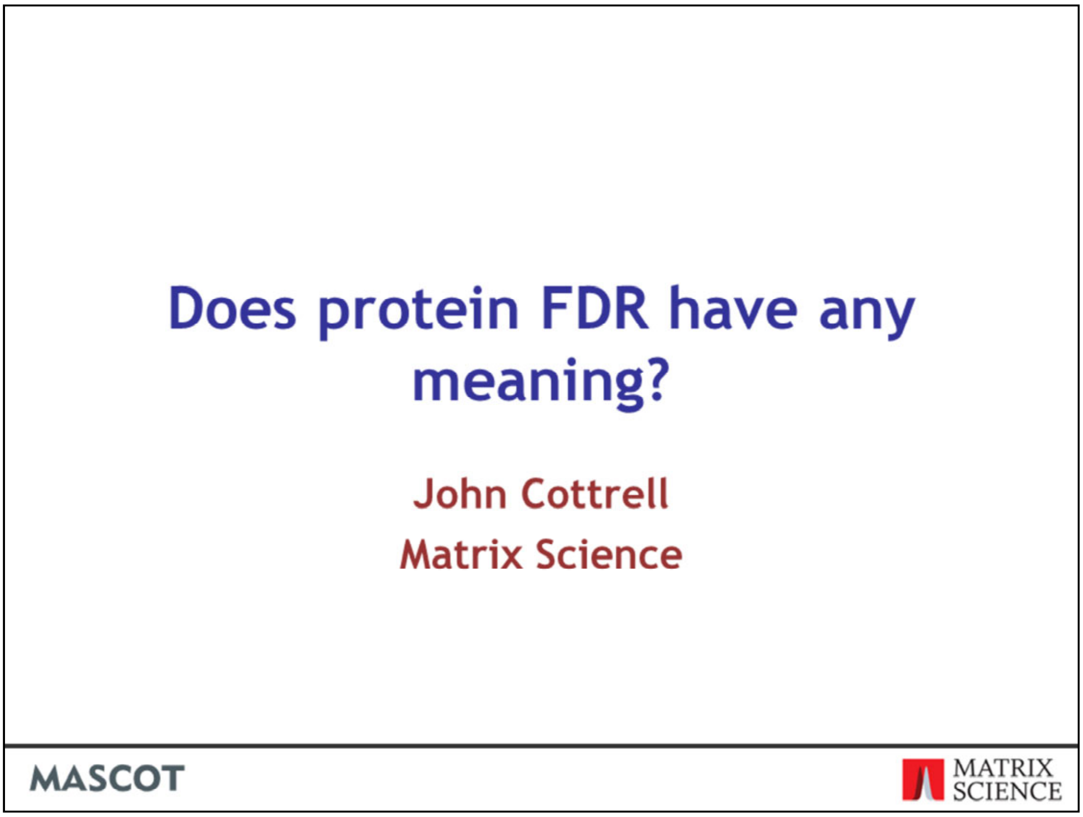


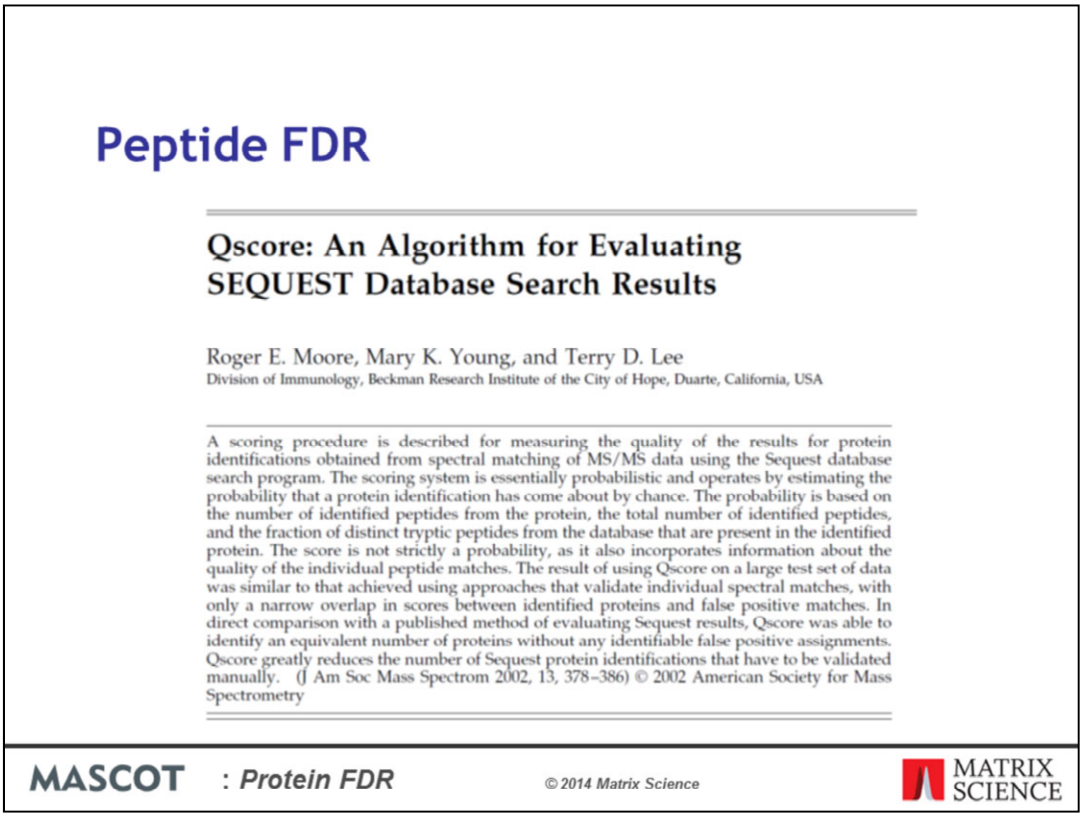
Does protein FDR have any meaning?

John Cottrell

Matrix Science

MASCOT

MATRIX SCIENCE



Peptide FDR

Qscore: An Algorithm for Evaluating SEQUEST Database Search Results

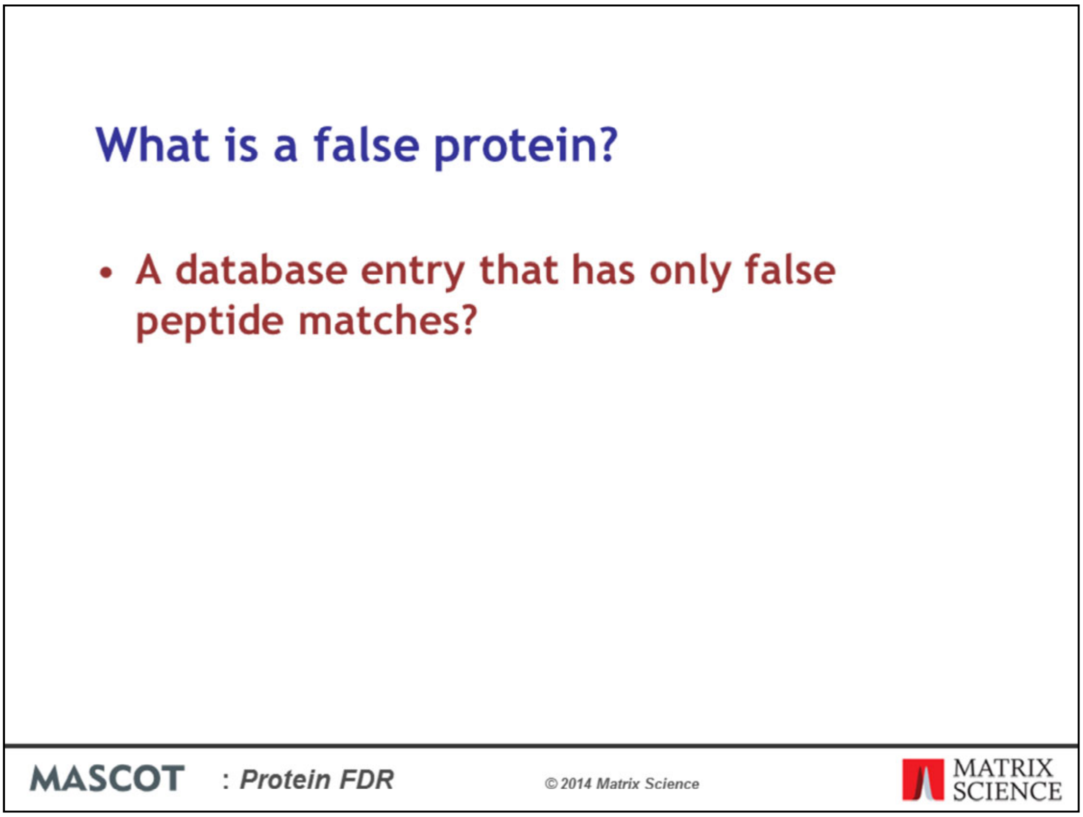
Roger E. Moore, Mary K. Young, and Terry D. Lee
Division of Immunology, Beckman Research Institute of the City of Hope, Duarte, California, USA

A scoring procedure is described for measuring the quality of the results for protein identifications obtained from spectral matching of MS/MS data using the Sequest database search program. The scoring system is essentially probabilistic and operates by estimating the probability that a protein identification has come about by chance. The probability is based on the number of identified peptides from the protein, the total number of identified peptides, and the fraction of distinct tryptic peptides from the database that are present in the identified protein. The score is not strictly a probability, as it also incorporates information about the quality of the individual peptide matches. The result of using Qscore on a large test set of data was similar to that achieved using approaches that validate individual spectral matches, with only a narrow overlap in scores between identified proteins and false positive matches. In direct comparison with a published method of evaluating Sequest results, Qscore was able to identify an equivalent number of proteins without any identifiable false positive assignments. Qscore greatly reduces the number of Sequest protein identifications that have to be validated manually. (J Am Soc Mass Spectrom 2002, 13, 378–386) © 2002 American Society for Mass Spectrometry

MATRIX SCIENCE

Its easy to grasp the concept of using a target/decoy search to estimate peptide false discovery rate. You search against a decoy database, in which there are no true matches available, so the number of observed matches provides a good estimate of the number of false matches from the target database. I think this is the first time this approach was applied to database search using MS/MS data, by Terry Lee's group in 2002, but the method only became widespread in proteomics after the publications from Steve Gygi's group.

The most popular way to create a decoy database is to reverse the protein sequences in the target database. When reversed entries are digested, we get a population of peptides that have most of the characteristics of target peptides. Certainly, in terms of the qualities scored by search engines, such peptides are perfect decoys.

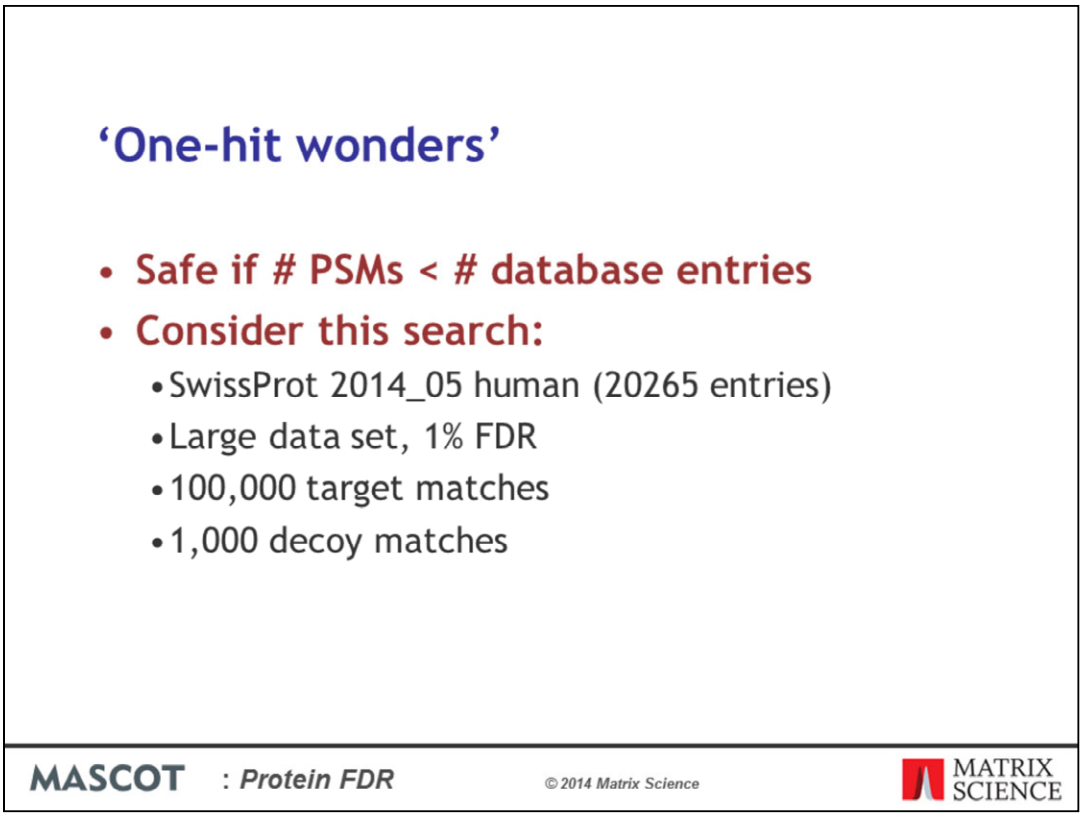


What is a false protein?

- A database entry that has only false peptide matches?

MASCOT : *Protein FDR* *© 2014 Matrix Science* MATRIX SCIENCE

Protein false discovery rate is not so easily estimated. First of all, what exactly do we mean by a false protein? A possible definition might be a database entry that has only false peptide matches. These are clearly junk, so best to filter them out by requiring every protein to have significant matches to two or more distinct peptide sequences. This eliminates the 'one-hit wonder' proteins, where a false peptide match has been assigned to a protein for which there is no other evidence.

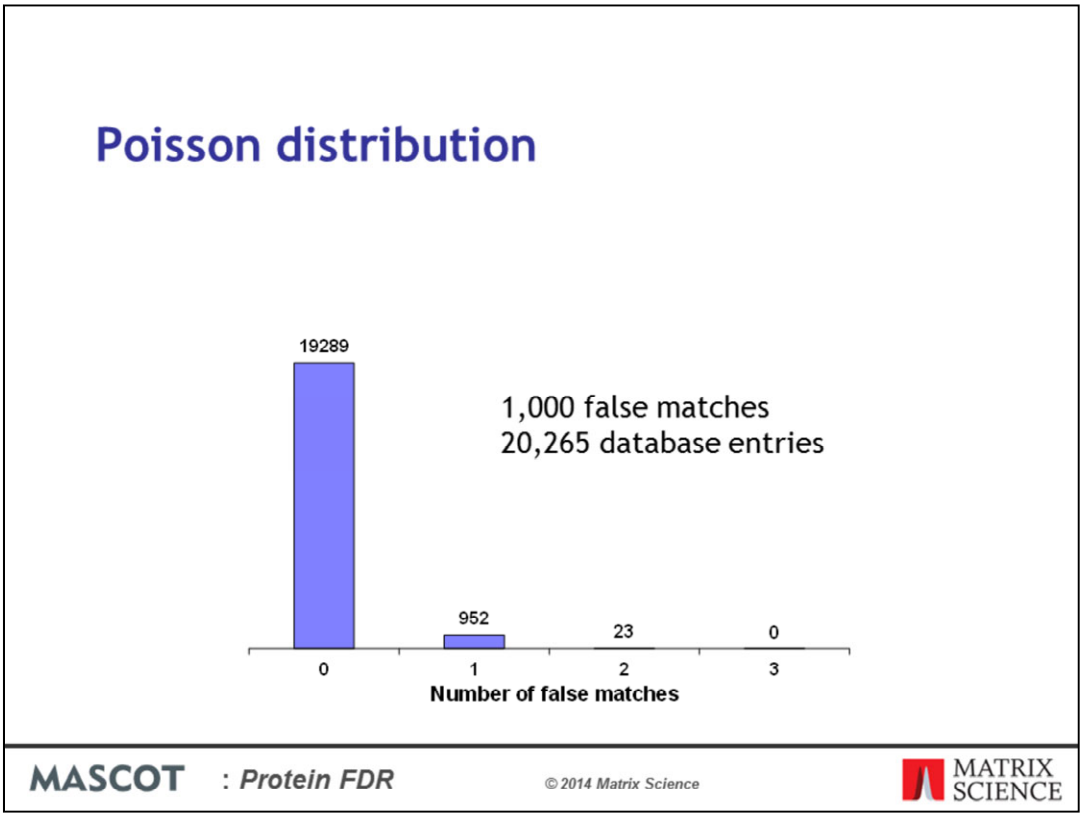


‘One-hit wonders’

- Safe if # PSMs < # database entries
- Consider this search:
  - SwissProt 2014_05 human (20265 entries)
  - Large data set, 1% FDR
  - 100,000 target matches
  - 1,000 decoy matches

MASCOT : *Protein FDR* © 2014 Matrix Science MATRIX SCIENCE

If you have a large number of matches in a search of a small database, filtering out one-hit wonders may not be enough. We can calculate the distribution of false peptide matches using Poisson statistics. SwissProt 2014_05 has 20,265 human entries. If we searched a large data set and got 100,000 matches at 1% peptide FDR, this would correspond to 1000 false peptide matches.



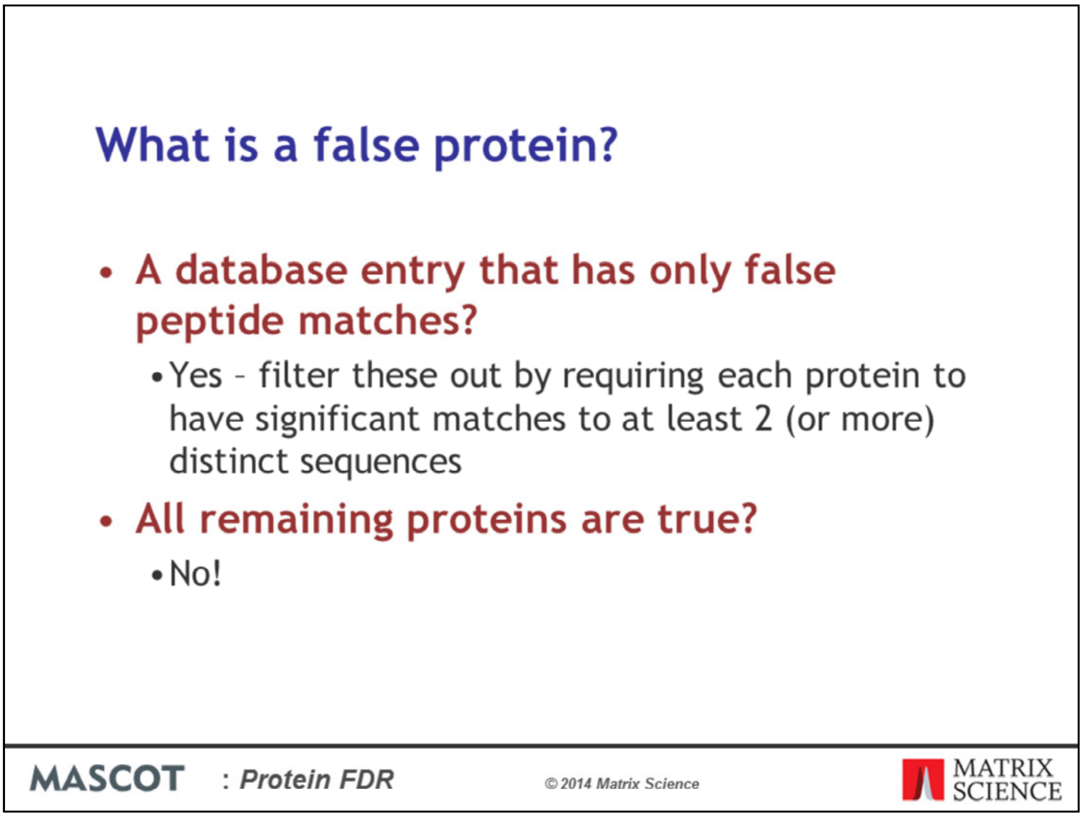
Poisson distribution

1,000 false matches
20,265 database entries

MASCOT : *Protein FDR*

MATRIX SCIENCE

The Poisson distribution predicts that, on average, 952 database entries will get one false match, 23 entries will get two, and less than one will get three. If an entry has two false matches to different sequences, it will pass a ‘one hit wonder’ filter, so we could have as many as 23 false proteins in our report. If this is too many, we raise the bar, require significant matches to three or more distinct peptide sequences, and the anticipated number of false proteins drops to less than one.

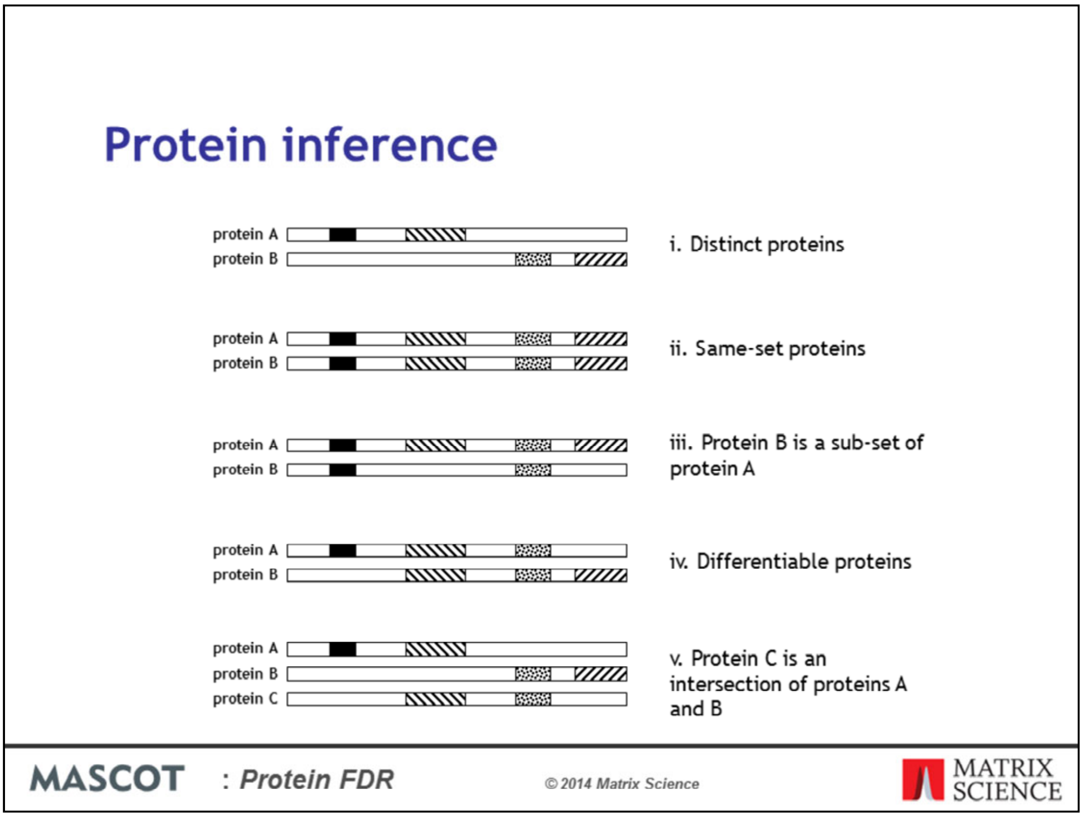


What is a false protein?

- A database entry that has only false peptide matches?
  - Yes – filter these out by requiring each protein to have significant matches to at least 2 (or more) distinct sequences
- All remaining proteins are true?
  - No!

MASCOT : *Protein FDR* © 2014 Matrix Science MATRIX SCIENCE

Problem solved? Not if our goal is to present an accurate list of the proteins that were present in the sample.



Protein inference

i. Distinct proteins

ii. Same-set proteins

iii. Protein B is a sub-set of protein A

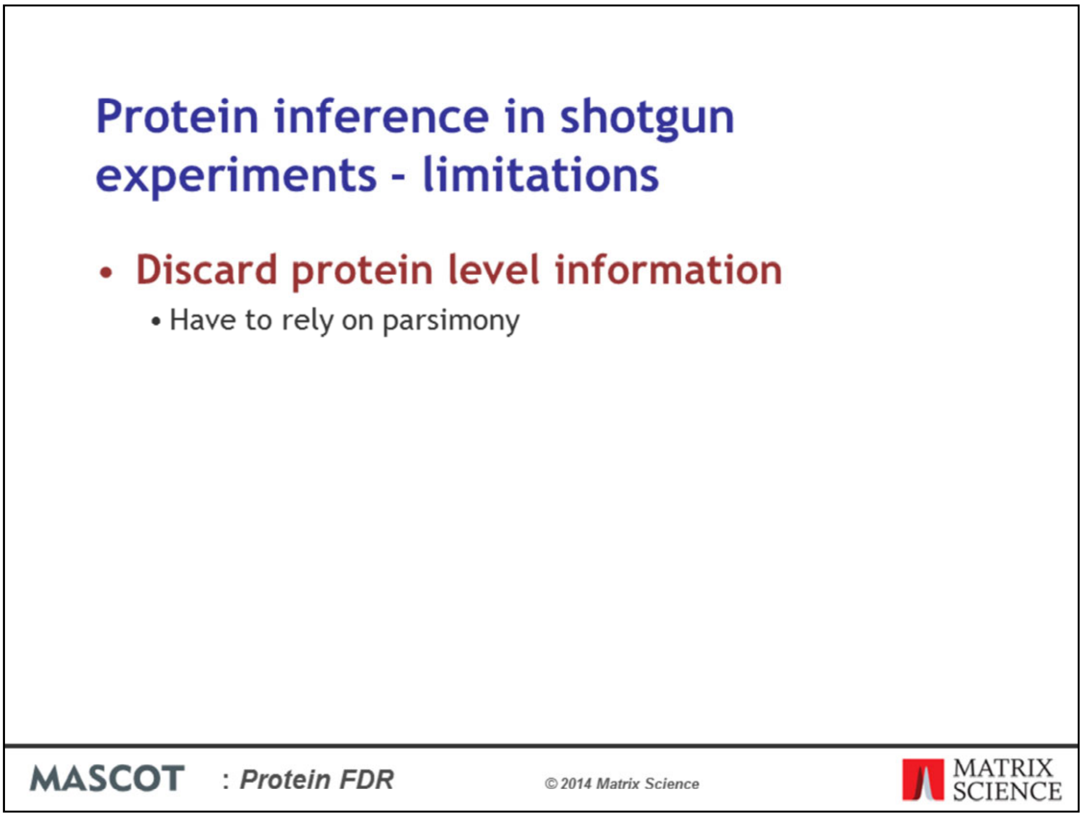
iv. Differentiable proteins

v. Protein C is an intersection of proteins A and B

MASCOT : *Protein FDR* © 2014 Matrix Science MATRIX SCIENCE

I'm sure everyone is familiar with the concept of same-set and sub-set proteins. If we search a comprehensive database, like NCBInr, same-set proteins will be common. Reporting just one of means that the count of proteins is probably correct, but we have no idea which one of the same-set proteins is actually present in the sample because protein inference only considers the peptide matches. It ignores the unmatched parts of the sequence and there is no penalty for the matches we fail to observe. So, even though the same-set proteins might be very different in any biological sense, we can only report that we have at least one out of the set.

However, its differentiable proteins that pose the real problem. Do we report one of them, or all of them?



Protein inference in shotgun experiments - limitations

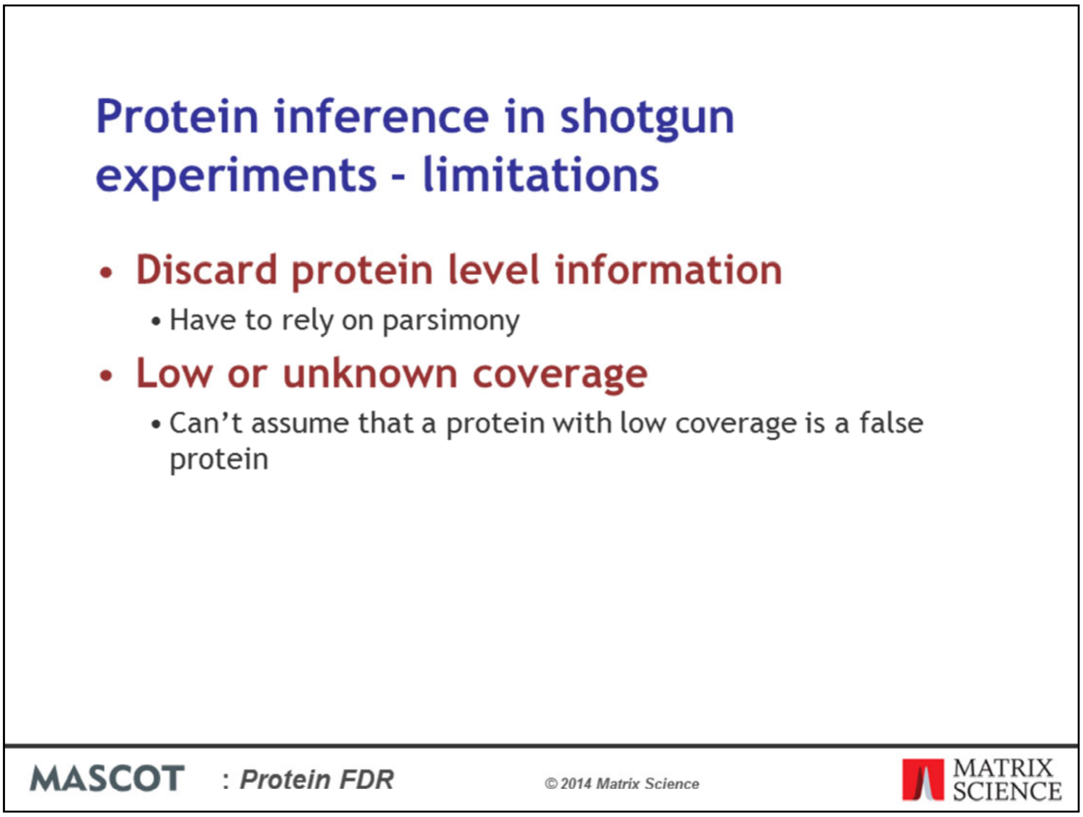
- Discard protein level information
  - Have to rely on parsimony

MASCOT : Protein FDR © 2014 Matrix Science MATRIX SCIENCE

We have to recognise that protein inference in shotgun proteomics is subject to some serious and fundamental limitations.

When we analyse a pure protein from a 2D gel spot, protein inference is much easier. If you can identify one peptide, you should be able to identify several, and with high coverage, one database entry becomes the clear winner. Other entries may contain some of the same peptides, but unless they also have similar protein mass and pI, they can be ruled out.

In shotgun proteomics, the protein level information is discarded in the interests of speed and scale, and protein inference comes to rely on parsimony, alone.



Protein inference in shotgun experiments - limitations

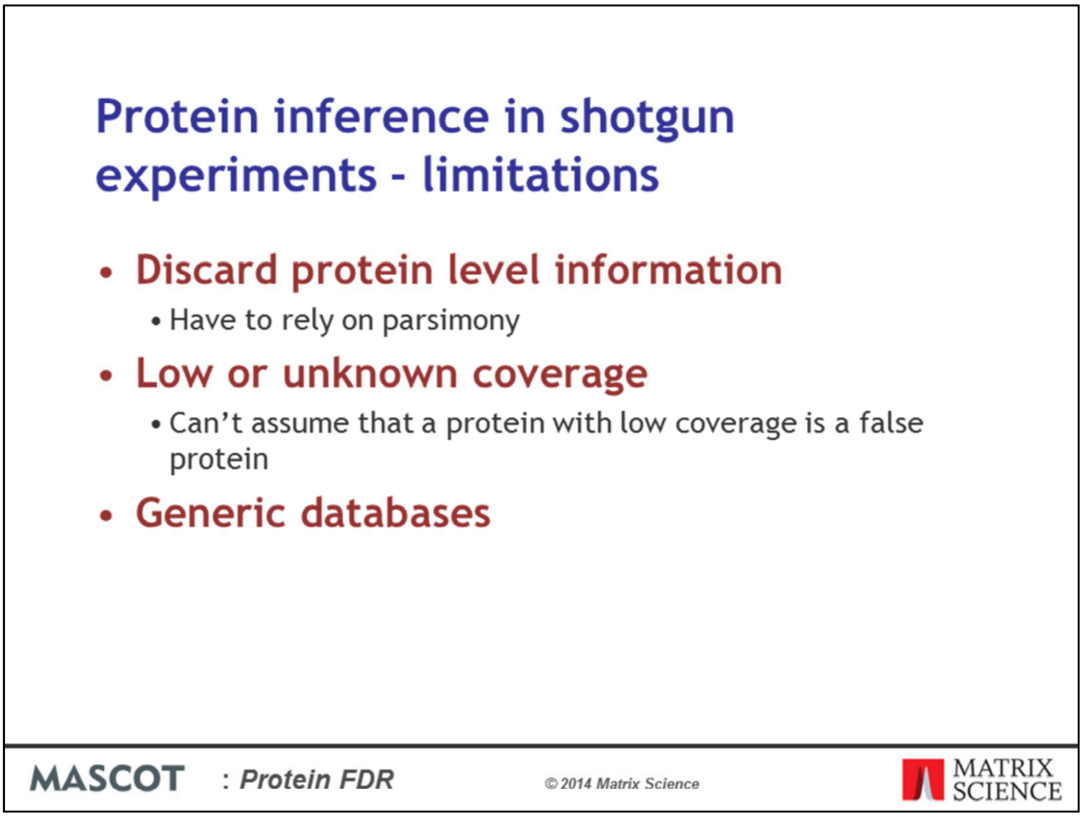
- Discard protein level information
  - Have to rely on parsimony
- Low or unknown coverage
  - Can’t assume that a protein with low coverage is a false protein

MASCOT : Protein FDR

© 2014 Matrix Science

MATRIX SCIENCE

In most shotgun experiments, the peptides are under-sampled. That is, MS/MS scans are acquired for the stronger peptide signals but the weaker ones get overlooked, and the number of different peptides observed for a particular protein depends on its abundance as well as its length. On the plus side, this is the basis of spectral counting as a method of quantitation. On the minus side, it means we can’t assume that a protein with low coverage is a false protein. It could be a true protein that happens to be present at a low level. Not that we actually know what the coverage is, because we don’t have masses for the proteins. When we talk about coverage, this means coverage for the database entry, not for the protein. Any attempt to use coverage in protein inference simply favours the shortest database entry that contains the observed matches.

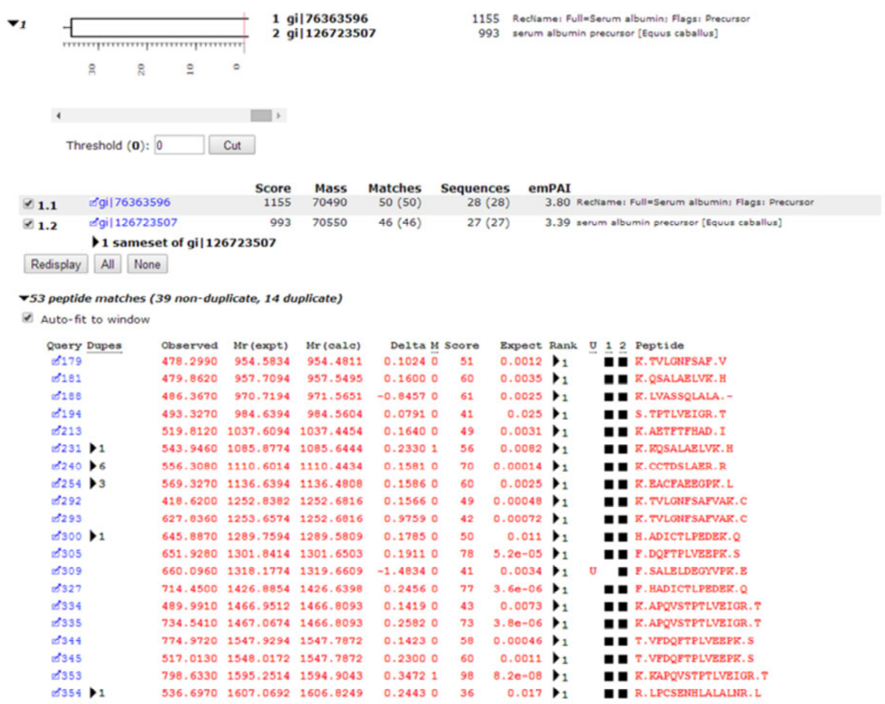


Protein inference in shotgun experiments - limitations

- Discard protein level information
  - Have to rely on parsimony
- Low or unknown coverage
  - Can't assume that a protein with low coverage is a false protein
- Generic databases

MASCOT : *Protein FDR* © 2014 Matrix Science MATRIX SCIENCE

Some day, it may become routine to create a custom database for the individual proteome under analysis using a technique such as RNA-Seq. Right now, most searches are against the public protein databases, and these will not contain perfectly correct sequences for many of the proteins in the sample. In the absence of the correct sequence, matches are assigned to a set of homologous entries.



1 gi|76363596 1155 RecName: Full=Serum albumin; Flags: Precursor
2 gi|126723507 993 serum albumin precursor [Equus caballus]

		Score	Mass	Matches	Sequences	emPAI	
☑ 1.1	gi|76363596	1155	70490	50 (50)	28 (28)	3.80	RecName: Full=Serum albumin; Flags: Precursor
☑ 1.2	gi|126723507	993	70550	46 (46)	27 (27)	3.39	serum albumin precursor [Equus caballus]

▶1 sameset of gi|126723507

▼53 peptide matches (39 non-duplicate, 14 duplicate)

☑ Auto-fit to window

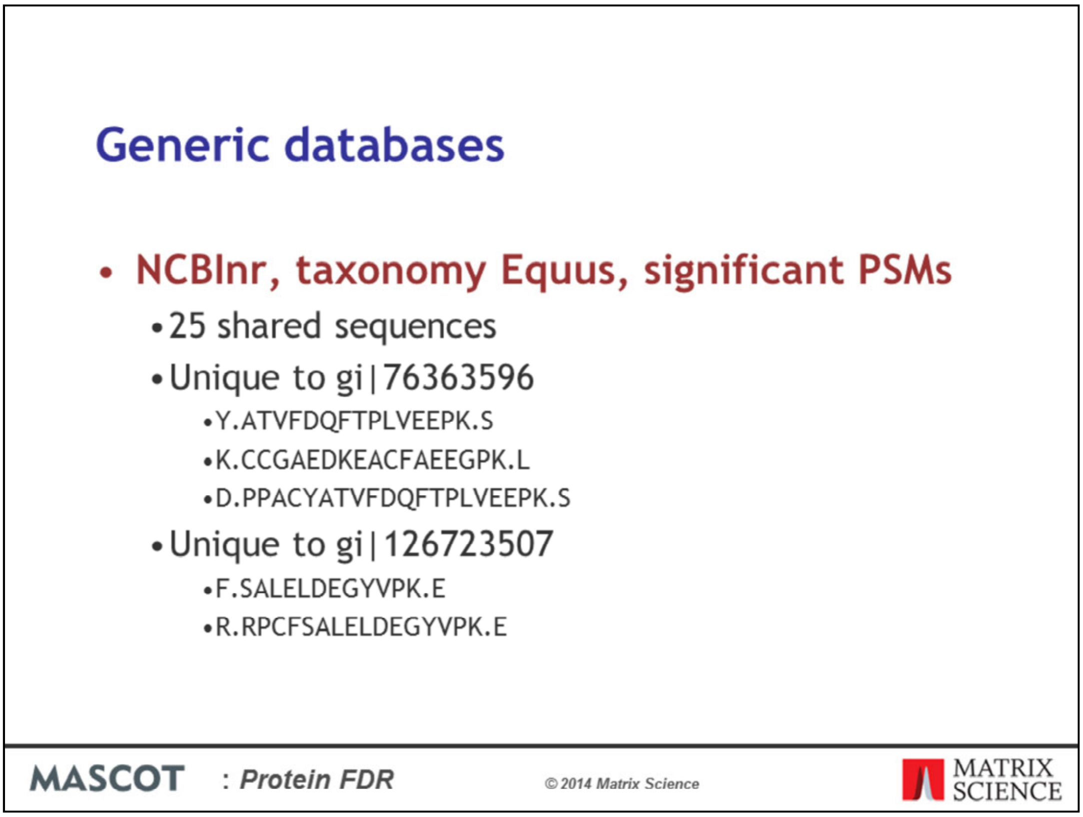
Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	Peptide
179		478.2990	954.5834	954.4811	0.1024	0	51	0.0012	▶1		■	■	K.TVLGNFSAF.V
181		479.8620	957.7094	957.5495	0.1600	0	60	0.0035	▶1		■	■	K.QSALAELVK.H
188		486.3670	970.7194	971.5651	-0.8457	0	61	0.0025	▶1		■	■	K.LVASSQLALA.-
194		493.3270	984.6394	984.5604	0.0791	0	41	0.025	▶1		■	■	S.TPTLVEIGR.T
213		519.8120	1037.6094	1037.4454	0.1640	0	49	0.0031	▶1		■	■	K.AETFTFHAD.I
231	▶1	543.9460	1085.8774	1085.6444	0.2330	1	56	0.0082	▶1		■	■	K.KQSALAELVK.H
240	▶6	556.3080	1110.6014	1110.4434	0.1581	0	70	0.00014	▶1		■	■	K.CCTDSLAER.R
254	▶3	569.3270	1136.6394	1136.4808	0.1586	0	60	0.0025	▶1		■	■	K.EACFAEEGPK.L
292		418.6200	1252.8382	1252.6816	0.1566	0	49	0.00048	▶1		■	■	K.TVLGNFSAFVAK.C
293		627.8360	1253.6574	1252.6816	0.9759	0	42	0.00072	▶1		■	■	K.TVLGNFSAFVAK.C
300	▶1	645.8870	1289.7594	1289.5809	0.1785	0	50	0.011	▶1		■	■	H.ADICTLPEDEK.Q
305		651.9280	1301.8414	1301.6503	0.1911	0	78	5.2e-05	▶1		■	■	F.DQFTPLVEEPK.S
309		660.0960	1318.1774	1319.6609	-1.4834	0	41	0.0034	▶1	U		■	F.SALELDEGYVPK.E
327		714.4500	1426.8854	1426.6398	0.2456	0	77	3.6e-06	▶1		■	■	F.HADICTLPEDEK.Q
334		489.9910	1466.9512	1466.8093	0.1419	0	43	0.0073	▶1		■	■	K.APQVSTPTLVEIGR.T
335		734.5410	1467.0674	1466.8093	0.2582	0	73	3.8e-06	▶1		■	■	K.APQVSTPTLVEIGR.T
344		774.9720	1547.9294	1547.7872	0.1423	0	58	0.00046	▶1		■	■	T.VFDQFTPLVEEPK.S
345		517.0130	1548.0172	1547.7872	0.2300	0	60	0.0011	▶1		■	■	T.VFDQFTPLVEEPK.S
353		798.6330	1595.2514	1594.9043	0.3472	1	98	8.2e-08	▶1		■	■	K.KAPQVSTPTLVEIGR.T
354	▶1	536.6970	1607.0692	1606.8249	0.2443	0	36	0.017	▶1		■	■	R.LPCSENHLALALNR.L

MASCOT : *Protein FDR*

© 2014 Matrix Science

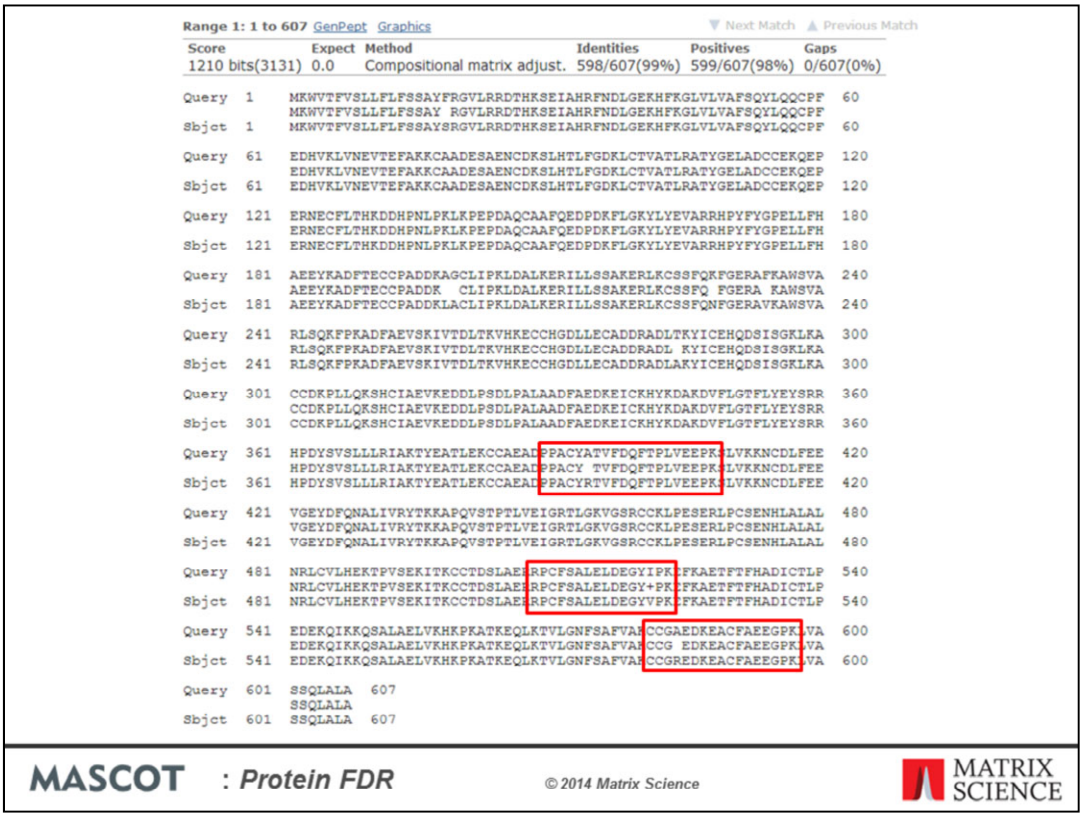
MATRIX SCIENCE

This search result illustrates. Much too small to estimate the peptide FDR with any accuracy, but the significance threshold has been set to a level where the count of decoy peptide matches is zero. In hit 1, most of the peptide matches are shared between two sequences, but each protein also has many significant matches that are not shared



Generic databases

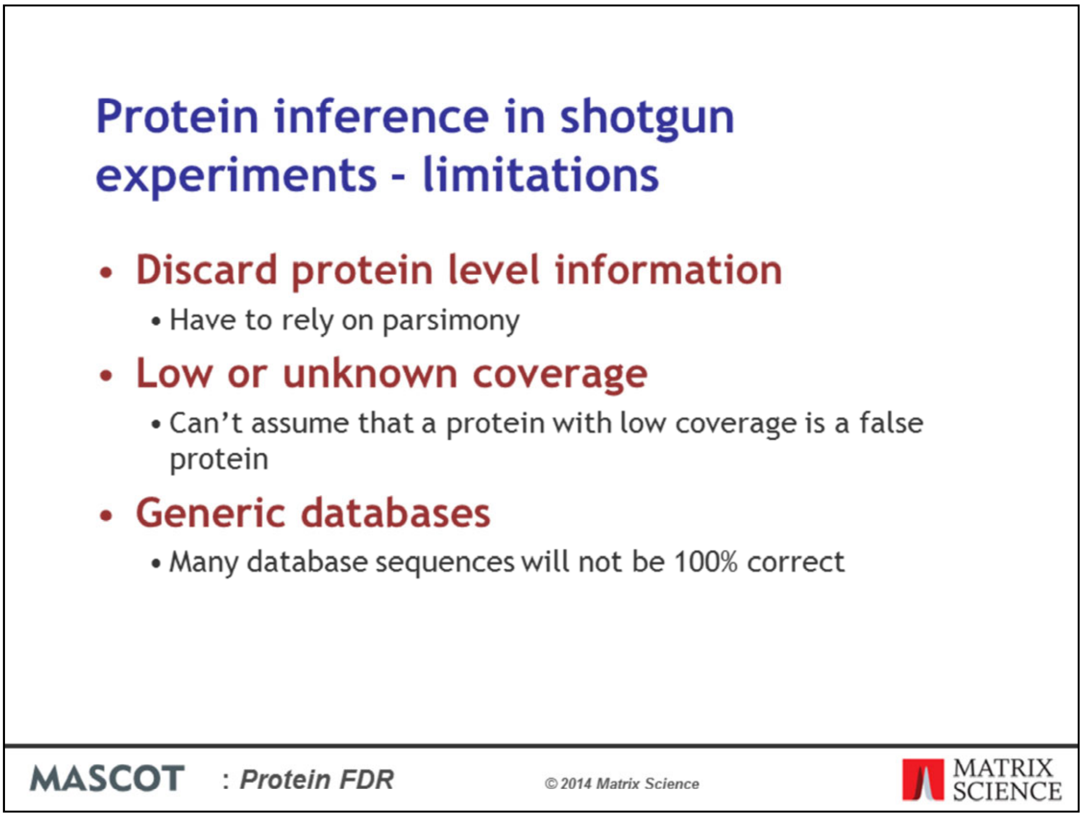
- NCBInr, taxonomy Equus, significant PSMs
  - 25 shared sequences
  - Unique to gi|76363596
    - Y.ATVFDQFTPLVEEPK.S
    - K.CCGAEDKEACFAEEGPK.L
    - D.PPACYATVFDQFTPLVEEPK.S
  - Unique to gi|126723507
    - F.SALELDEGYVPK.E
    - R.RPCFSALELDEGYVPK.E



```
Range 1: 1 to 607 GenPept  Graphics                                  ▼ Next Match  ▲ Previous Match
 Score            Expect  Method                        Identities     Positives      Gaps
 1210 bits(3131)  0.0     Compositional matrix adjust.  598/607(99%)   599/607(98%)   0/607(0%)

Query  1    MKWVTFVSLLFLFSSAYFRGVLRRDTHKSEIAHRFNDLGEKHFKGLVLVAFSQYLQQCPF  60
            MKWVTFVSLLFLFSSAY RGVLRRDTHKSEIAHRFNDLGEKHFKGLVLVAFSQYLQQCPF
Sbjct  1    MKWVTFVSLLFLFSSAYSRGVLRRDTHKSEIAHRFNDLGEKHFKGLVLVAFSQYLQQCPF  60

Query  61   EDHVKLVNEVTEFAKKCAADESAENCDKSLHTLFGDKLCTVATLRATYGELADCCEKQEP  120
            EDHVKLVNEVTEFAKKCAADESAENCDKSLHTLFGDKLCTVATLRATYGELADCCEKQEP
Sbjct  61   EDHVKLVNEVTEFAKKCAADESAENCDKSLHTLFGDKLCTVATLRATYGELADCCEKQEP  120

Query  121  ERNECFLTHKDDHPNLPKLKPEPDAQCAAFQEDPDKFLGKYLYEVARRHPYFYGPELLFH  180
            ERNECFLTHKDDHPNLPKLKPEPDAQCAAFQEDPDKFLGKYLYEVARRHPYFYGPELLFH
Sbjct  121  ERNECFLTHKDDHPNLPKLKPEPDAQCAAFQEDPDKFLGKYLYEVARRHPYFYGPELLFH  180

Query  181  AEEYKADFTECCPADDKAGCLIPKLDALKERILLSSAKERLKCSSFQKFGERAFKAWSVA  240
            AEEYKADFTECCPADDK  CLIPKLDALKERILLSSAKERLKCSSFQ FGERA KAWSVA
Sbjct  181  AEEYKADFTECCPADDKLACLIPKLDALKERILLSSAKERLKCSSFQNFGERAVKAWSVA  240

Query  241  RLSQKFPKADFAEVSKIVTDLTKVHKECCHGDLLECADDRADLTKYICEHQDSISGKLKA  300
            RLSQKFPKADFAEVSKIVTDLTKVHKECCHGDLLECADDRADL KYICEHQDSISGKLKA
Sbjct  241  RLSQKFPKADFAEVSKIVTDLTKVHKECCHGDLLECADDRADLAKYICEHQDSISGKLKA  300

Query  301  CCDKPLLQKSHCIAEVKEDDLPSDLPALAADFAEDKEICKHYKDAKDVFLGTFLYEYSRR  360
            CCDKPLLQKSHCIAEVKEDDLPSDLPALAADFAEDKEICKHYKDAKDVFLGTFLYEYSRR
Sbjct  301  CCDKPLLQKSHCIAEVKEDDLPSDLPALAADFAEDKEICKHYKDAKDVFLGTFLYEYSRR  360

Query  361  HPDYSVSLLLRIAKTYEATLEKCCAEADPPACYATVFDQFTPLVEEPKSLVKKNCDLFEE  420
            HPDYSVSLLLRIAKTYEATLEKCCAEADPPACY TVFDQFTPLVEEPKSLVKKNCDLFEE
Sbjct  361  HPDYSVSLLLRIAKTYEATLEKCCAEADPPACYRTVFDQFTPLVEEPKSLVKKNCDLFEE  420

Query  421  VGEYDFQNALIVRYTKKAPQVSTPTLVEIGRTLGKVGSRCCKLPESERLPCSENHLALAL  480
            VGEYDFQNALIVRYTKKAPQVSTPTLVEIGRTLGKVGSRCCKLPESERLPCSENHLALAL
Sbjct  421  VGEYDFQNALIVRYTKKAPQVSTPTLVEIGRTLGKVGSRCCKLPESERLPCSENHLALAL  480

Query  481  NRLCVLHEKTPVSEKITKCCTDSLAERRPCFSALELDEGYIPKEFKAETFTFHADICTLP  540
            NRLCVLHEKTPVSEKITKCCTDSLAERRPCFSALELDEGY+PKEFKAETFTFHADICTLP
Sbjct  481  NRLCVLHEKTPVSEKITKCCTDSLAERRPCFSALELDEGYVPKEFKAETFTFHADICTLP  540

Query  541  EDEKQIKKQSALAELVKHKPKATKEQLKTVLGNFSAFVAKCCGAEDKEACFAEEGPKLVA  600
            EDEKQIKKQSALAELVKHKPKATKEQLKTVLGNFSAFVAKCCG EDKEACFAEEGPKLVA
Sbjct  541  EDEKQIKKQSALAELVKHKPKATKEQLKTVLGNFSAFVAKCCGREDKEACFAEEGPKLVA  600

Query  601  SSQLALA  607
            SSQLALA
Sbjct  601  SSQLALA  607
```

MASCOT : Protein FDR © 2014 Matrix Science MATRIX SCIENCE

BLAST alignment between the two protein sequences shows them to be 99% identical. The alignments for the 'unique' peptides are highlighted. When you look at it like this, it becomes clear that we don't have two distinct proteins, just a variant that is not 100% identical to either of the two database entries. In other cases, these differences might have corresponded to splice variants, and there are indeed two different, but homologous proteins. In other words, deciding whether a pair of differentiable proteins should be reported as one or two isn't a matter of numbers or statistics. It requires an understanding of the relationship between the database sequences and whether the true protein is a third sequence, not present in the database.



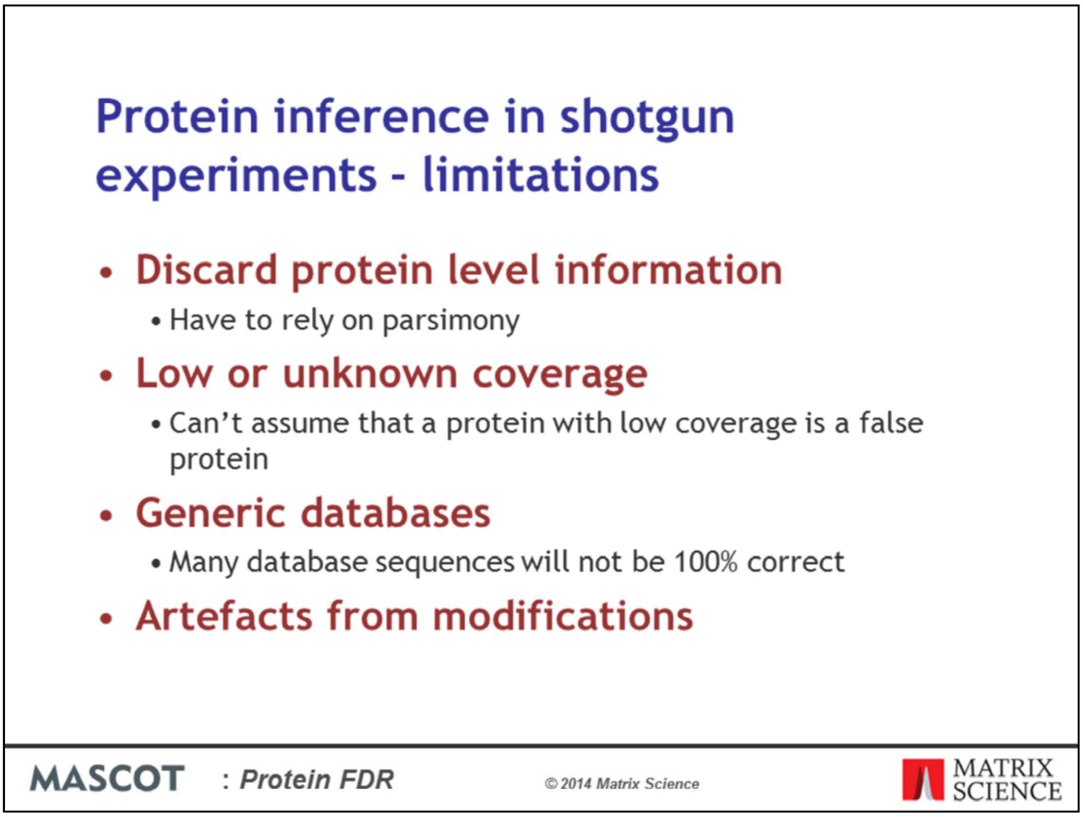
Protein inference in shotgun experiments - limitations

- Discard protein level information
  - Have to rely on parsimony
- Low or unknown coverage
  - Can’t assume that a protein with low coverage is a false protein
- Generic databases
  - Many database sequences will not be 100% correct

MASCOT : *Protein FDR*

MATRIX SCIENCE

If you don’t have time to study every hit, one way to simplify things is to search a non-redundant database. If your sample is from a well characterised organism, then SwissProt is always a good choice. Some peptide matches will be lost, which could lead to the loss of true proteins that had very low coverage, but the list of proteins with reasonable coverage will be more reliable in that you are less likely to over-report.



Protein inference in shotgun experiments - limitations

- Discard protein level information
  - Have to rely on parsimony
- Low or unknown coverage
  - Can’t assume that a protein with low coverage is a false protein
- Generic databases
  - Many database sequences will not be 100% correct
- Artefacts from modifications

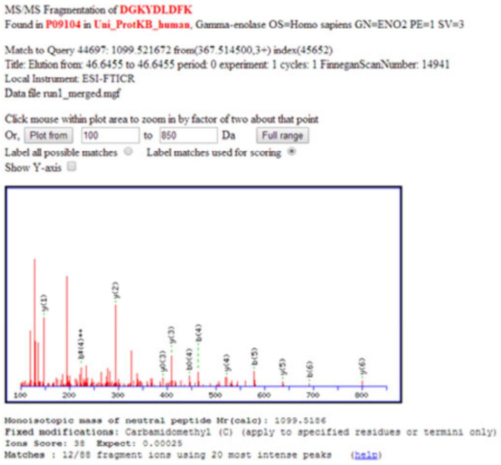
MASCOT : *Protein FDR*

© 2014 Matrix Science

MATRIX SCIENCE

Including an unnecessary modification in a search or omitting a modification that is actually present in the sample can cause false peptide matches that lead to the wrong protein being inferred.

Artefacts from modifications

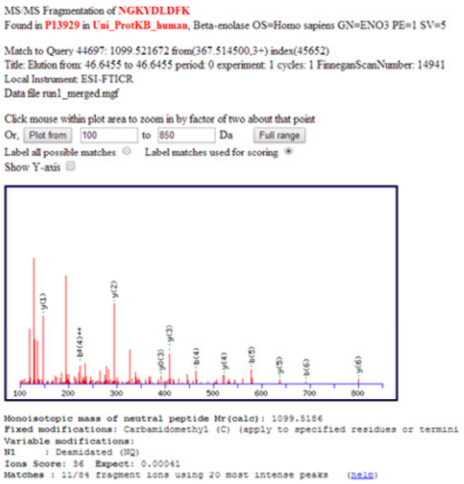


MS/MS Fragmentation of **DGKYDLDFK**
Found in **P09104** in **Uni_ProtKB_human**, Gamma-enolase OS=Homo sapiens GN=ENO2 PE=1 SV=3

Match to Query 44697: 1099.521672 from(367.514500,3+) index(45652)
Title: Elution from: 46.6455 to 46.6455 period: 0 experiment: 1 cycles: 1 FinneganScanNumber: 14941
Local Instrument: ESI-FTICR
Data file run1_merged.mgf

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 100 to 850 Da Full range
Label all possible matches ○ Label matches used for scoring ●
Show Y-axis □

```
Monoisotopic mass of neutral peptide Mr(calc): 1099.5186
Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini only)
Ions Score: 38  Expect: 0.00025
Matches : 12/88 fragment ions using 20 most intense peaks   (help)
```



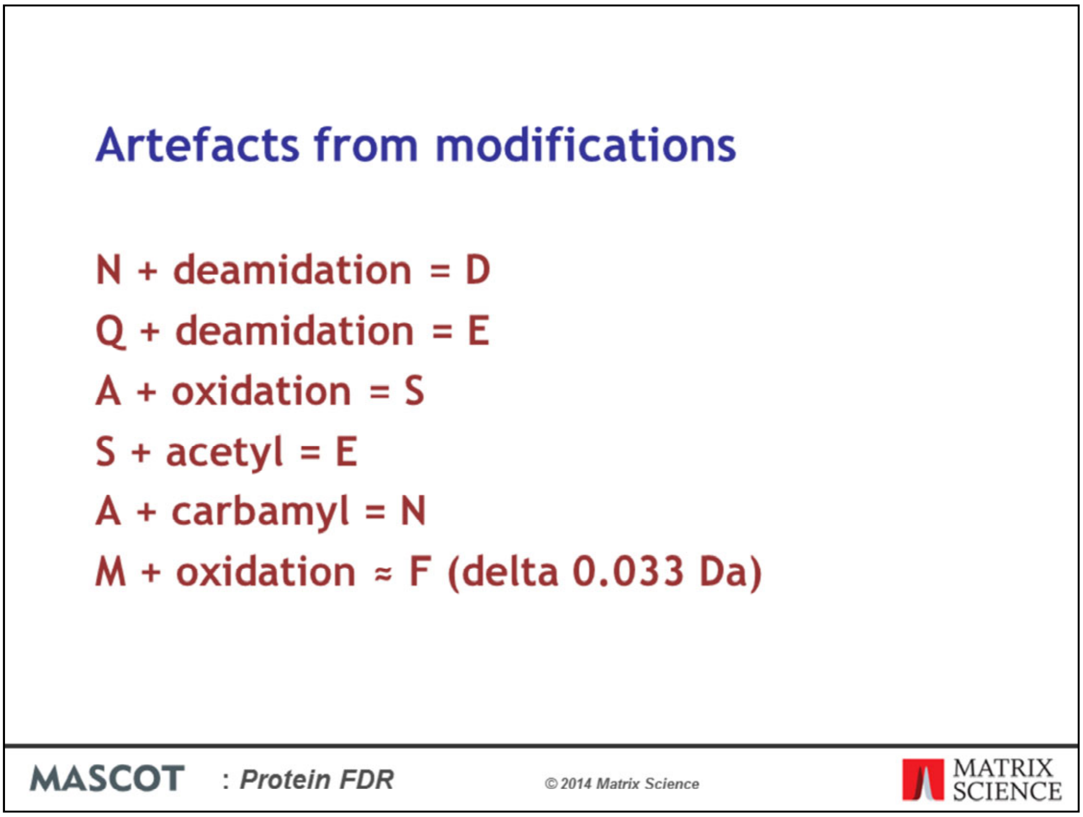
MS/MS Fragmentation of **NGKYDLDFK**
Found in **P13929** in **Uni_ProtKB_human**, Beta-enolase OS=Homo sapiens GN=ENO3 PE=1 SV=5

Match to Query 44697: 1099.521672 from(367.514500,3+) index(45652)
Title: Elution from: 46.6455 to 46.6455 period: 0 experiment: 1 cycles: 1 FinneganScanNumber: 14941
Local Instrument: ESI-FTICR
Data file run1_merged.mgf

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 100 to 850 Da Full range
Label all possible matches ○ Label matches used for scoring ●
Show Y-axis □

```
Monoisotopic mass of neutral peptide Mr(calc): 1099.5186
Fixed modifications: Carbamidomethyl (C) (apply to specified residues or termini
Variable modifications:
N1      : Deamidated (NQ)
Ions Score: 36  Expect: 0.00041
Matches : 11/84 fragment ions using 20 most intense peaks   (help)
```

The most frequent culprit is deamidation. The same peptide sequence may occur in two different proteins except that in one it has a D at a particular position and in the other an N. If the true protein is the one with the D, but the search included deamidation, we get an equally good match for the false protein. If the true protein is the one with the N, but it is mostly deamidated, we may not see the match for the true protein unless the search includes deamidation.



Artefacts from modifications

N + deamidation = D

Q + deamidation = E

A + oxidation = S

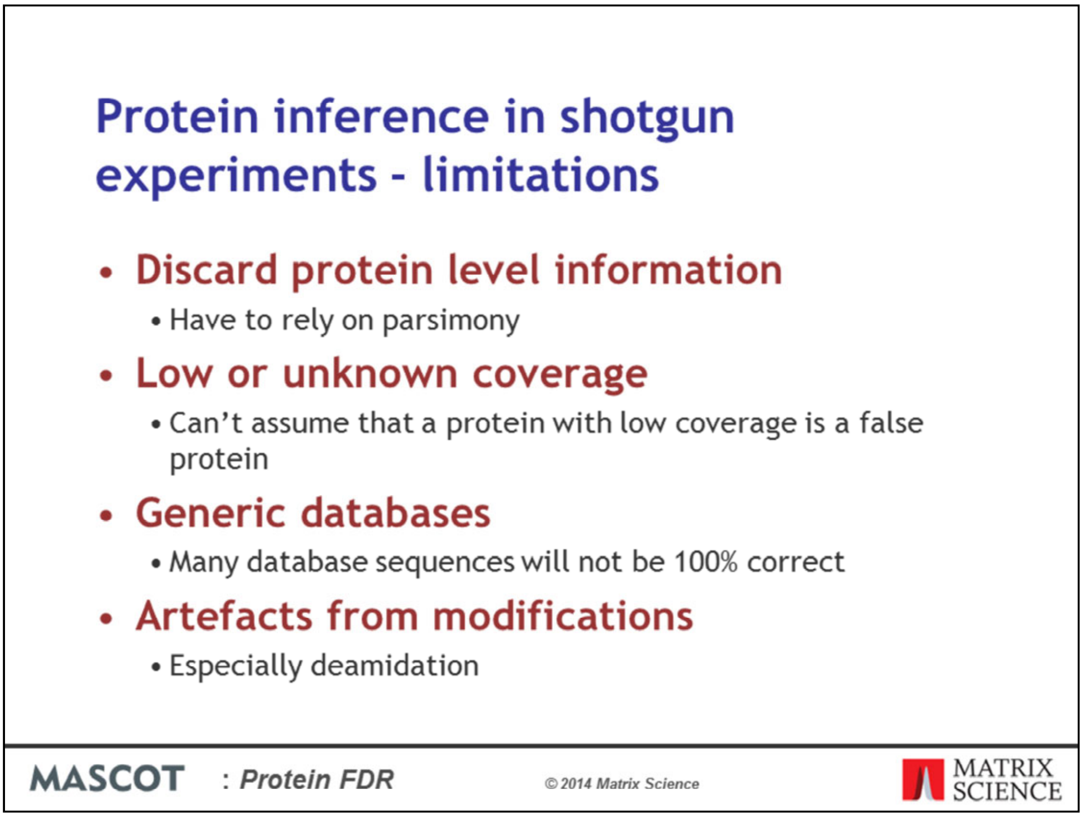
S + acetyl = E

A + carbamyl = N

M + oxidation ≈ F (delta 0.033 Da)

MASCOT : *Protein FDR* © 2014 Matrix Science MATRIX SCIENCE

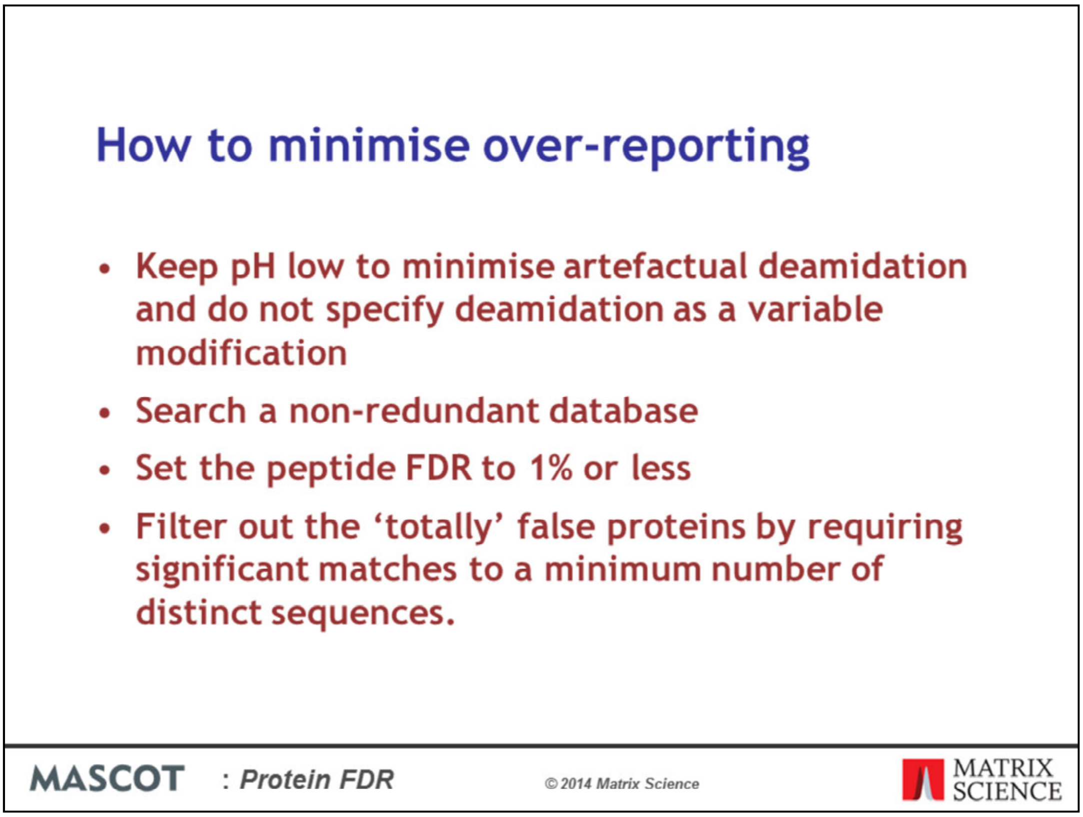
Deamidation is insidious because the substituted residue is also the site for the modification. There are many other cases where a common modification exactly compensates for a residue substitution, such as A + oxidation = S, S + acetyl = E, and A + carbamyl = N. But, the residue itself is not a common site for the modification, so the score for the match will suffer unless the modification can be located adjacent to the substitution, which will happen less frequently. The other common example is M + oxidation = F. The mass difference is 0.033 Da, so this can give an equally good match unless the mass accuracy is very high.



Protein inference in shotgun experiments - limitations

- **Discard protein level information**
  - Have to rely on parsimony
- **Low or unknown coverage**
  - Can't assume that a protein with low coverage is a false protein
- **Generic databases**
  - Many database sequences will not be 100% correct
- **Artefacts from modifications**
  - Especially deamidation

Protein inference is a complex problem, which can be made even more difficult by conflicting goals. A shotgun survey of the total protein complement of a complex sample is one thing. Detailed characterisation of individual proteins of interest is another. We cannot expect to get both from a single experiment.

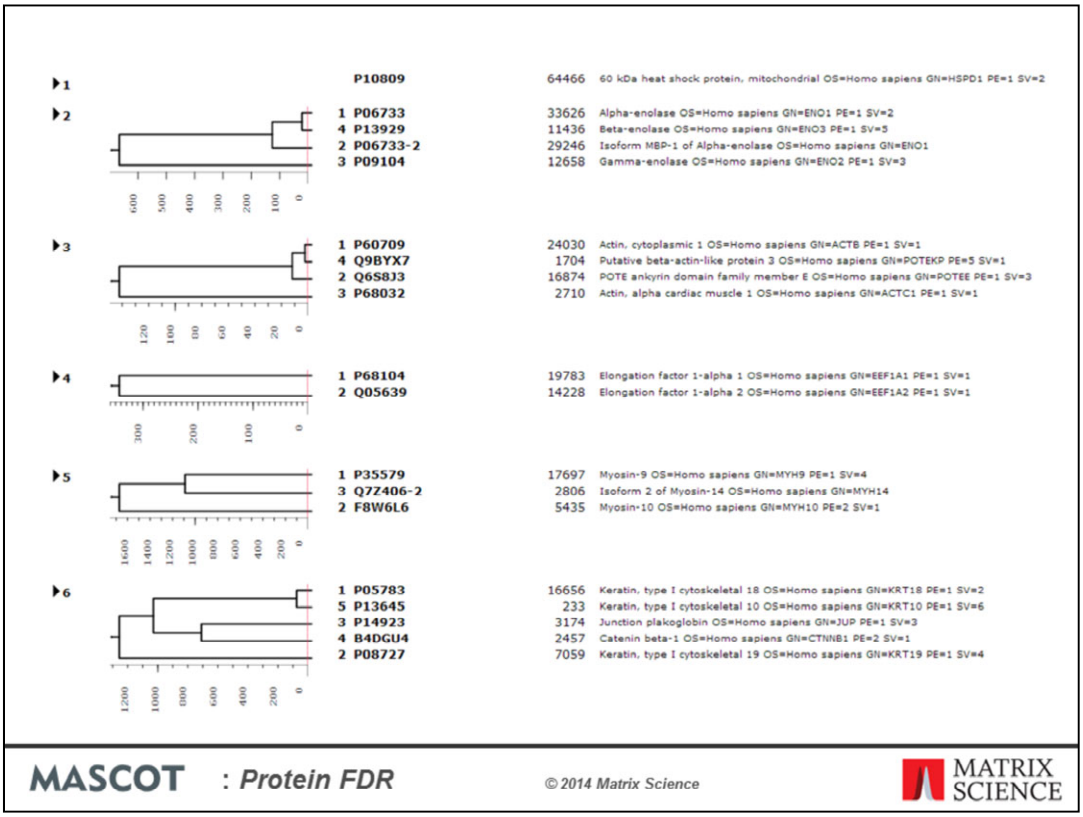


How to minimise over-reporting

- Keep pH low to minimise artefactual deamidation and do not specify deamidation as a variable modification
- Search a non-redundant database
- Set the peptide FDR to 1% or less
- Filter out the ‘totally’ false proteins by requiring significant matches to a minimum number of distinct sequences.

MASCOT : *Protein FDR* © 2014 *Matrix Science* MATRIX SCIENCE

If the primary aim is an accurate list of the proteins in a complex sample, there are several steps we can take to minimise over-reporting:



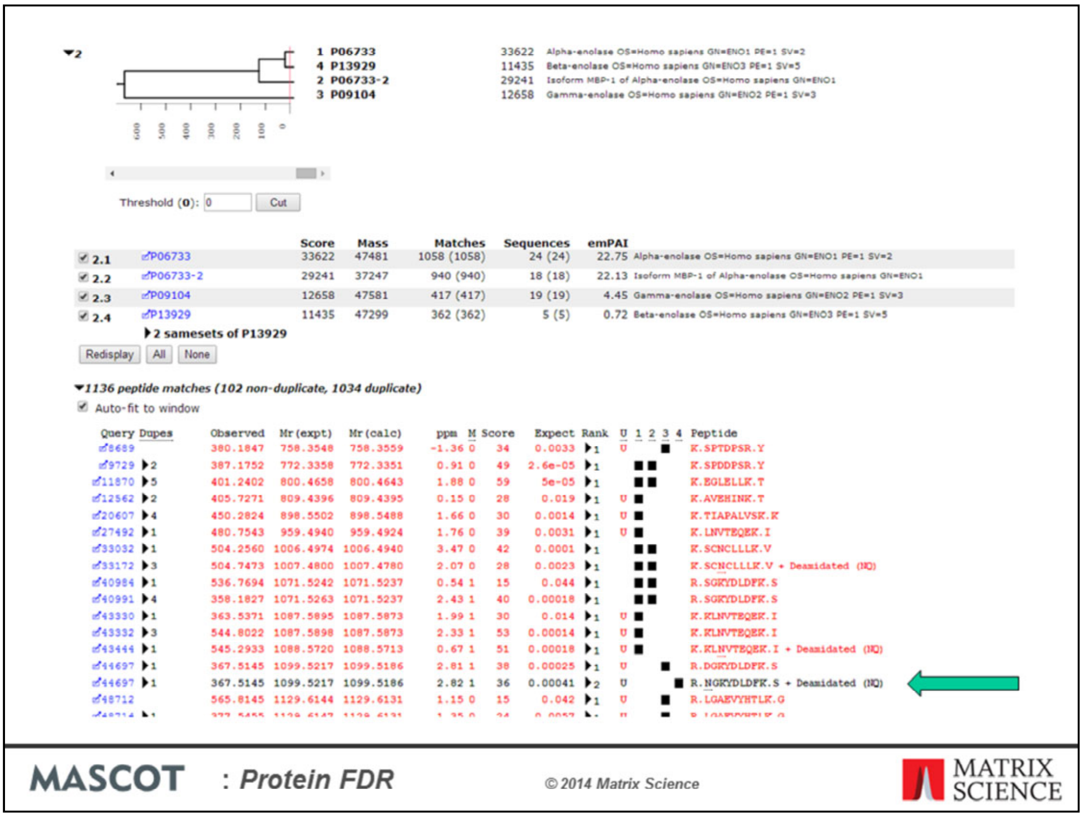
```
▸1        P10809     64466  60 kDa heat shock protein, mitochondrial OS=Homo sapiens GN=HSPD1 PE=1 SV=2

▸2      1 P06733     33626  Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2
        4 P13929     11436  Beta-enolase OS=Homo sapiens GN=ENO3 PE=1 SV=5
        2 P06733-2   29246  Isoform MBP-1 of Alpha-enolase OS=Homo sapiens GN=ENO1
        3 P09104     12658  Gamma-enolase OS=Homo sapiens GN=ENO2 PE=1 SV=3

▸3      1 P60709     24030  Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1
        4 Q9BYX7      1704  Putative beta-actin-like protein 3 OS=Homo sapiens GN=POTEKP PE=5 SV=1
        2 Q6S8J3     16874  POTE ankyrin domain family member E OS=Homo sapiens GN=POTEE PE=1 SV=3
        3 P68032      2710  Actin, alpha cardiac muscle 1 OS=Homo sapiens GN=ACTC1 PE=1 SV=1

▸4      1 P68104     19783  Elongation factor 1-alpha 1 OS=Homo sapiens GN=EEF1A1 PE=1 SV=1
        2 Q05639     14228  Elongation factor 1-alpha 2 OS=Homo sapiens GN=EEF1A2 PE=1 SV=1

▸5      1 P35579     17697  Myosin-9 OS=Homo sapiens GN=MYH9 PE=1 SV=4
        3 Q7Z406-2    2806  Isoform 2 of Myosin-14 OS=Homo sapiens GN=MYH14
        2 F8W6L6      5435  Myosin-10 OS=Homo sapiens GN=MYH10 PE=2 SV=1

▸6      1 P05783     16656  Keratin, type I cytoskeletal 18 OS=Homo sapiens GN=KRT18 PE=1 SV=2
        5 P13645       233  Keratin, type I cytoskeletal 10 OS=Homo sapiens GN=KRT10 PE=1 SV=6
        3 P14923      3174  Junction plakoglobin OS=Homo sapiens GN=JUP PE=1 SV=3
        4 B4DGU4      2457  Catenin beta-1 OS=Homo sapiens GN=CTNNB1 PE=2 SV=1
        2 P08727      7059  Keratin, type I cytoskeletal 19 OS=Homo sapiens GN=KRT19 PE=1 SV=4
```

MASCOT : *Protein FDR*

© 2014 Matrix Science

MATRIX SCIENCE

If there is still ambiguity for a protein of interest, additional experiments will be required. The protein family summary, introduced in Mascot 2.3, attempts to present the search results as clearly as possible, so that you can make up your own mind about what to believe.



▼2

1	P06733	33622	Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2
4	P13929	11435	Beta-enolase OS=Homo sapiens GN=ENO3 PE=1 SV=5
2	P06733-2	29241	Isoform MBP-1 of Alpha-enolase OS=Homo sapiens GN=ENO1
3	P09104	12658	Gamma-enolase OS=Homo sapiens GN=ENO2 PE=1 SV=3

Threshold (0): 0 Cut

		Score	Mass	Matches	Sequences	emPAI	
☑ 2.1	P06733	33622	47481	1058 (1058)	24 (24)	22.75	Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2
☑ 2.2	P06733-2	29241	37247	940 (940)	18 (18)	22.13	Isoform MBP-1 of Alpha-enolase OS=Homo sapiens GN=ENO1
☑ 2.3	P09104	12658	47581	417 (417)	19 (19)	4.45	Gamma-enolase OS=Homo sapiens GN=ENO2 PE=1 SV=3
☑ 2.4	P13929	11435	47299	362 (362)	5 (5)	0.72	Beta-enolase OS=Homo sapiens GN=ENO3 PE=1 SV=5

▶2 samesets of P13929

Redisplay All None

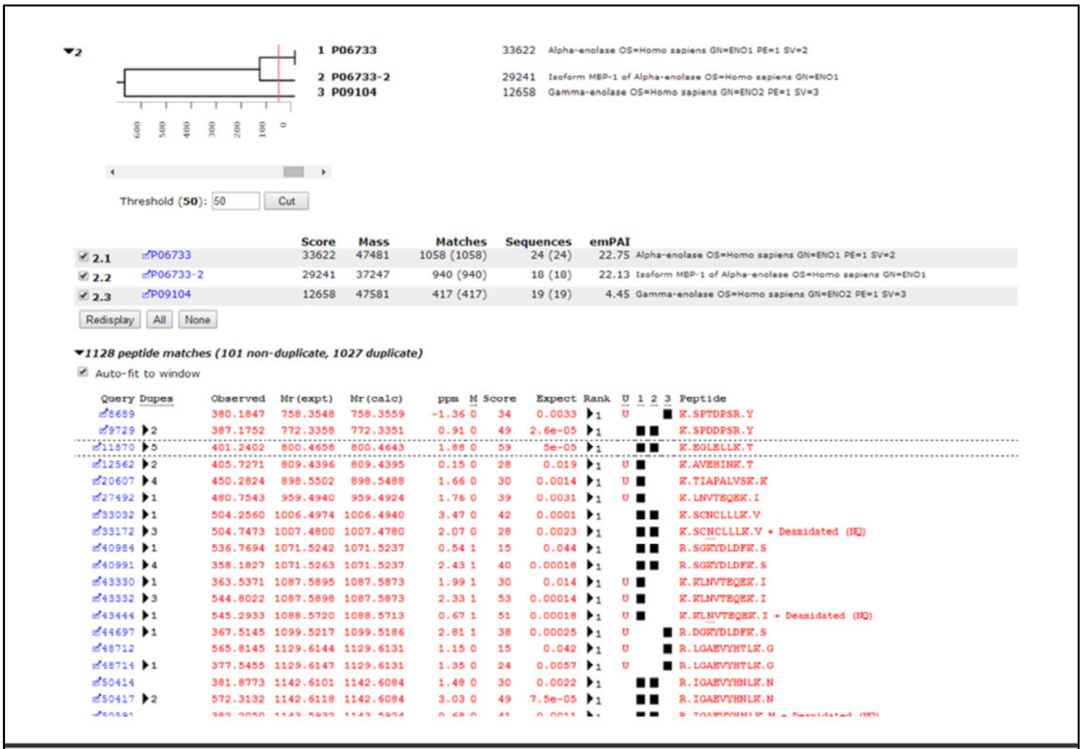
▼*1136 peptide matches (102 non-duplicate, 1034 duplicate)*

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	1	2	3	4	Peptide
8689		380.1847	758.3548	758.3559	-1.36	0	34	0.0033	▶1	U			■		K.SPTDPSR.Y
9729	▶2	387.1752	772.3358	772.3351	0.91	0	49	2.6e-05	▶1		■	■			K.SPDDPSR.Y
11870	▶5	401.2402	800.4658	800.4643	1.88	0	59	5e-05	▶1		■	■			K.EGLELLK.T
12562	▶2	405.7271	809.4396	809.4395	0.15	0	28	0.019	▶1	U	■				K.AVEHINK.T
20607	▶4	450.2824	898.5502	898.5488	1.66	0	30	0.0014	▶1	U	■				K.TIAPALVSK.K
27492	▶1	480.7543	959.4940	959.4924	1.76	0	39	0.0031	▶1	U	■				K.LNVTEQEK.I
33032	▶1	504.2560	1006.4974	1006.4940	3.47	0	42	0.0001	▶1		■	■			K.SCNCLLLK.V
33172	▶3	504.7473	1007.4800	1007.4780	2.07	0	28	0.0023	▶1		■	■			K.SCNCLLLK.V + Deamidated (NQ)
40984	▶1	536.7694	1071.5242	1071.5237	0.54	1	15	0.044	▶1		■	■			R.SGKYDLDFK.S
40991	▶4	358.1827	1071.5263	1071.5237	2.43	1	40	0.00018	▶1		■	■			R.SGKYDLDFK.S
43330	▶1	363.5371	1087.5895	1087.5873	1.99	1	30	0.014	▶1	U	■				K.KLNVTEQEK.I
43332	▶3	544.8022	1087.5898	1087.5873	2.33	1	53	0.00014	▶1	U	■				K.KLNVTEQEK.I
43444	▶1	545.2933	1088.5720	1088.5713	0.67	1	51	0.00018	▶1	U	■				K.KLNVTEQEK.I + Deamidated (NQ)
44697	▶1	367.5145	1099.5217	1099.5186	2.81	1	38	0.00025	▶1	U			■		R.DGKYDLDFK.S
44697	▶1	367.5145	1099.5217	1099.5186	2.82	1	36	0.00041	▶2	U				■	R.NGKYDLDFK.S + Deamidated (NQ)
48712		565.8145	1129.6144	1129.6131	1.15	0	15	0.042	▶1	U			■		R.LGAEVYHTLK.G

MASCOT : *Protein FDR* © 2014 Matrix Science MATRIX SCIENCE

The members of each family are differentiable proteins, connected by shared matches but with at least one unique match each. In this family, there is little difference between alpha and beta enolase. You can drop beta enolase automatically by cutting the dendrograms at a score of (say) 50. In this case, this would be a very smart move, because studying the results shows that the only match unique to beta enolase is the deamidated peptide we were looking at earlier.



1 P06733 33622 Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2
2 P06733-2 29241 Isoform MBP-1 of Alpha-enolase OS=Homo sapiens GN=ENO1
3 P09104 12658 Gamma-enolase OS=Homo sapiens GN=ENO2 PE=1 SV=3

Threshold (**50**): 50

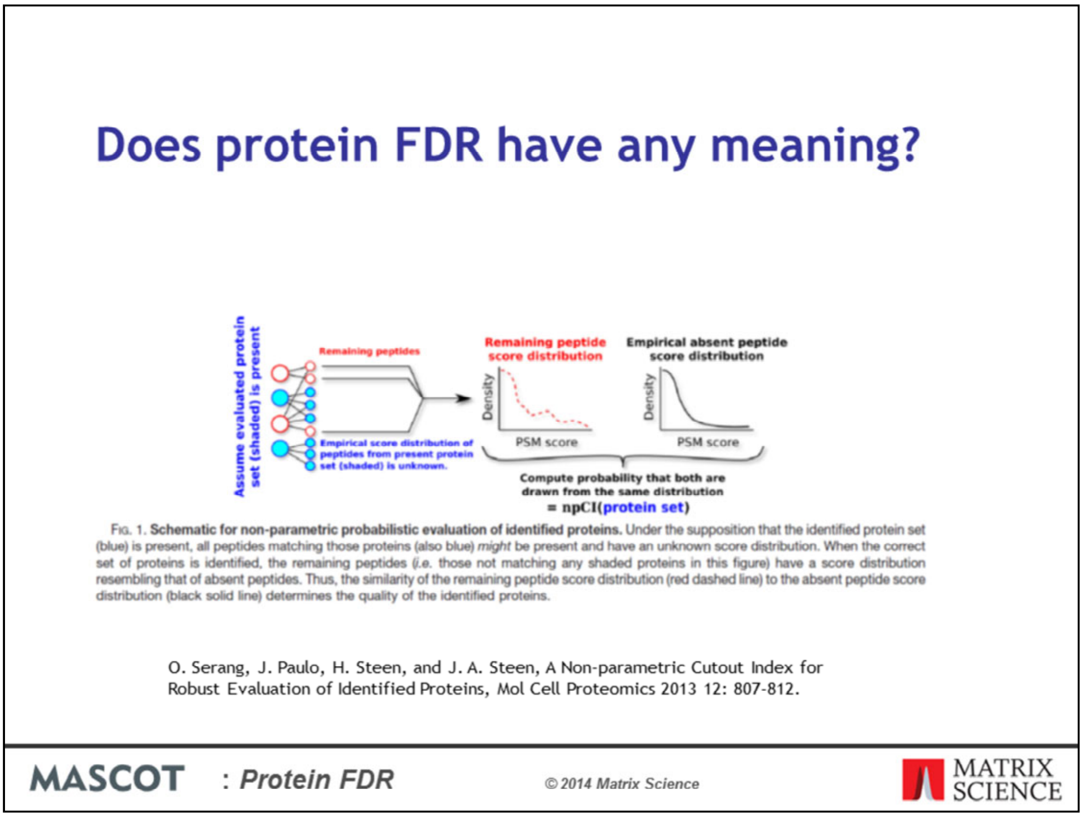
		Score	Mass	Matches	Sequences	emPAI	
☑ **2.1**	P06733	33622	47481	1058 (1058)	24 (24)	22.75	Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2
☑ **2.2**	P06733-2	29241	37247	940 (940)	18 (18)	22.13	Isoform MBP-1 of Alpha-enolase OS=Homo sapiens GN=ENO1
☑ **2.3**	P09104	12658	47581	417 (417)	19 (19)	4.45	Gamma-enolase OS=Homo sapiens GN=ENO2 PE=1 SV=3

▾***1128 peptide matches (101 non-duplicate, 1027 duplicate)***

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	1	2	3	Peptide
8689		380.1847	758.3548	758.3559	-1.36	0	34	0.0033	▸1	U			■	K.SPTDPSR.Y
9729	▸2	387.1752	772.3358	772.3351	0.91	0	49	2.6e-05	▸1		■	■		K.SPDDPSR.Y
11870	▸5	401.2402	800.4658	800.4643	1.88	0	59	5e-05	▸1		■	■		K.EGLELLK.T
12562	▸2	405.7271	809.4396	809.4395	0.15	0	28	0.019	▸1	U	■			K.AVEHINK.T
20607	▸4	450.2824	898.5502	898.5488	1.66	0	30	0.0014	▸1	U	■			K.TIAPALVSK.K
27492	▸1	480.7543	959.4940	959.4924	1.76	0	39	0.0031	▸1	U	■			K.LNVTEQEK.I
33032	▸1	504.2560	1006.4974	1006.4940	3.47	0	42	0.0001	▸1		■	■		K.SCNCLLLK.V
33172	▸3	504.7473	1007.4800	1007.4780	2.07	0	28	0.0023	▸1		■	■		K.SCNCLLLK.V + Deamidated (NQ)
40984	▸1	536.7694	1071.5242	1071.5237	0.54	1	15	0.044	▸1		■	■		R.SGKYDLDFK.S
40991	▸4	358.1827	1071.5263	1071.5237	2.43	1	40	0.00018	▸1		■	■		R.SGKYDLDFK.S
43330	▸1	363.5371	1087.5895	1087.5873	1.99	1	30	0.014	▸1	U	■			K.KLNVTEQEK.I
43332	▸3	544.8022	1087.5898	1087.5873	2.33	1	53	0.00014	▸1	U	■			K.KLNVTEQEK.I
43444	▸1	545.2933	1088.5720	1088.5713	0.67	1	51	0.00018	▸1	U	■			K.KLNVTEQEK.I + Deamidated (NQ)
44697	▸1	367.5145	1099.5217	1099.5186	2.81	1	38	0.00025	▸1	U			■	R.DGKYDLDFK.S
48712		565.8145	1129.6144	1129.6131	1.15	0	15	0.042	▸1	U			■	R.LGAEVYHTLK.G
48714	▸1	377.5455	1129.6147	1129.6131	1.35	0	24	0.0057	▸1	U			■	R.LGAEVYHTLK.G
50414		381.8773	1142.6101	1142.6084	1.48	0	30	0.0022	▸1		■	■		R.IGAEVYHNLK.N
50417	▸2	572.3132	1142.6118	1142.6084	3.03	0	49	7.5e-05	▸1		■	■		R.IGAEVYHNLK.N

By cutting the dendrogram at a score of 50, beta enolase becomes a sub-set protein. As it would have been if the search hadn’t included deamidation.



Does protein FDR have any meaning?

Assume evaluated protein set (shaded) is present
Remaining peptides
Empirical score distribution of peptides from present protein set (shaded) is unknown.
Remaining peptide score distribution
Empirical absent peptide score distribution
Density
PSM score
Density
PSM score
Compute probability that both are drawn from the same distribution
= npCI(protein set)

FIG. 1. **Schematic for non-parametric probabilistic evaluation of identified proteins.** Under the supposition that the identified protein set (blue) is present, all peptides matching those proteins (also blue) *might* be present and have an unknown score distribution. When the correct set of proteins is identified, the remaining peptides (*i.e.* those not matching any shaded proteins in this figure) have a score distribution resembling that of absent peptides. Thus, the similarity of the remaining peptide score distribution (red dashed line) to the absent peptide score distribution (black solid line) determines the quality of the identified proteins.

O. Serang, J. Paulo, H. Steen, and J. A. Steen, A Non-parametric Cutout Index for Robust Evaluation of Identified Proteins, Mol Cell Proteomics 2013 12: 807-812.

MASCOT : *Protein FDR*

© 2014 Matrix Science

MATRIX SCIENCE

And the question in the title? I think the answer is no, unless you are willing to accept the (not very useful) definition of false proteins as database entries that have only false peptide matches. If we are trying to present a list of proteins that is accurate in any biological sense, it is very important to be aware of the issues associated with protein inference in shotgun proteomics. Statistics can give us a handle on how many proteins might be present, as in this ingenious approach from Hanno Steen’s lab. But, knowing which proteins are present out of the same-sets and sub-sets and differentiable sets, which I think is implicit in the concept of protein FDR, is a very different matter.