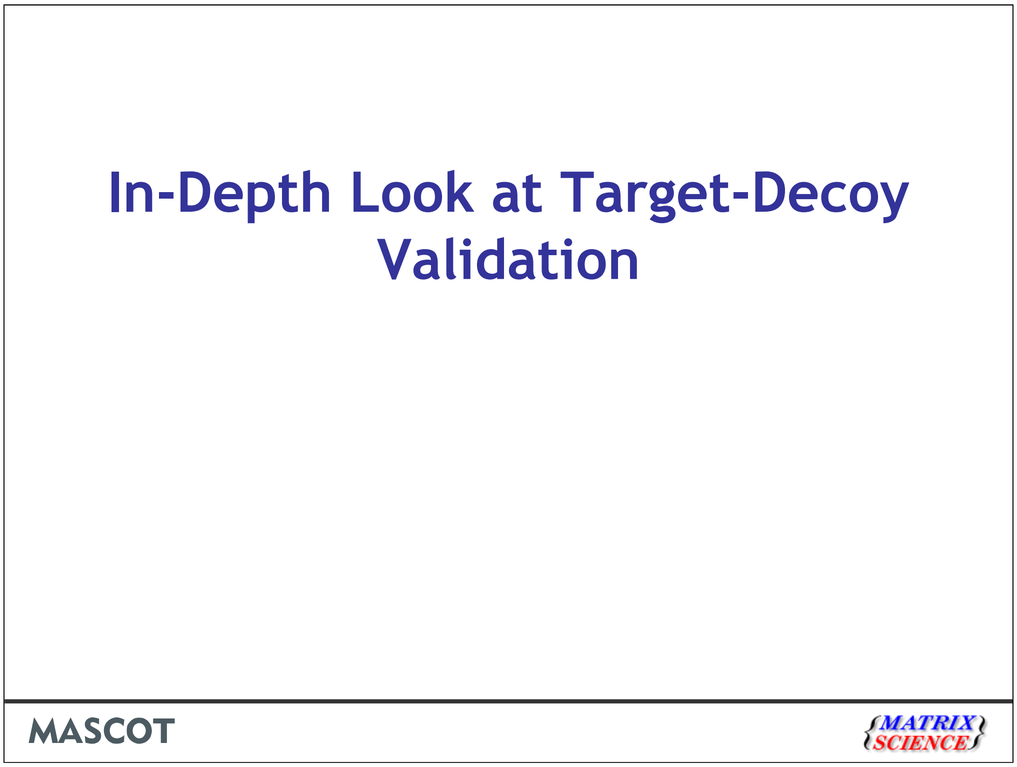
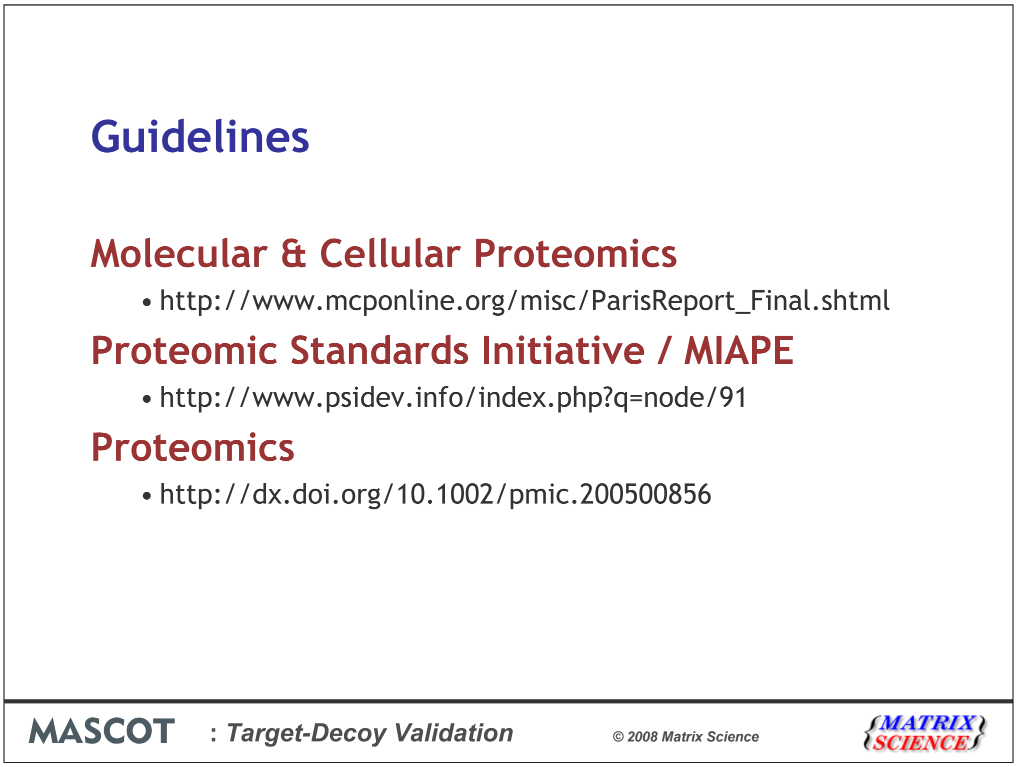


In-Depth Look at Target-Decoy Validation



Guidelines

Molecular & Cellular Proteomics

- http://www.mcponline.org/misc/ParisReport_Final.shtml

Proteomic Standards Initiative / MIAPE

- http://www.psidev.info/index.php?q=node/91

Proteomics

- http://dx.doi.org/10.1002/pmic.200500856

MASCOT : *Target-Decoy Validation* © 2008 Matrix Science MATRIX SCIENCE

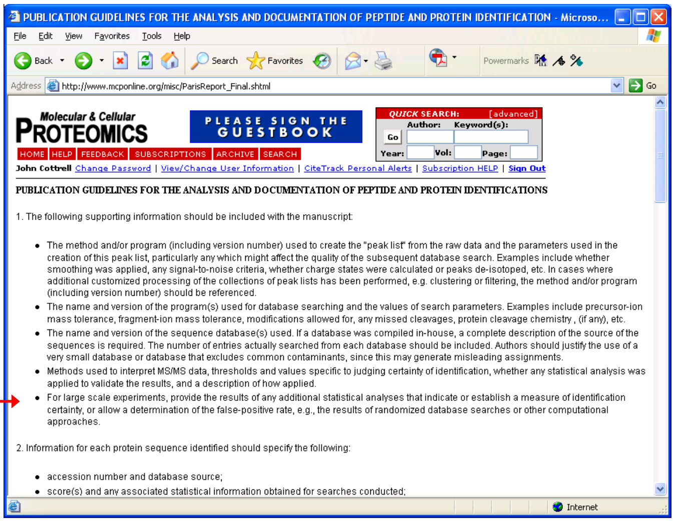
Whatever search algorithm you use, it is simply good scientific practice to validate the results. In particular, verify that the false discovery rate is not greater than claimed.

There is growing concern that some of the results in the literature, particularly from large scale searches, are not as reliable as they could be.

This has led to calls for greater stringency in the reporting of proteomics results. Most notably, the initiative taken by the Editors of Molecular and Cellular Proteomics, who held a workshop in 2005 to define a set of guidelines. The MCP guidelines have also been adopted by the Journal of Proteome Research

The PSI guidelines, which are part of the MIAPE standard, are very similar to those from MCP. The editorial board of Proteomics has also drafted a similar document. Hopefully, at some point, these parties and others will agree to a common standard.

Guidelines


PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATION - Microso...
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Powermarks
Address
http://www.mcponline.org/misc/ParisReport_Final.shtml
Go
Molecular & Cellular PROTEOMICS
PLEASE SIGN THE GUESTBOOK
QUICK SEARCH: [advanced]
Author:
Keyword(s):
Go
Year:
Vol:
Page:
HOME
HELP
FEEDBACK
SUBSCRIPTIONS
ARCHIVE
SEARCH
John Cottrell Change Password | View/Change User Information | CiteTrack Personal Alerts | Subscription HELP | Sign Out
Internet

PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATIONS

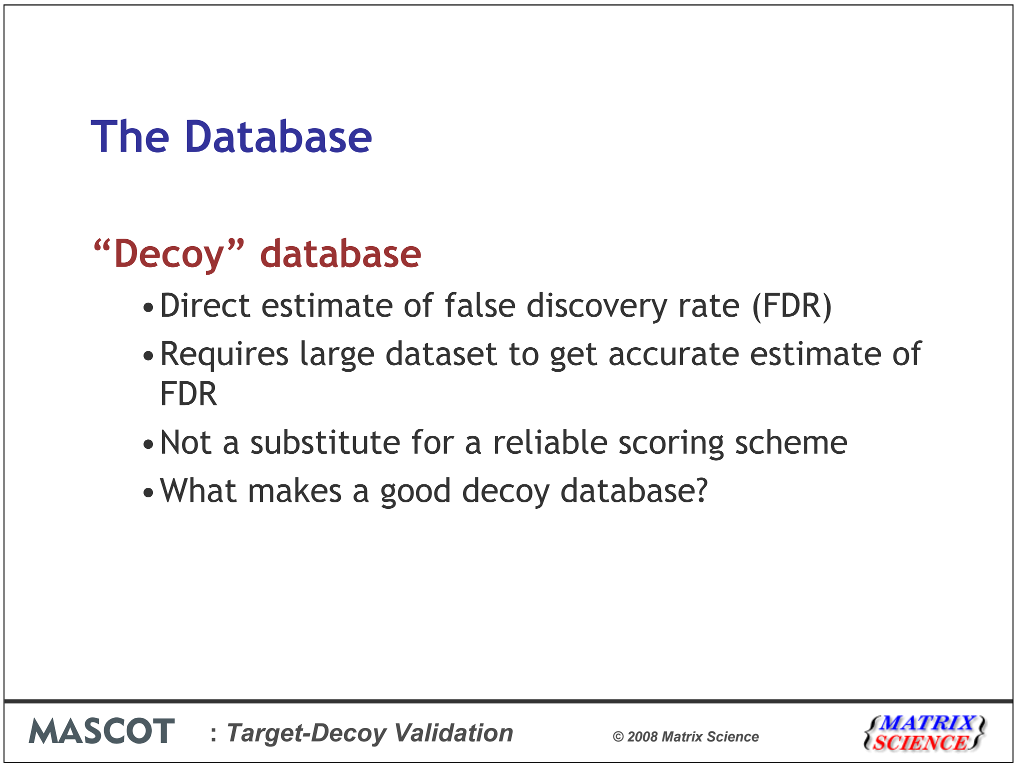
1. The following supporting information should be included with the manuscript:

- The method and/or program (including version number) used to create the "peak list" from the raw data and the parameters used in the creation of this peak list, particularly any which might affect the quality of the subsequent database search. Examples include whether smoothing was applied, any signal-to-noise criteria, whether charge states were calculated or peaks de-isotoped, etc. In cases where additional customized processing of the collections of peak lists has been performed, e.g. clustering or filtering, the method and/or program (including version number) should be referenced.
- The name and version of the program(s) used for database searching and the values of search parameters. Examples include precursor-ion mass tolerance, fragment-ion mass tolerance, modifications allowed for, any missed cleavages, protein cleavage chemistry , (if any), etc.
- The name and version of the sequence database(s) used. If a database was compiled in-house, a complete description of the source of the sequences is required. The number of entries actually searched from each database should be included. Authors should justify the use of a very small database or database that excludes common contaminants, since this may generate misleading assignments.
- Methods used to interpret MS/MS data, thresholds and values specific to judging certainty of identification, whether any statistical analysis was applied to validate the results, and a description of how applied.
- For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, e.g., the results of randomized database searches or other computational approaches.

2. Information for each protein sequence identified should specify the following:

- accession number and database source;
- score(s) and any associated statistical information obtained for searches conducted;

One of the specific recommendations in the MCP guidelines is “For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, e.g., the results of reversed or randomized database searches or other computational approaches.”



The Database

“Decoy” database

- Direct estimate of false discovery rate (FDR)
- Requires large dataset to get accurate estimate of FDR
- Not a substitute for a reliable scoring scheme
- What makes a good decoy database?

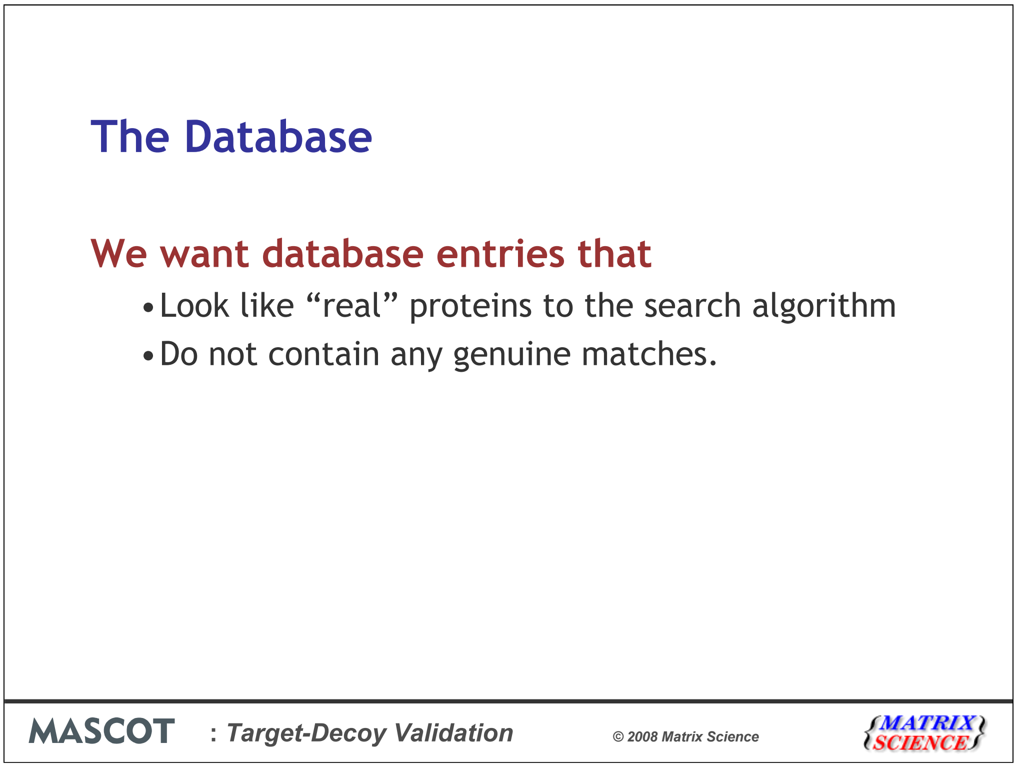
MASCOT : *Target-Decoy Validation* © 2008 Matrix Science {MATRIX SCIENCE}

I think it was the Gygi group first coined the term “decoy database search” for this approach. The idea is to repeat the search, using identical search parameters, against a database in which the sequences have been reversed or scrambled.

You do not expect to get any significant matches from the decoy database. So, the number of matches that are found is a good estimate of the number of false positives in the results from the target database.

This is an excellent validation method for MS/MS searches of large data sets. It is not as useful for a search of a small number of spectra, because the numbers are too small to give an accurate estimate of the false discovery rate. Hence, it is not a substitute for a reliable scoring algorithm.

What are the requirements for a decoy database?



The Database

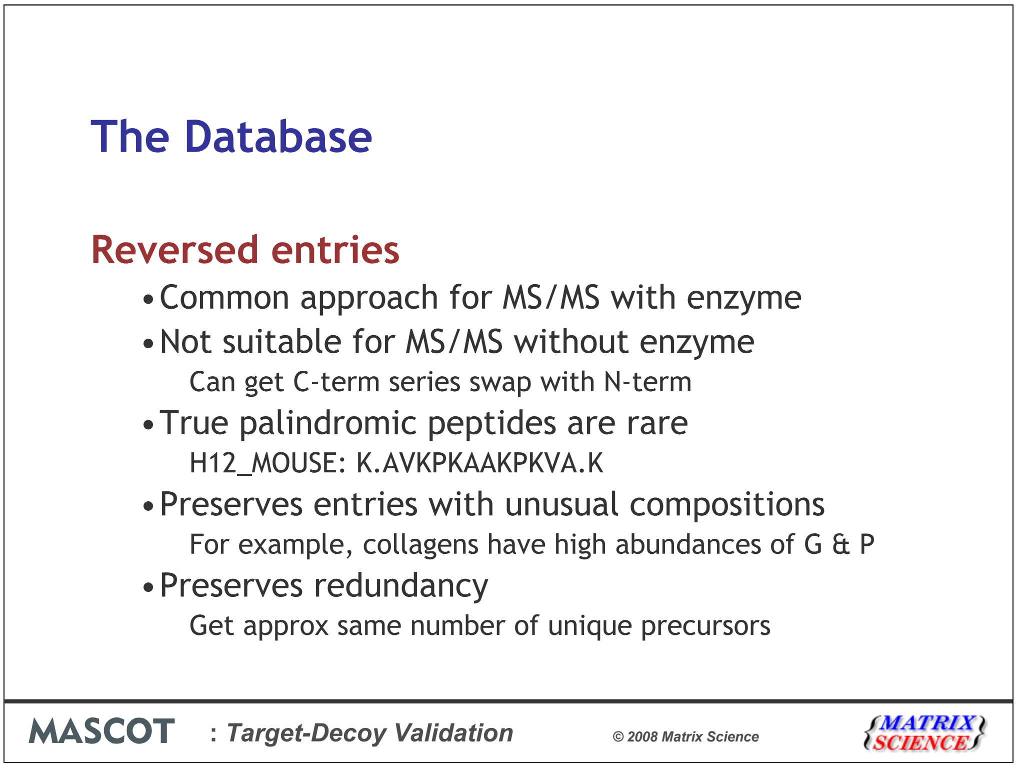
We want database entries that

- Look like “real” proteins to the search algorithm
- Do not contain any genuine matches.

This is actually quite a difficult question.

We want database entries that look like “real” proteins to the search algorithm

However, we want database entries that don’t contain genuine peptide sequences



The Database

Reversed entries

- Common approach for MS/MS with enzyme
- Not suitable for MS/MS without enzyme
  Can get C-term series swap with N-term
- True palindromic peptides are rare
  H12_MOUSE: K.AVKPKAAKPKVA.K
- Preserves entries with unusual compositions
  For example, collagens have high abundances of G & P
- Preserves redundancy
  Get approx same number of unique precursors

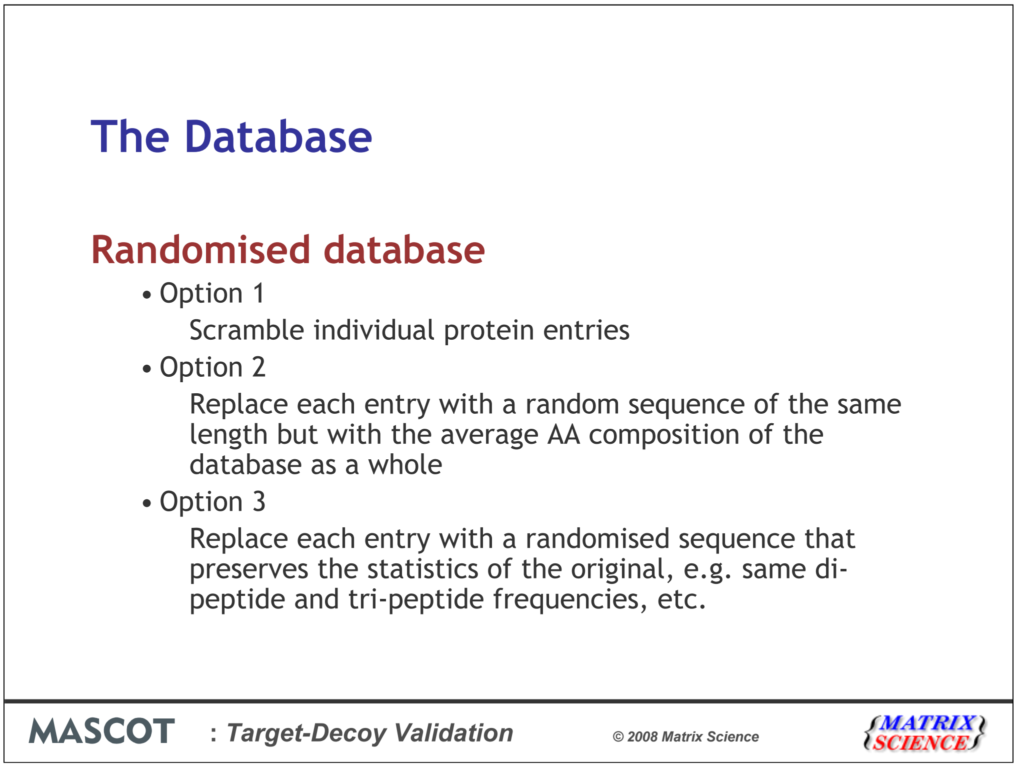
MASCOT : *Target-Decoy Validation* © 2008 Matrix Science

The Gygi group advocate simply reversing the entries in the original database. This is a reasonable approach for an MS/MS search where a specific enzyme has been used.

It is not suitable for no-enzyme MS/MS searches, especially when there are several variable mods, because it is possible to get mass shift at each end of a reversed peptide sequence that just happens to transform a genuine y series match into a false b series match or vice versa. (True palindromic peptides also exist, but are rare).

One advantage of using reversed entries is that entries with unusual amino acid compositions are preserved. For example, the very high levels of G and P in collagens or the runs of near poly-G in many keratins

Reversing the entries also preserves the degree of redundancy present in the target database, which means that the number of unique sequences will be similar. This will not be the case if the sequences are randomised.



The Database

Randomised database

- Option 1
  Scramble individual protein entries
- Option 2
  Replace each entry with a random sequence of the same length but with the average AA composition of the database as a whole
- Option 3
  Replace each entry with a randomised sequence that preserves the statistics of the original, e.g. same di-peptide and tri-peptide frequencies, etc.

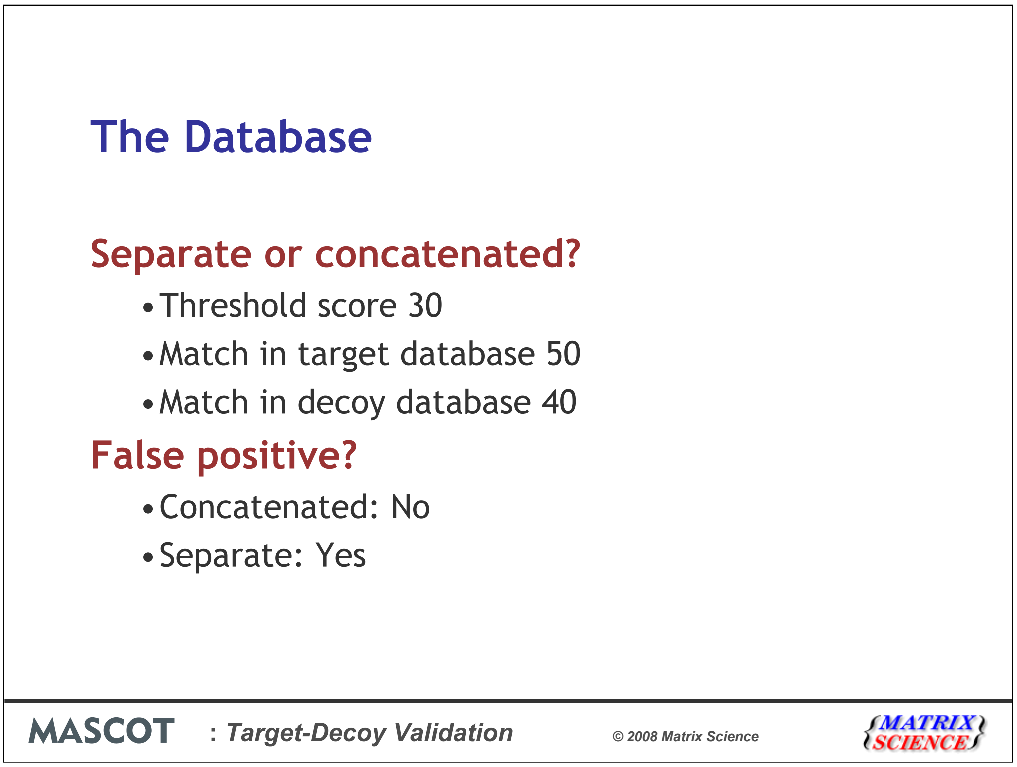
MASCOT : *Target-Decoy Validation* © 2008 Matrix Science MATRIX SCIENCE

The other approach is to randomise the sequences, rather than simply reverse them. This still leaves us with a number of possibilities.:

With option 1, an entry that is Cys rich or Gly rich will remain so. This can create problems with entries which have long runs of poly R or K. When scrambled, these go from producing fewer tryptic peptides than you would expect for a protein of that size to producing far too many

Option 2 is our preferred approach and has been implemented as an automatic part of a Mascot search, as I will describe shortly.

Option 3 is unnecessary unless the search algorithm is known to take di-peptide and tri-peptide frequencies into account. This is not the case with Mascot and I suspect not the case with any other algorithm. The only time this might have a noticeable effect is with enzyme cleavage specificity. To get the same number of peptides with strict trypsin, you need to preserve the frequencies of KP and RP when randomising.



The Database

Separate or concatenated?

- Threshold score 30
- Match in target database 50
- Match in decoy database 40

False positive?

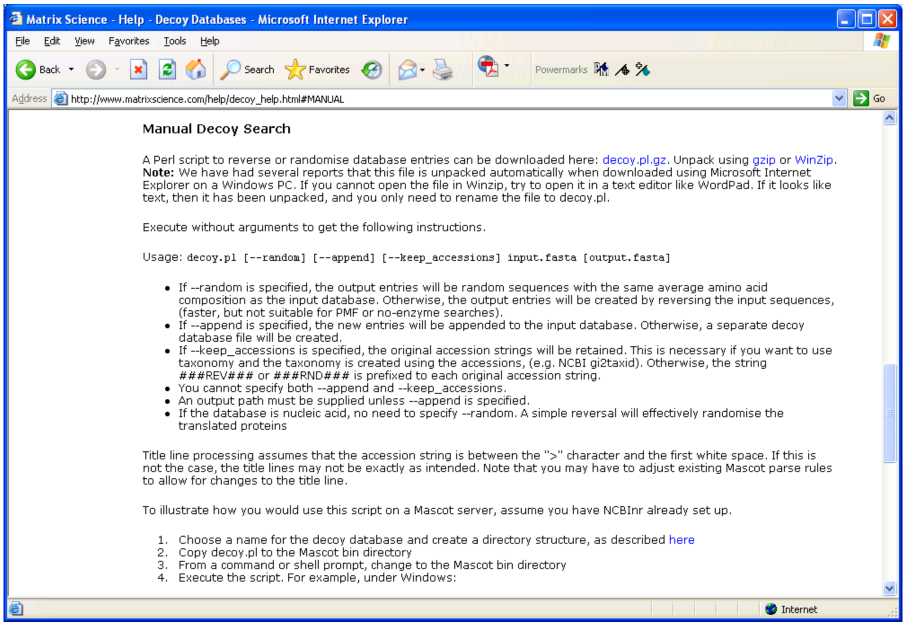
- Concatenated: No
- Separate: Yes

MASCOT : *Target-Decoy Validation*

MATRIX SCIENCE

The Gygi group advocate searching a database in which the real and decoy sequences have been concatenated. A more conservative approach is to search the two databases independently. If the Mascot score threshold for a given spectrum is (say) 30, and we get a match of 50 from the real database and 40 from the decoy database, this would not count as a false positive from a concatenated database, but it would count as a false positive if the two had been searched independently.

Note that, when you search a concatenated database, you must double the number of matches found in the decoy because a random match is equally likely to occur in the target half.



Matrix Science - Help - Decoy Databases - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Powermarks
Address http://www.matrixscience.com/help/decoy_help.html#MANUAL
Go
Internet

Manual Decoy Search

A Perl script to reverse or randomise database entries can be downloaded here: decoy.pl.gz. Unpack using gzip or WinZip. **Note:** We have had several reports that this file is unpacked automatically when downloaded using Microsoft Internet Explorer on a Windows PC. If you cannot open the file in Winzip, try to open it in a text editor like WordPad. If it looks like text, then it has been unpacked, and you only need to rename the file to decoy.pl.

Execute without arguments to get the following instructions.

Usage: `decoy.pl [--random] [--append] [--keep_accessions] input.fasta [output.fasta]`

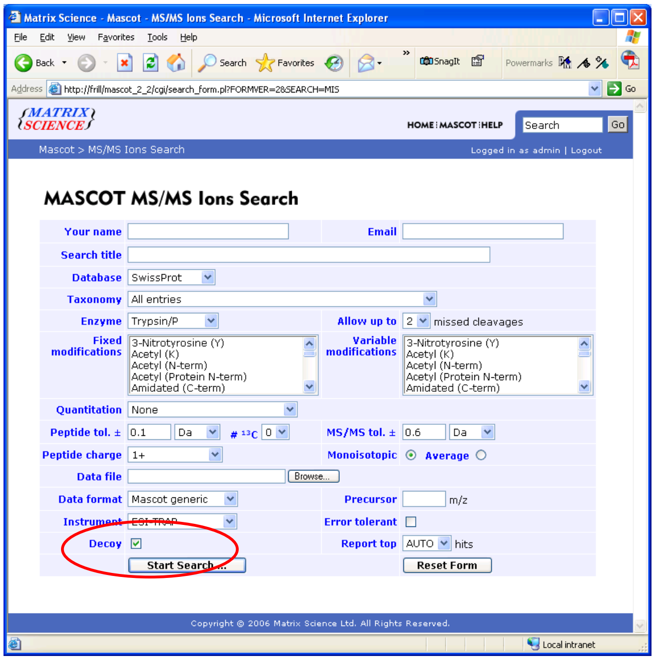
- If --random is specified, the output entries will be random sequences with the same average amino acid composition as the input database. Otherwise, the output entries will be created by reversing the input sequences, (faster, but not suitable for PMF or no-enzyme searches).
- If --append is specified, the new entries will be appended to the input database. Otherwise, a separate decoy database file will be created.
- If --keep_accessions is specified, the original accession strings will be retained. This is necessary if you want to use taxonomy and the taxonomy is created using the accessions, (e.g. NCBI gi2taxid). Otherwise, the string ###REV### or ###RND### is prefixed to each original accession string.
- You cannot specify both --append and --keep_accessions.
- An output path must be supplied unless --append is specified.
- If the database is nucleic acid, no need to specify --random. A simple reversal will effectively randomise the translated proteins

Title line processing assumes that the accession string is between the ">" character and the first white space. If this is not the case, the title lines may not be exactly as intended. Note that you may have to adjust existing Mascot parse rules to allow for changes to the title line.

To illustrate how you would use this script on a Mascot server, assume you have NCBInr already set up.

1. Choose a name for the decoy database and create a directory structure, as described here
2. Copy decoy.pl to the Mascot bin directory
3. From a command or shell prompt, change to the Mascot bin directory
4. Execute the script. For example, under Windows:

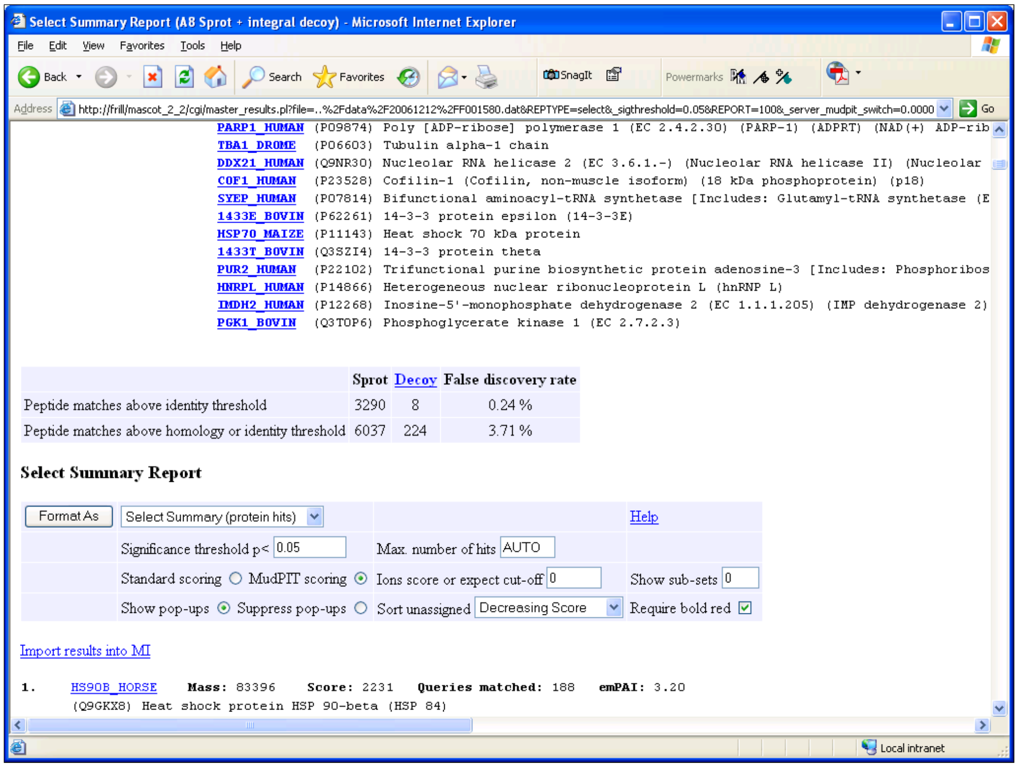
On our public web site there is a help page devoted to decoy database searches. You can download a utility program from this page that allows you to create a randomised or reversed database. You can use this if you have an old version of Mascot, or if you want to use a decoy with reversed entries.


Matrix Science - Mascot - MS/MS Ions Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites SnagIt Powermarks
Address http://frill/mascot_2_2/cgi/search_form.pl?FORMVER=2&SEARCH=MIS
Go
MATRIX SCIENCE
HOME | MASCOT | HELP
Search Go
Mascot > MS/MS Ions Search
Logged in as admin | Logout
Copyright © 2006 Matrix Science Ltd. All Rights Reserved.
Local intranet

MASCOT MS/MS Ions Search

Your name		Email	
Search title			
Database	SwissProt		
Taxonomy	All entries		
Enzyme	Trypsin/P	Allow up to	2 missed cleavages
Fixed modifications	3-Nitrotyrosine (Y) Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term)	Variable modifications	3-Nitrotyrosine (Y) Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term)
Quantitation	None		
Peptide tol. ±	0.1 Da # ^{13}C 0	MS/MS tol. ±	0.6 Da
Peptide charge	1+	Monoisotopic ●	Average ○
Data file	Browse...		
Data format	Mascot generic	Precursor	m/z
Instrument	ESI-TRAP	Error tolerant	□
Decoy	☑	Report top	AUTO hits
Start Search ...		Reset Form	

Because more and more people wish to perform decoy searches routinely, we added this into Mascot 2.2 as a built-in part of the search. If you choose the Decoy checkbox on the search form, every time a protein sequence from the target database is tested, a random sequence of the same length is automatically generated and tested. The average amino acid composition of each random sequence is the same as the average composition of the target database. The matches and scores for the random sequences are recorded separately in the result file. The net result is identical to searching a separate, randomised database.



Address http://frill/mascot_2_2/cgi/master_results.pl?file=..%2Fdata%2F20061212%2FF001580.dat&REPTYPE=select&_sigthreshold=0.05&REPORT=100&_server_mudpit_switch=0.0000

```
PARP1_HUMAN (P09874) Poly [ADP-ribose] polymerase 1 (EC 2.4.2.30) (PARP-1) (ADPRT) (NAD(+) ADP-rib
TBA1_DROME  (P06603) Tubulin alpha-1 chain
DDX21_HUMAN (Q9NR30) Nucleolar RNA helicase 2 (EC 3.6.1.-) (Nucleolar RNA helicase II) (Nucleolar
COF1_HUMAN  (P23528) Cofilin-1 (Cofilin, non-muscle isoform) (18 kDa phosphoprotein) (p18)
SYEP_HUMAN  (P07814) Bifunctional aminoacyl-tRNA synthetase [Includes: Glutamyl-tRNA synthetase (E
1433E_BOVIN (P62261) 14-3-3 protein epsilon (14-3-3E)
HSP70_MAIZE (P11143) Heat shock 70 kDa protein
1433T_BOVIN (Q3SZI4) 14-3-3 protein theta
PUR2_HUMAN  (P22102) Trifunctional purine biosynthetic protein adenosine-3 [Includes: Phosphoribos
HNRPL_HUMAN (P14866) Heterogeneous nuclear ribonucleoprotein L (hnRNP L)
IMDH2_HUMAN (P12268) Inosine-5'-monophosphate dehydrogenase 2 (EC 1.1.1.205) (IMP dehydrogenase 2)
PGK1_BOVIN  (Q3TOP6) Phosphoglycerate kinase 1 (EC 2.7.2.3)
```

	Sprot	Decoy	False discovery rate
Peptide matches above identity threshold	3290	8	0.24 %
Peptide matches above homology or identity threshold	6037	224	3.71 %

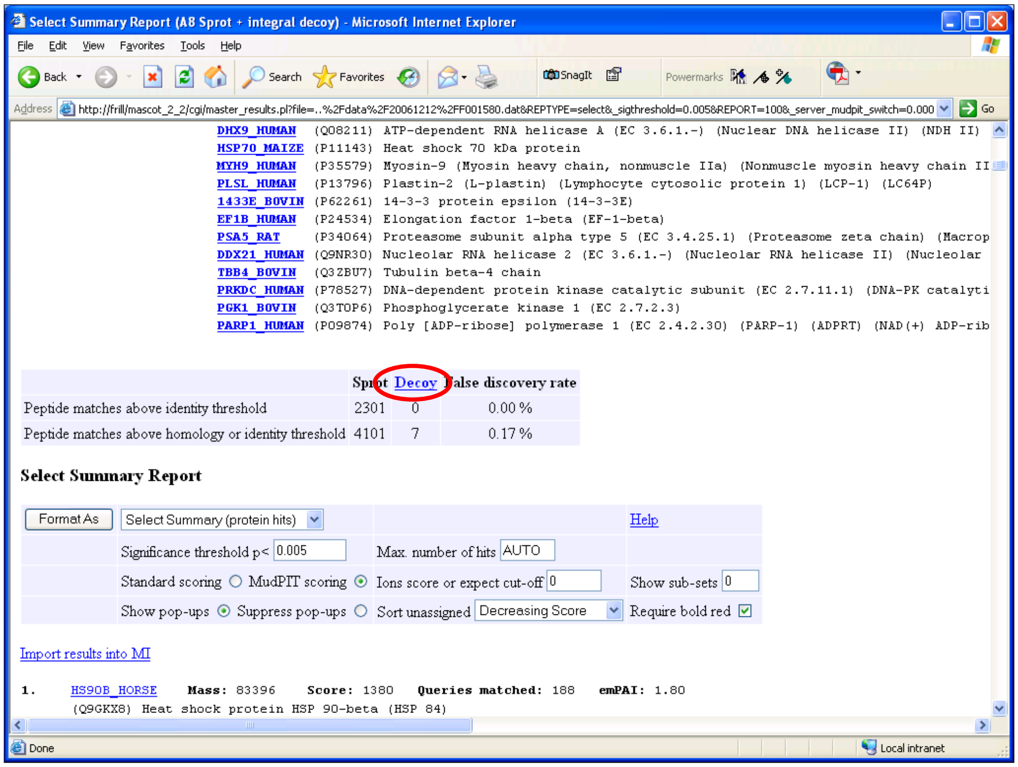
Select Summary Report

Format As	Select Summary (protein hits)		Help
	Significance threshold p< 0.05	Max. number of hits AUTO	
	Standard scoring ○ MudPIT scoring ●	Ions score or expect cut-off 0	Show sub-sets 0
	Show pop-ups ● Suppress pop-ups ○	Sort unassigned Decreasing Score	Require bold red ☑

Import results into MI

```
1.     HS90B_HORSE     Mass: 83396     Score: 2231    Queries matched: 188    emPAI: 3.20
       (Q9GKX8) Heat shock protein HSP 90-beta (HSP 84)
```

When the search is complete, the statistics for matches to the random sequences, which are effectively sequences from a decoy database, are reported in the result header. If you change the significance threshold, the numbers are recalculated. For example, if we change the threshold from 5% to 0.5% …



Select Summary Report (A8 Sprot + integral decoy) - Microsoft Internet Explorer

http://frill/mascot_2_2/cgi/master_results.pl?file=..%2Fdata%2F20061212%2FF001580.dat&REPTYPE=select&_sigthreshold=0.005&REPORT=100&_server_mudpit_switch=0.000

```
DHX9_HUMAN   (Q08211) ATP-dependent RNA helicase A (EC 3.6.1.-) (Nuclear DNA helicase II) (NDH II)
HSP70_MAIZE  (P11143) Heat shock 70 kDa protein
MYH9_HUMAN   (P35579) Myosin-9 (Myosin heavy chain, nonmuscle IIa) (Nonmuscle myosin heavy chain II
PLSL_HUMAN   (P13796) Plastin-2 (L-plastin) (Lymphocyte cytosolic protein 1) (LCP-1) (LC64P)
1433E_BOVIN  (P62261) 14-3-3 protein epsilon (14-3-3E)
EF1B_HUMAN   (P24534) Elongation factor 1-beta (EF-1-beta)
PSA5_RAT     (P34064) Proteasome subunit alpha type 5 (EC 3.4.25.1) (Proteasome zeta chain) (Macrop
DDX21_HUMAN  (Q9NR30) Nucleolar RNA helicase 2 (EC 3.6.1.-) (Nucleolar RNA helicase II) (Nucleolar
TBB4_BOVIN   (Q3ZBU7) Tubulin beta-4 chain
PRKDC_HUMAN  (P78527) DNA-dependent protein kinase catalytic subunit (EC 2.7.11.1) (DNA-PK catalyti
PGK1_BOVIN   (Q3TOP6) Phosphoglycerate kinase 1 (EC 2.7.2.3)
PARP1_HUMAN  (P09874) Poly [ADP-ribose] polymerase 1 (EC 2.4.2.30) (PARP-1) (ADPRT) (NAD(+) ADP-rib
```

	Sprot	Decoy	False discovery rate
Peptide matches above identity threshold	2301	0	0.00 %
Peptide matches above homology or identity threshold	4101	7	0.17 %

Select Summary Report

Format As	Select Summary (protein hits)		Help
	Significance threshold p< 0.005	Max. number of hits AUTO	
	Standard scoring ○ MudPIT scoring ●	Ions score or expect cut-off 0	Show sub-sets 0
	Show pop-ups ● Suppress pop-ups ○	Sort unassigned Decreasing Score	Require bold red ☑

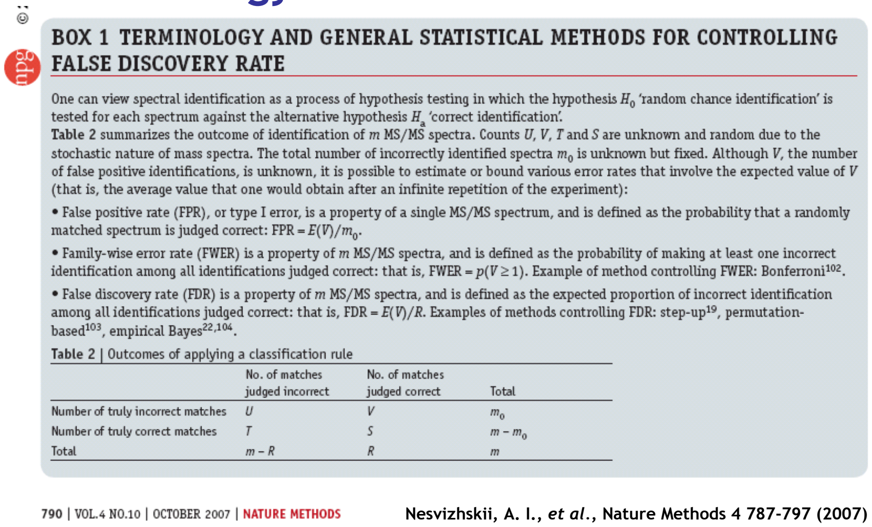
Import results into MI

```
1.   HS90B_HORSE   Mass: 83396   Score: 1380   Queries matched: 188   emPAI: 1.80
     (Q9GKX8) Heat shock protein HSP 90-beta (HSP 84)
```

Done   Local intranet

The false discovery rate drops accordingly. Of course, so does the number of true positives If you click the hyperlink, you display the results from searching the randomised database.

Terminology



BOX 1 TERMINOLOGY AND GENERAL STATISTICAL METHODS FOR CONTROLLING FALSE DISCOVERY RATE

One can view spectral identification as a process of hypothesis testing in which the hypothesis H_0 'random chance identification' is tested for each spectrum against the alternative hypothesis H_a 'correct identification'.

Table 2 summarizes the outcome of identification of m MS/MS spectra. Counts U, V, T and S are unknown and random due to the stochastic nature of mass spectra. The total number of incorrectly identified spectra m_0 is unknown but fixed. Although V, the number of false positive identifications, is unknown, it is possible to estimate or bound various error rates that involve the expected value of V (that is, the average value that one would obtain after an infinite repetition of the experiment):

- False positive rate (FPR), or type I error, is a property of a single MS/MS spectrum, and is defined as the probability that a randomly matched spectrum is judged correct: $\mathrm{FPR} = E(V)/m_0$.
- Family-wise error rate (FWER) is a property of m MS/MS spectra, and is defined as the probability of making at least one incorrect identification among all identifications judged correct: that is, $\mathrm{FWER} = p(V \geq 1)$. Example of method controlling FWER: Bonferroni[102].
- False discovery rate (FDR) is a property of m MS/MS spectra, and is defined as the expected proportion of incorrect identification among all identifications judged correct: that is, $\mathrm{FDR} = E(V)/R$. Examples of methods controlling FDR: step-up[19], permutation-based[103], empirical Bayes[22,104].

Table 2 | Outcomes of applying a classification rule

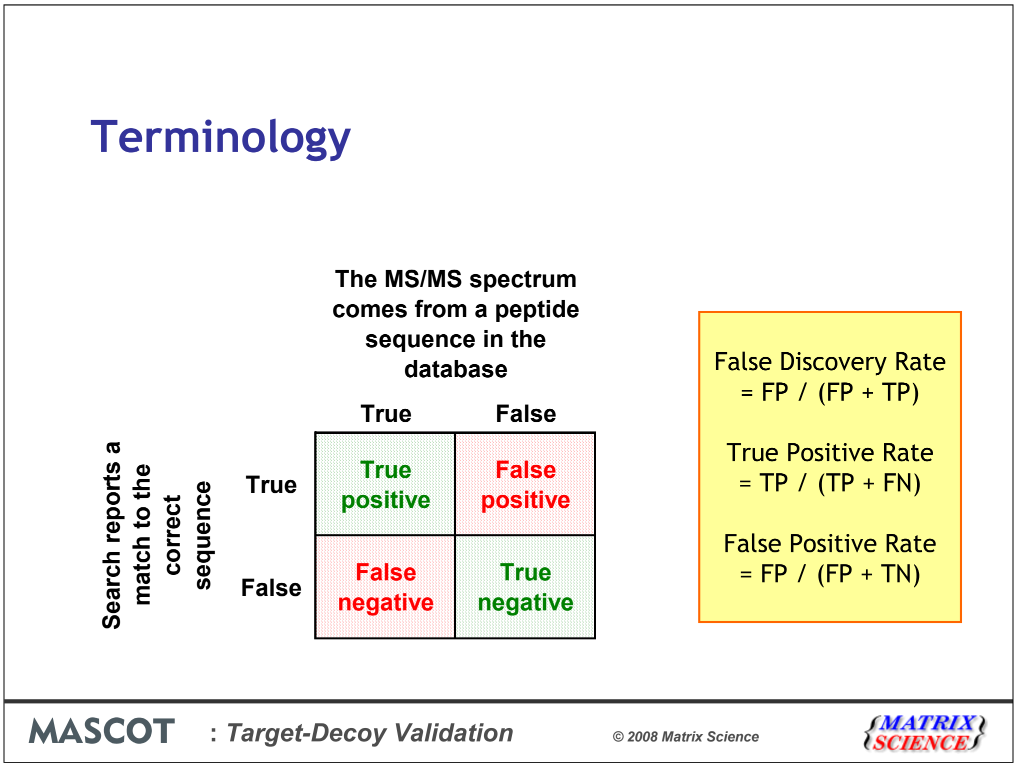
	No. of matches judged incorrect	No. of matches judged correct	Total
Number of truly incorrect matches	U	V	m_0
Number of truly correct matches	T	S	$m - m_0$
Total	$m - R$	R	m

790 | VOL.4 NO.10 | OCTOBER 2007 | NATURE METHODS

Nesvizhskii, A. I., *et al.*, Nature Methods 4 787-797 (2007)

MASCOT : *Target-Decoy Validation* © 2008 Matrix Science MATRIX SCIENCE

There is some confusion in the literature over terminology. Should we talk about false discovery rate or false positive rate? Some explanations are quite difficult to understand. This recent review gives a rather mathematical definition.



Terminology

		The MS/MS spectrum comes from a peptide sequence in the database	
		True	False
Search reports a match to the correct sequence	True	True positive	False positive
	False	False negative	True negative

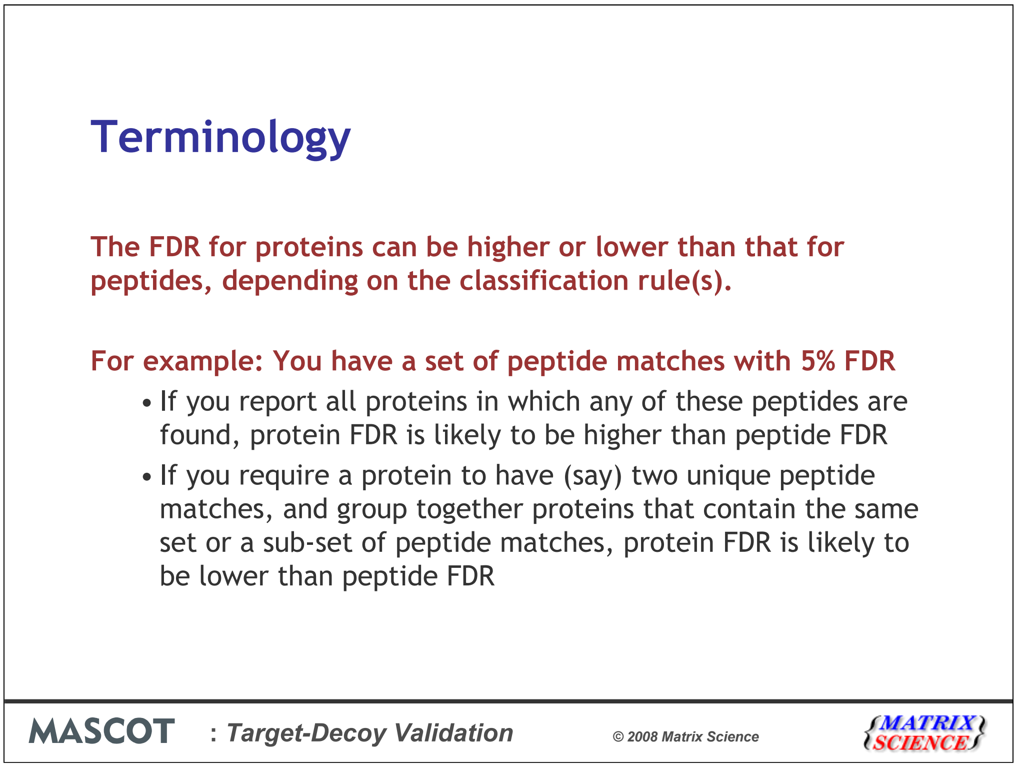
False Discovery Rate
= FP / (FP + TP)

True Positive Rate
= TP / (TP + FN)

False Positive Rate
= FP / (FP + TN)

If TP is true positive matches and FP is false positive matches, the number of matches in the target database is TP + FP and the number of matches in the decoy database is FP. The quantity that is reported is the False Discovery Rate = FP / (FP + TP)

True Positive Rate and False Positive Rate are different quantities, and I'll return to this topic later.



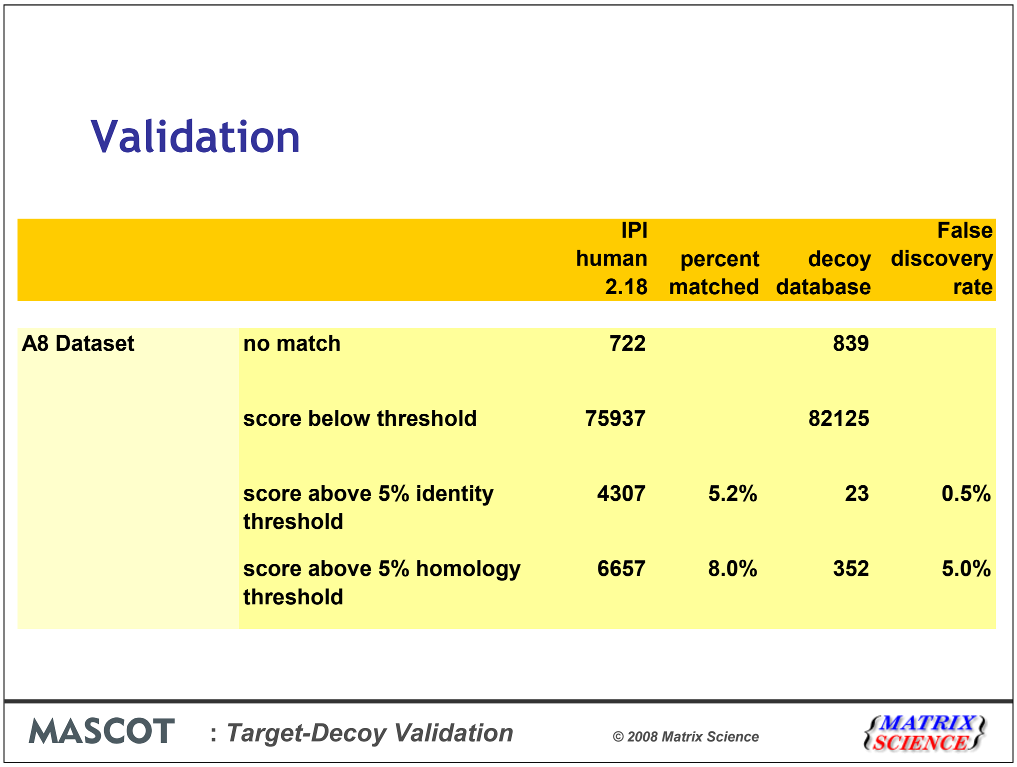
Terminology

The FDR for proteins can be higher or lower than that for peptides, depending on the classification rule(s).

For example: You have a set of peptide matches with 5% FDR

- If you report all proteins in which any of these peptides are found, protein FDR is likely to be higher than peptide FDR
- If you require a protein to have (say) two unique peptide matches, and group together proteins that contain the same set or a sub-set of peptide matches, protein FDR is likely to be lower than peptide FDR

It is very important to distinguish between peptide and protein FDR. They are usually very different.



Validation

		IPI human 2.18	percent matched	decoy database	False discovery rate
A8 Dataset	no match	722		839	
	score below threshold	75937		82125	
	score above 5% identity threshold	4307	5.2%	23	0.5%
	score above 5% homology threshold	6657	8.0%	352	5.0%

MASCOT : *Target-Decoy Validation* © 2008 Matrix Science

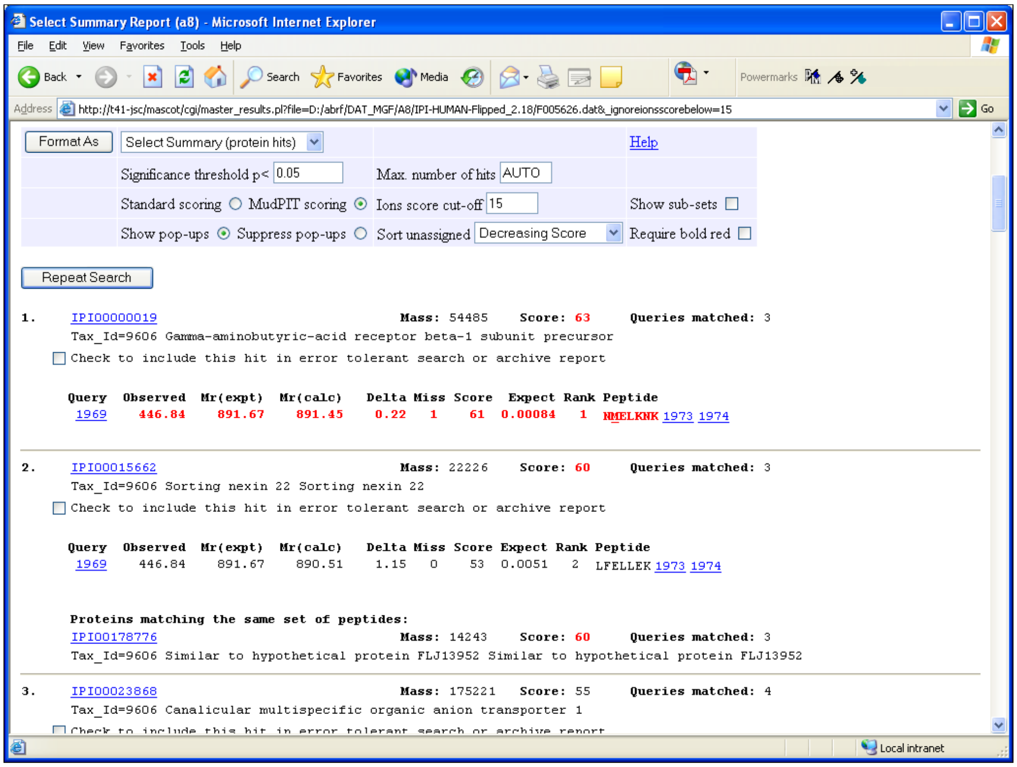
Lets look at some typical numbers for a Mascot search of MudPIT data from a standard ion trap. The significance threshold was the default setting of 5%.

The first column of figures is for a search of IPI human. The third column of figures is for a search of the same data against the reversed database using identical search parameters.

You can see that the false discovery rate for the identity threshold is very conservative. A factor of 10 below the predicted rate. This is often the case for ion trap data, because the mass accuracy and signal to noise are limited.

The false discovery rate for the homology threshold is spot on. By using the homology threshold, we get a large number of additional true positives without exceeding our 5% false discovery rate.

Notice that only some 8% of the spectra give significant matches. This is not unusual. In fact, it is quite good. I would say 5% is average.


Select Summary Report (a8) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Powermarks
Address
Go
Format As
Help
Repeat Search
Local intranet

http://t41-jsc/mascot/cgi/master_results.pl?file=D:/abrf/DAT_MGF/A8/IPI-HUMAN-Flipped_2.18/F005626.dat&_ignoreionsscorebelow=15

Select Summary (protein hits)

Significance threshold p< 0.05 | Max. number of hits AUTO

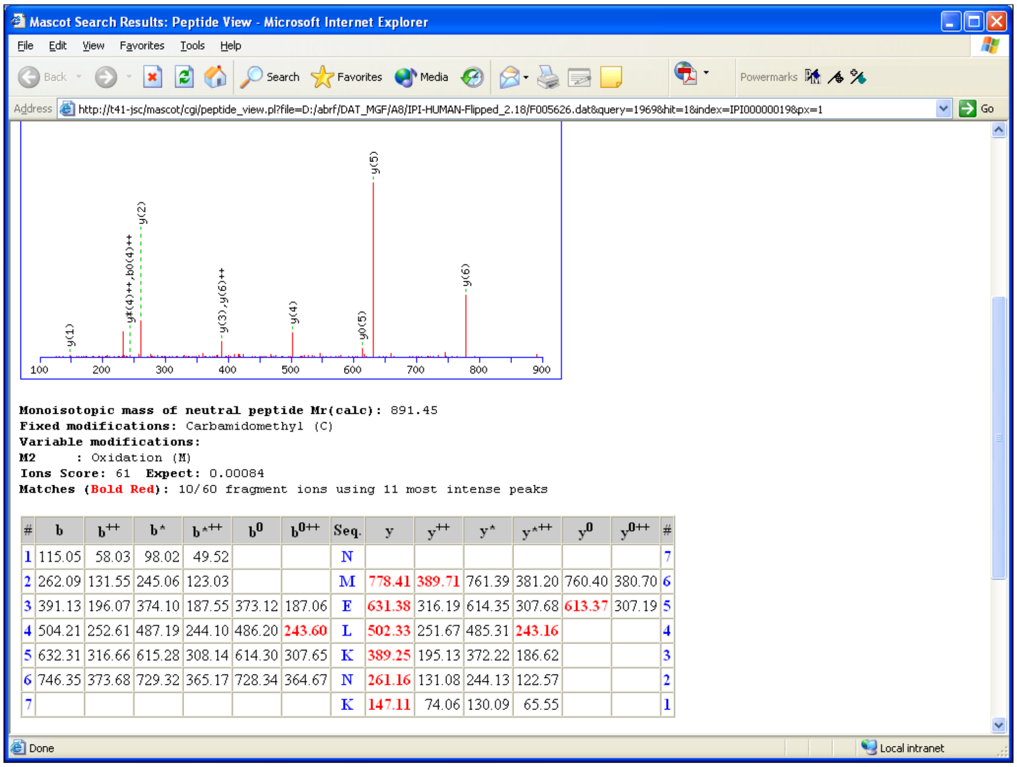
Standard scoring ○ MudPIT scoring ● | Ions score cut-off 15 | Show sub-sets □

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

```
1.     IPI00000019                          Mass: 54485    Score: 63     Queries matched: 3
       Tax_Id=9606 Gamma-aminobutyric-acid receptor beta-1 subunit precursor
     □ Check to include this hit in error tolerant search or archive report

      Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score  Expect Rank Peptide
       1969    446.84    891.67    891.45    0.22    1    61  0.00084    1  NMELKNK 1973 1974

2.     IPI00015662                          Mass: 22226    Score: 60     Queries matched: 3
       Tax_Id=9606 Sorting nexin 22 Sorting nexin 22
     □ Check to include this hit in error tolerant search or archive report

      Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
       1969    446.84    891.67    890.51    1.15    0    53 0.0051    2  LFELLEK 1973 1974


       Proteins matching the same set of peptides:
       IPI00178776                          Mass: 14243    Score: 60     Queries matched: 3
       Tax_Id=9606 Similar to hypothetical protein FLJ13952 Similar to hypothetical protein FLJ13952

3.     IPI00023868                          Mass: 175221   Score: 55     Queries matched: 4
       Tax_Id=9606 Canalicular multispecific organic anion transporter 1
     □ Check to include this hit in error tolerant search or archive report
```

What about those false positives? Let’s have a closer look.

This is the select report for the a8 search of the reversed database. Our highest scoring false positive has a score of 61



http://t41-jsc/mascot/cgi/peptide_view.pl?file=D:/abrf/DAT_MGF/A8/IPI-HUMAN-Flipped_2.18/F005626.dat&query=1969&hit=1&index=IPI00000019&px=1

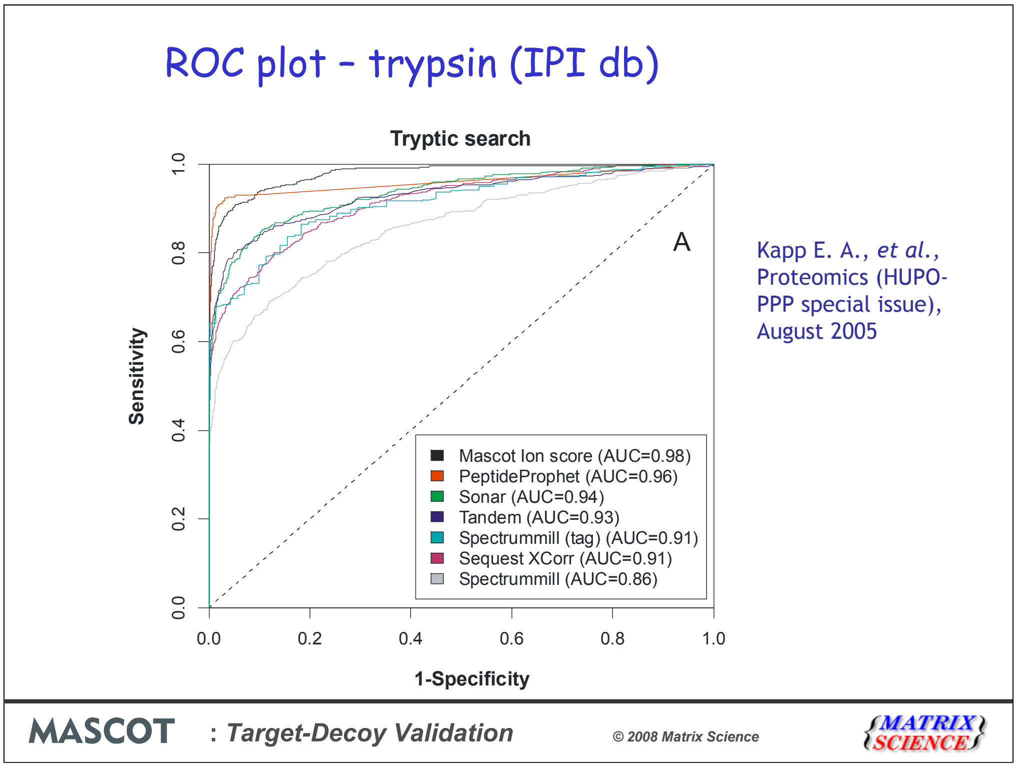
```
Monoisotopic mass of neutral peptide Mr(calc): 891.45
Fixed modifications: Carbamidomethyl (C)
Variable modifications:
M2      : Oxidation (M)
Ions Score: 61  Expect: 0.00084
Matches (Bold Red): 10/60 fragment ions using 11 most intense peaks
```

#	b	b^{++}	b^{*}	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	115.05	58.03	98.02	49.52			N							7
2	262.09	131.55	245.06	123.03			M	**778.41**	**389.71**	761.39	381.20	760.40	380.70	6
3	391.13	196.07	374.10	187.55	373.12	187.06	E	**631.38**	316.19	614.35	307.68	**613.37**	307.19	5
4	504.21	252.61	487.19	244.10	486.20	**243.60**	L	**502.33**	251.67	485.31	**243.16**			4
5	632.31	316.66	615.28	308.14	614.30	307.65	K	**389.25**	195.13	372.22	186.62			3
6	746.35	373.68	729.32	365.17	728.34	364.67	N	**261.16**	131.08	244.13	122.57			2
7							K	**147.11**	74.06	130.09	65.55			1

And this is what it looks like. A near perfect match from the reversed database. Tryptic peptide, complete run of y ions, only one large peak left unmatched.

Asking whether it is correct or wrong becomes almost a philosophical question.

The fact is, when we search large numbers of spectra against large sequence databases, we can get such matches by chance. No amount of expert manual inspection will prevent this. Database matching is a statistical process and, for this search, the number and magnitude of the false positives is within the predicted range.

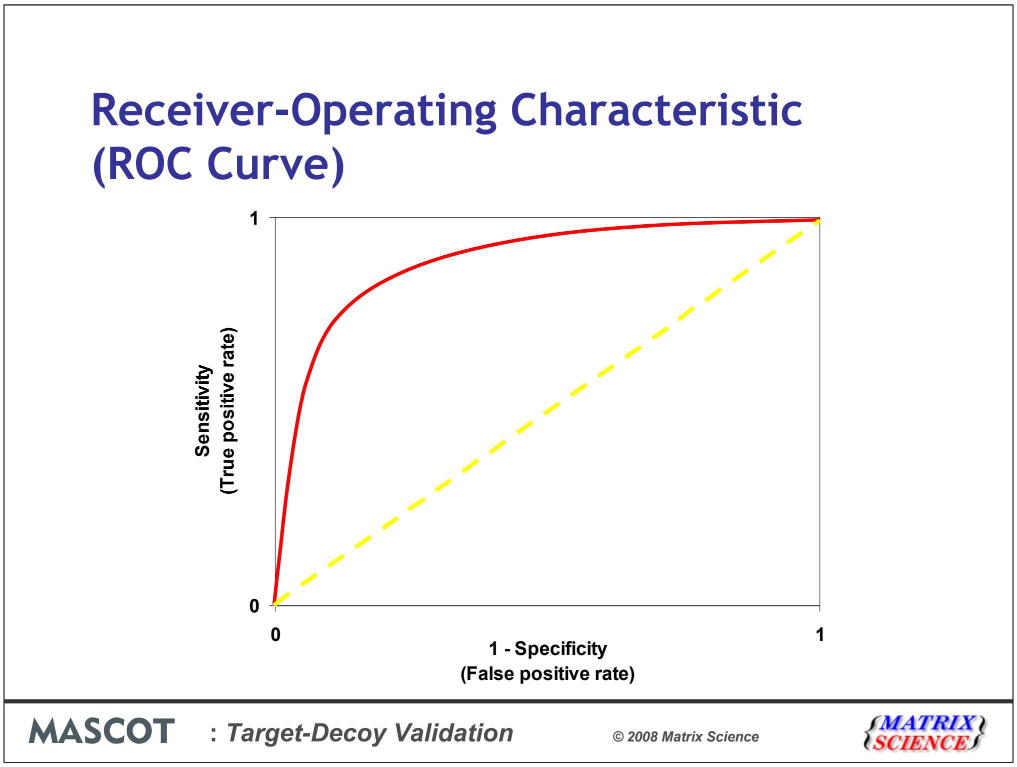


ROC plot - trypsin (IPI db)

Kapp E. A., *et al.*, Proteomics (HUPO-PPP special issue), August 2005

MASCOT : *Target-Decoy Validation* © 2008 Matrix Science MATRIX SCIENCE

The performance of a scoring scheme is sometimes illustrated as a Receiver-Operating Characteristic or ROC Curve. Here is an example from the publication summarising the statistics for the data collected for the HUPO plasma proteome project.

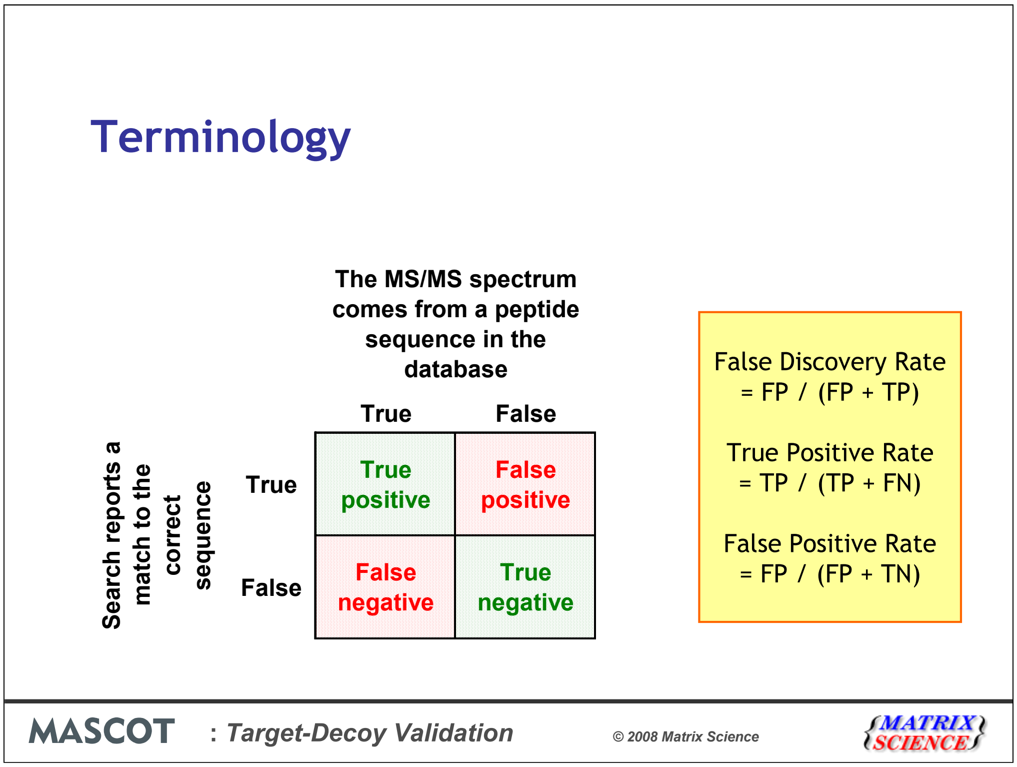


Receiver-Operating Characteristic (ROC Curve)

MASCOT : *Target-Decoy Validation* *© 2008 Matrix Science* MATRIX SCIENCE

A ROC Curve plots true positive rate and false positive rate as a function of a discriminator, such as a score threshold. A good scoring scheme will try to follow the axes, as illustrated by the red curve, pushing its way up into the top left corner. A useless scoring algorithm, that cannot distinguish correct and incorrect matches, would follow the yellow dashed diagonal line.

The origin of the ROC curve has unit specificity, i.e. zero false positives, but also zero true positives. Not a useful place to be. The top right of the ROC curve has unit sensitivity, i.e. 100% true positives, but also 100% false positives, which is equally useless. By setting a significance threshold in Mascot, you effectively choose where you want to be on the curve.



Terminology

		The MS/MS spectrum comes from a peptide sequence in the database	
		True	False
Search reports a match to the correct sequence	True	True positive	False positive
	False	False negative	True negative

False Discovery Rate
= FP / (FP + TP)

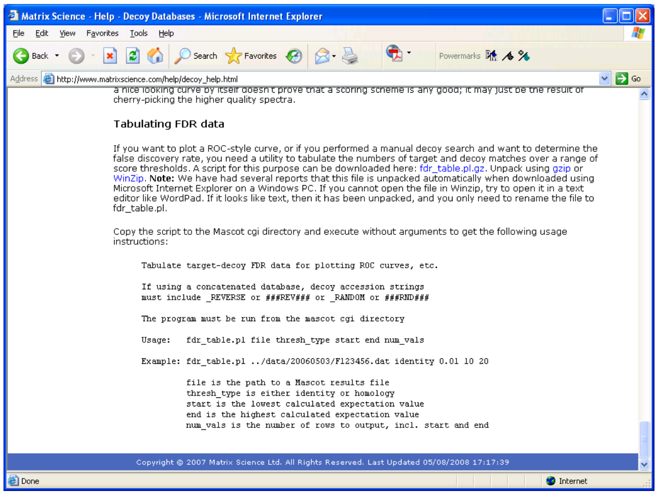
True Positive Rate
= TP / (TP + FN)

False Positive Rate
= FP / (FP + TN)

MASCOT : *Target-Decoy Validation* © 2008 Matrix Science MATRIX SCIENCE

You will remember seeing this slide earlier. To plot an authentic ROC curve, we need estimates of the numbers of true negatives (TN) and false negatives (FN), because true positive rate = TP / (TP + FN) and false positive rate = FP / (FP + TN). However, for real-life datasets, where we are dealing with unknown samples, we do not know TN and FN. So, what is presented as a ROC curve is often just a plot of the fraction of spectra matched in the target database versus the fraction matched in the decoy, or something similar.

Receiver-Operating Characteristic (ROC Curve)


Matrix Science - Help - Decoy Databases - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Powermarks
Address
http://www.matrixscience.com/help/decoy_help.html
Go
Done
Internet

a nice looking curve by itself doesn't prove that a scoring scheme is any good; it may just be the result of cherry-picking the higher quality spectra.

Tabulating FDR data

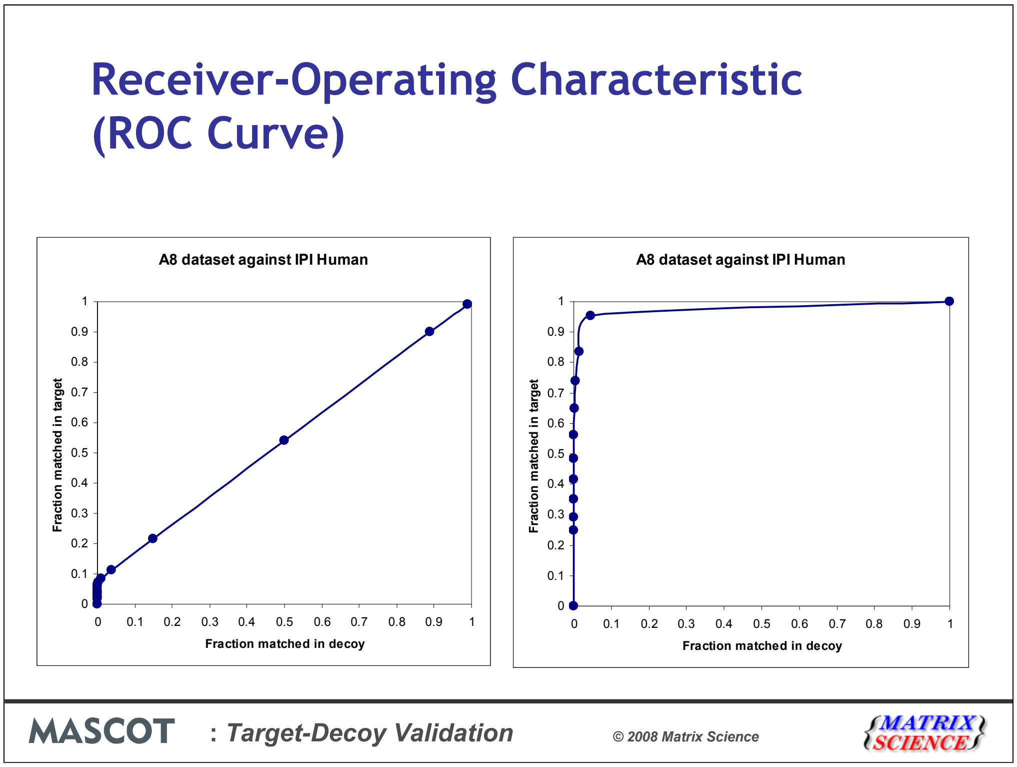
If you want to plot a ROC-style curve, or if you performed a manual decoy search and want to determine the false discovery rate, you need a utility to tabulate the numbers of target and decoy matches over a range of score thresholds. A script for this purpose can be downloaded here: fdr_table.pl.gz. Unpack using gzip or WinZip. **Note:** We have had several reports that this file is unpacked automatically when downloaded using Microsoft Internet Explorer on a Windows PC. If you cannot open the file in Winzip, try to open it in a text editor like WordPad. If it looks like text, then it has been unpacked, and you only need to rename the file to fdr_table.pl.

Copy the script to the Mascot cgi directory and execute without arguments to get the following usage instructions:

```
Tabulate target-decoy FDR data for plotting ROC curves, etc.

If using a concatenated database, decoy accession strings
must include _REVERSE or ###REV### or _RANDOM or ###RND###

The program must be run from the mascot cgi directory

Usage:   fdr_table.pl file thresh_type start end num_vals

Example: fdr_table.pl ../data/20060503/F123456.dat identity 0.01 10 20

         file is the path to a Mascot results file
         thresh_type is either identity or homology
         start is the lowest calculated expectation value
         end is the highest calculated expectation value
         num_vals is the number of rows to output, incl. start and end
```

Copyright © 2007 Matrix Science Ltd. All Rights Reserved. Last Updated 05/08/2008 17:17:39

If you want to plot a ROC-style curve, or if you performed a manual decoy search and want to determine the false discovery rate, you need a utility to tabulate the numbers of target and decoy matches over a range of score thresholds. A script for this purpose can be downloaded from the decoy help page on our web site



Receiver-Operating Characteristic (ROC Curve)

MASCOT : *Target-Decoy Validation* *© 2008 Matrix Science* MATRIX SCIENCE

If we take the MudPIT search I showed you a few slides back, and plot a ROC-style curve for the entire data set, you will get a very poor looking curve, like this one, because no score can discriminate the unmatchable spectra. In other words, as the score threshold is reduced towards zero, additional matches are equally likely to come from the decoy as from the target, and the ROC curve tends towards a diagonal line.

If you exclude the unmatchable spectra, then you can get a nice looking curve from exactly the same set of search results, like the one on the right. However, deciding which spectra to include is arbitrary. I could choose just the 10 strongest spectra, in which case any scoring scheme will give a beautiful ROC curve.

As long as all curves use the same dataset and assumptions, ROC curves are fine for comparisons. But, a nice looking ROC curve by itself doesn't necessarily prove that a scoring scheme is any good.