Maximising Sensitivity with Percolator

MASCOT

MATRIX SCIENCE

Terminology

The MS/MS spectrum comes from a peptide sequence in the database

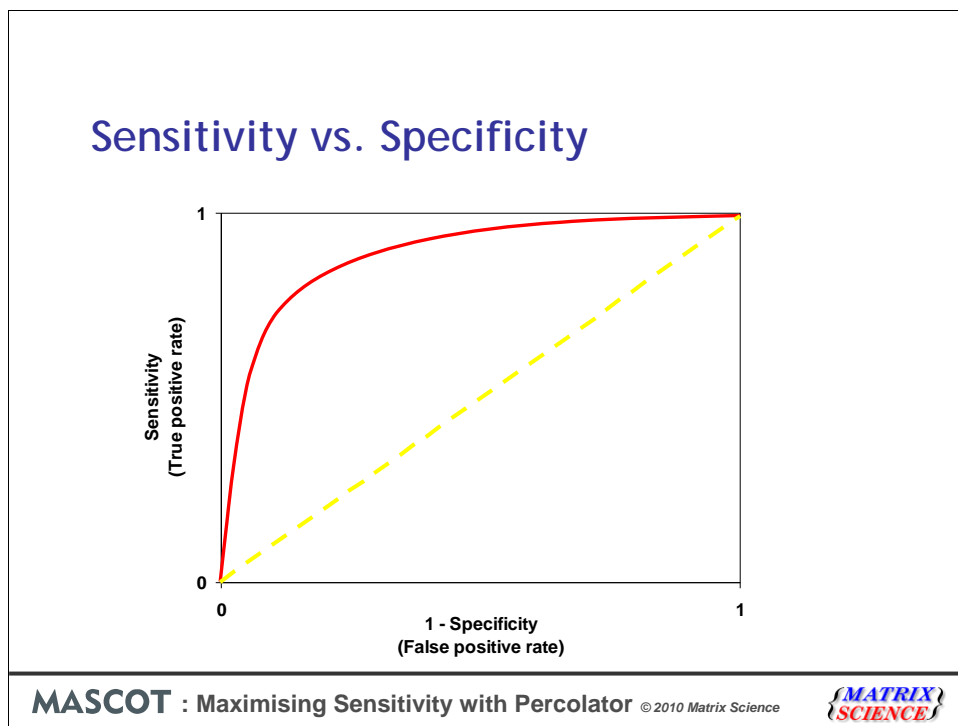|		True	False
Search reports a match to the correct sequence	True	True positive	False positive
	False	False negative	True negative

False Discovery Rate
= FP / (FP + TP)

True Positive Rate
= TP / (TP + FN)

False Positive Rate
= FP / (FP + TN)

MASCOT : Maximising Sensitivity with Percolator *© 2010 Matrix Science*

MATRIX SCIENCE

Database searching is a statistical process. Most MS/MS spectra do not encode the complete peptide sequence; there are gaps and ambiguities. Hopefully, most of the time, we are able to report the correct match, a true positive, but not always. If the sequence of the peptide is not in the database, and we report a match below our score or significance threshold, that's also OK, and we have a true negative. The other two quadrants represent failure. A false positive is when we report a significant match to the wrong sequence. A false negative is when we fail to report a match even though the correct sequence is in the database. For real-life datasets, where we cannot be certain that all the correct sequences are present in the database, we don't know whether a failure to get a match to a spectrum is a TN or a FN. When we do a decoy search, we make an estimate of TP and FP, and report a false discovery rate, which is defined as the count of significant matches in the decoy sequences divided by the total count of significant matches in both target and decoy.

Sensitivity vs. Specificity

MASCOT : **Maximising Sensitivity with Percolator** *© 2010 Matrix Science*

The characteristic attributes of any scoring algorithm are sensitivity and specificity. That is, you want as many correct matches as possible, that and as few incorrect matches as possible. This curve, that illustrates the relationship between sensitivity and specificity, is called a ROC curve, which stands for Receiver Operating Characteristic.

This 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 or a score threshold, you effectively choose where you want to be on the curve.

Sensitivity vs. Specificity

MASCOT : **Maximising Sensitivity with Percolator** *© 2010 Matrix Science*
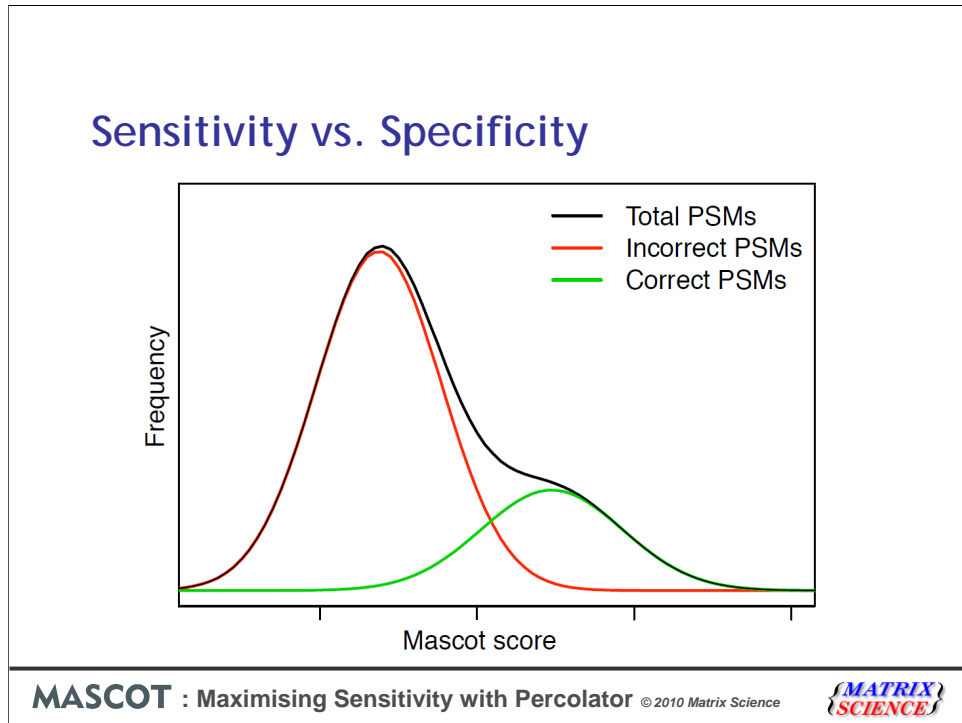
This is another way to look at it. Even the best scoring scheme cannot fully separate the correct and incorrect matches, as shown here in a schematic way. The score distribution for the correct matches, in green, overlaps that of the incorrect matches, in red. The observed score distribution is the sum of these two curves, in black
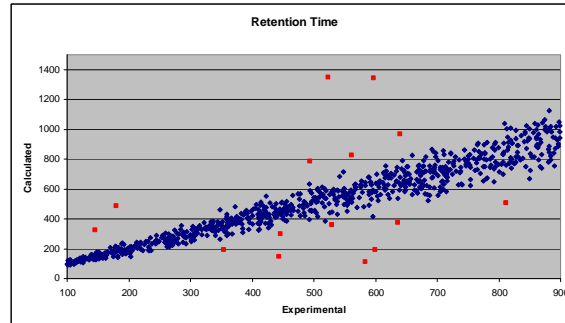
When we set a score threshold, we are trying to separate the green and red curves as cleanly as possible. But, the lower the threshold, the more incorrect matches are reported. The higher the threshold, the fewer correct matches

But, what if we could find ways to pull these two distributions further apart, or make the distributions narrower? In other words, better resolve the two distributions. This would allow us to improve the sensitivity for a given false discovery rate.

This is perfectly possible. There are many observables that the Mascot scoring algorithm doesn't include. For example, HPLC retention time. If the experimental retention times are generally close to the calculated values, we might suspect outliers are false positive matches

Sensitivity vs. Specificity

Mascot scoring ignores
• Systematic mass errors

High scoring match

Fragment mass tolerance : ± 0.6 Da

Low scoring match
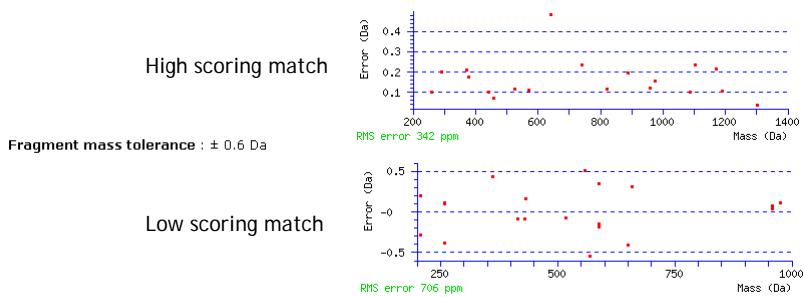
MASCOT : Maximising Sensitivity with Percolator © 2010 Matrix Science

The more accurate the mass values, the tighter the mass tolerance can be in a Mascot search. But, Mascot only cares about whether the mass values fall within the specified window. In this example, we are searching trap data with a tolerance of +/- 0.6 Da. When we look at a strong match, the scatter of fragment mass values appears to be much tighter, maybe +/- 0.1 Da, assuming the single high value is random match. When we look at a low scoring, random match, the errors are uniformly scattered across the tolerance window. So, if we had a match that was close to threshold, the scatter on the fragment mass values would be an indication as to whether it was a correct match or not.

Sensitivity vs. Specificity

Mascot scoring ignores

- Counts of modifications

1.   CH60_CRIGR   **Mass:** 76623   **Score:** 404   **Matches:** 8(5)   **Sequences:** 7(5)
    60 kDa heat shock protein, mitochondrial OS=Cricetulus griseus GN=HSPD1 PE=2 SV=1
    ☐ Check to include this hit in error tolerant search or archive report

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
☑	11	417.1822	832.3498	832.3828	-0.0329	0	45	0.04	1	U	K.APGFGDNR.K
☑	15	451.2499	900.4853	900.5280	-0.0428	0	52	0.018	1	U	K.LSDGVAVLK.V
☑	16	456.7806	911.5467	911.5804	-0.0337	0	59	0.0011	1	U	K.VGLQVVAVK.A
☑	21	480.7447	959.4748	959.5036	-0.0288	0	45	0.06	1	U	R.VTDALNATR.A
☑	24	595.7855	1189.5565	1189.6013	-0.0448	0	57	0.0036	1	U	K.DIGNIISDAMK.K + Methyl (DE)
☑	51	1057.0537	2112.0929	2112.1323	-0.0394	0	116	4.4e-009	1	U	R.ALMLQGVDLLADAVAVTMGPK.G
☑	61	828.1238	2481.3495	2481.3942	-0.0447	0	(12)	1e+002	1	U	R.TALLDAAGVASLLTTAEAVVTEIPK.E + 2 Methyl (DE)
☑	62	828.1322	2481.3748	2481.3942	-0.0194	0	31	1.2	1	U	R.TALLDAAGVASLLTTAEAVVTEIPK.E + 2 Methyl (DE)

MASCOT : Maximising Sensitivity with Percolator *© 2010 Matrix Science*   MATRIX SCIENCE

Here are some results from a search with 3 variable modifications. If we look at the confident matches, most peptides are unmodified. One carries a single modification and a long peptide carries the same modification at two locations.

Sensitivity vs. Specificity

Mascot scoring ignores

- Counts of modifications

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
☑	52	1065.0399	2128.0653	2128.1365	-0.0712	1	3	1.6e+003	1		TSVVQGIHTDSNKLVLQAR
☑	45	640.3281	1917.9625	1917.9624	0.0001	1	12	2.3e+002	1		EQNDRVLELVSQLLR
☑	49	1020.9879	2039.9613	2039.9220	0.0392	0	4	1.4e+003	1		EEEDDLLNELDEMEK + 8 Methyl (DE); Carbamidomethyl (N-term)
☑	39	752.8643	1503.7141	1503.7715	-0.0574	1	18	79	1		ESRDVELLCNTK + 3 Methyl (DE); Carbamidomethyl (N-term)
☑	34	714.8884	1427.7623	1427.7619	0.0004	0	20	46	1		EEQEAAISALAAK + 3 Methyl (DE); Carbamidomethyl (N-term)
☑	40	760.8461	1519.6777	1519.7646	-0.0869	1	18	69	1		EKYTLDIFTER + 2 Methyl (DE); Carbamidomethyl (N-term)
☑	60	789.1094	2364.3063	2364.2916	0.0147	1	31	1.8	1		LGWTVTDALESIKETILEVK + 3 Methyl (DE); Carbamidomethyl (N-term)

MASCOT : Maximising Sensitivity with Percolator *© 2010 Matrix Science*

MATRIX SCIENCE

Now look down at the low scoring, random matches on the unassigned list. Some are unmodified, of course, but others are heavily modified. One has 8 methyls plus another modification at the terminus. This is to be expected. Peptides that have a large number of potential modification sites support many possible arrangements and permutations of modifications, some of which match quite well by chance. In other words, there are more degrees of freedom. So, if two matches had the same score, and both had 8 Ds and Es, but one was unmodified and the other had 4 methylations, we might feel greater confidence in the match to the unmodified peptide.

Sensitivity vs. Specificity

Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search

Andrew Keller,*,† Alexey I. Nesvizhskii,*,† Eugene Kolker, and Ruedi Aebersold

Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103

Peptide Prophet

- Expectation maximization
- No-enzyme search
- Positive training set: fully tryptic matches
- Negative training set: non-specific matches

MASCOT : Maximising Sensitivity with Percolator *© 2010 Matrix Science* **MATRIX SCIENCE**


The common factor in these properties is that you have to learn how to use them by looking at a set of results of reasonable size, because the rules are likely to change from search to search. Using a count of modifications might not be such a good idea if you were analysing highly modified histones.

The pioneer of using machine learning on a collection of characteristics was Peptide Prophet from the Institute for Systems Biology. This was, and still is, popular for transforming Sequest scores into probabilities.

It takes information about the matches in addition to the score, and uses an algorithm called expectation maximization to learn what distinguishes correct from incorrect matches. Originally, a widely used approach was to run the Sequest search without enzyme specificity and then assume that matches to fully tryptic peptides were correct and matches to non-specific peptides were incorrect.

Sensitivity vs. Specificity
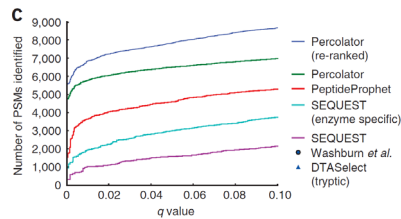
Semi-supervised learning for peptide identification from shotgun proteomics datasets

Lukas Käll[1], Jesse D Canterbury[1], Jason Weston[2], William Stafford Noble[1,3] & Michael J MacCoss[1]
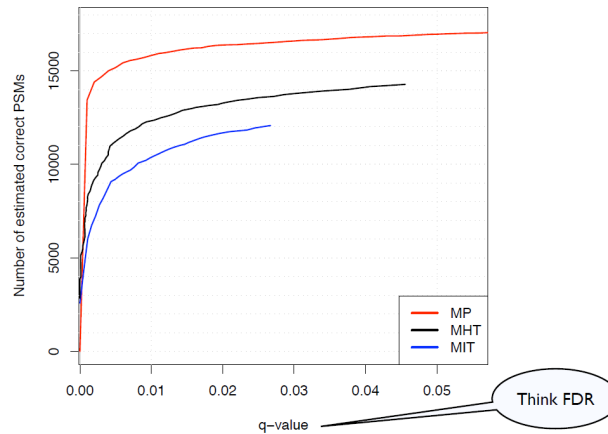
Percolator

- Support vector machine
- Target decoy search
- Positive training set: high scoring matches from target
- Negative training set: matches from decoy

MASCOT : **Maximising Sensitivity with Percolator** *© 2010 Matrix Science*   *MATRIX SCIENCE*

A more recent development has been to use the matches from a decoy database as negative examples for the classifier. Percolator trains a machine learning algorithm called a support vector machine to discriminate between a sub-set of the high-scoring matches from the target database, assumed correct, and the matches from the decoy database, assumed incorrect. Percolator was developed by the MacCoss group at U. Washington. Lukas Kall is now in Sweden, at the University of Stockholm.

Sensitivity optimisation

M. Brosch, L. Yu, T. Hubbard, J. Choudhary, *J Proteome Res* (2009).

MASCOT : **Maximising Sensitivity with Percolator** *© 2010 Matrix Science*

This can give very substantial improvements in sensitivity. The original Percolator was implemented mainly with Sequest in mind, but Markus Brosch at the Sanger Centre wrote a wrapper that allowed it to be used with Mascot results and published results such as this. The black trace is the sensitivity using the Mascot homology threshold (MHT) and the red trace is the sensitivity after processing through Percolator (MP). It doesn't work for every single data set. But, when it does work, the improvements can be most impressive.

Those of you who attended this meeting last year will remember that Markus gave a presentation on this topic

(PSM = peptide sequence match, MIT = Mascot identity threshold)

Percolator

MASCOT : **Maximising Sensitivity with Percolator** *© 2010 Matrix Science*

Using a decoy database is particularly convenient with Mascot, because it can be done automatically as part of any search

Sensitivity optimisation

MASCOT : **Maximising Sensitivity with Percolator** © *2010 Matrix Science*

The developers of Percolator have kindly agreed to allow us to distribute and install Percolator as part of Mascot 2.3. This option is available for any search that has at least 100 MS/MS spectra and auto-decoy results, but it works best if there are several thousand spectra. To switch to Percolator scores, just check the box and then choose Filter. This is the example search that is linked from the MS/MS Summary report help page

Sensitivity optimisation

▼ *Decoy search summary*

Peptide matches	in target	in Decoy	FDR
– above identity threshold	1484	15	1.01%
– above identity or homology threshold	1837	20	1.09%

Decoy results are available in ☑the decoy report.

▼ *Decoy search summary*

Peptide matches	in target	in Decoy	FDR
– above identity threshold	1985	19	0.96%
– above identity or homology threshold	1985	19	0.96%

Decoy results are available in ☑the decoy report.

Delta M Score	Expect	Rank	U 1 2 3	Peptide
0.1363 0 33	0.00049	▶1	U ■ ■ ■	R.LIGDAAK.N
0.3020 0 14	0.039	▶1	U ■	K.VQVEYK.G
0.2841 0 17	0.018	▶	U ▬	R.GRLRDIEK.A
0.4581 0 18	0.015	▶1		Score > 13 indicates **identity**
0.0517 0 21	0.0087	▶1	U ■	K.VLEDSDLK.K
0.2227 0 25	0.0031	▶1	U ■	K.ITITNDQNR.L

MASCOT : Maximising Sensitivity with Percolator *© 2010 Matrix Science*   {MATRIX SCIENCE}

Using the Mascot homology threshold for a 1% false discovery rate, there are 1837 peptide matches. Re-scoring with Percolator gives a useful increase to 1985 matches.

Note that, in general, the scores are lower after switching to Percolator. The value in the expect column is the Posterior Error Probability (PEP) output by Percolator. A Mascot score is calculated from this and there is a single score threshold, which we will continue to call the identity threshold, with a fixed value of 13 (-10 log 0.05). By keeping the score, threshold, and expect value consistent, we aim not to break any third party software that expects to find these values.
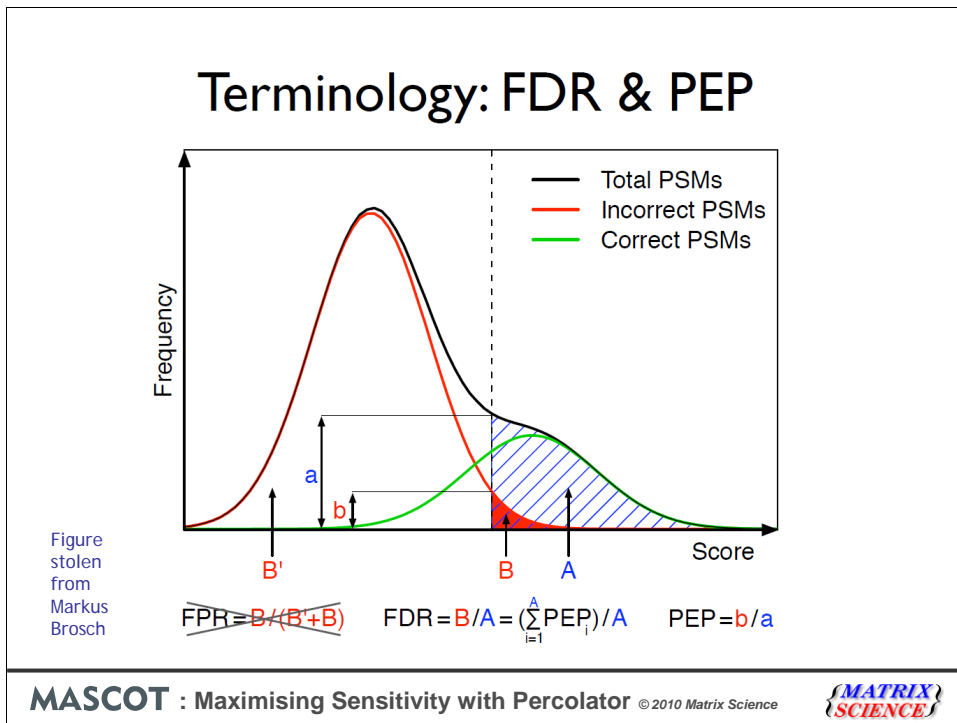
I've stolen this slide from the talk Markus gave last year because it makes the difference between FDR and PEP very clear. The vertical dashed line is our significance threshold, chosen to give an acceptable false discovery rate (FDR or q value). This is the ratio of the areas under the black and red curves, B/A. That is, it is a property of the set of matches, not of an individual match.

For any particular match, the chance of it being incorrect, given its score, is the Posterior Error Probability (PEP). This corresponds to the ratio of the heights b/a, although we cannot measure a and b directly.

Returning to the previous slide. After Percolator processing, the count of all matches with a q value equal to or less than the significance threshold gives us our false discovery rate. This is a population of matches, some of which, individually, will have greater or lesser chances of being incorrect. The measure for individual matches is the Percolator PEP value, which is tabulated in the expect column. The PEP is converted to a score by setting a fixed threshold score of 13.

The Mechanics

- **All binaries installed as part of Mascot 2.3**
- **Currently shipping Percolator 1.14**
- **After any suitable search:**
  1. ms-createpip.exe runs, reading the result file and creating a Percolator input file (*.pip) containing a list of features for every query
  2. Percolator runs, taking input from the *.pip file and writing output to two output files (*.target.pop, *.decoy.pop)
  3. When a report is generated, Mascot Parser transparently opens the *.pop file as required
  4. If you view a report from an old result file that is suitable for Percolator, the report script automatically triggers the creation of *.pip and *.pop files

MASCOT : **Maximising Sensitivity with Percolator** © *2010 Matrix Science*   *MATRIX SCIENCE*

The architecture of the integration between Mascot and Percolator.

Features are the observables, e.g. retention time, mass error, count of modifications or missed cleavages, etc.

The Mechanics

```
Command Prompt

C:\Inetpub\mascot\bin>ms-createpip.exe --help
Version 2.3.0

Usage: ms-createpip.exe [OPTION] -i filename

  Options:
    -h, --help         display this help page and exit
    -f, --features     display list of features defined in mascot.dat and exit
    --sessionID <id>   not normally used because this is run from command line
    -o output_file     default is filename.pip
    -q <#queries>      override minimum number of queries set in mascot.dat
    -s <#sequences>    override minimum number of sequences set in mascot.dat
    -a <feature>       add a feature to the list specified in mascot.dat
    -r <feature>       remove a feature to the list specified in mascot.dat
    -c                 use cached results
    -p <interval>      progress reports of Process:X% every <interval> seconds
    --nocache          do not use cached results
    --version          display version number and exit


Features
    retentionTime   Retention time in seconds if available
             dM     Calculated minus observed peptide mass in Da
         mScore     Mascot score (always on)
        lgDScore    Mascot score minus Mascot score of next best non-isobaric
                    peptide hit
         mrCalc     Calculated Mr
         charge     Charge
          dMppm     Calculated minus observed peptide mass in ppm
          absDM     Absolute value of calculated minus observed peptide mass in
                    Da
        absDMppm    Absolute value of calculated minus observed peptide mass in
                    ppm
          isoDM     Calculated minus observed peptide mass, after eliminating
                    possible isotope errors up to 2 Da, in Da
        isoDMppm    Calculated minus observed peptide mass, after eliminating
                    possible isotope errors up to 2 Da, in ppm
             mc     Number of missed cleavages (always 0 if no enzyme)
         varmods    Number of modified sites divided by number of modifiable
                    sites
```

MASCOT : **Maximising Sensitivity with Percolator** *© 2010 Matrix Science*   {MATRIX SCIENCE}

Configuration information is in mascot.dat. This controls which features are used , paths to executables and other files, logging levels, etc. There is some documention in the Mascot Setup & Installation manual. You can also get help by executing ms-createpip.exe and percolator.exe with the argument --help

The Mechanics

Creating the input file can be time consuming for a large result file, but is a one-time operation

Defaults are set in mascot.dat

- Whether to show Mascot scores or Percolator scores when report first loaded
- Whether to use retention time information if available
- Which features to include

MASCOT : Maximising Sensitivity with Percolator *© 2010 Matrix Science* {MATRIX SCIENCE}

Some miscellaneous points

Limitations

'Protein Features' carry some risk and are currently not implemented (Mascot 2.3.00)

- Feature is essentially a count of the number of sequences assigned to the parent protein, normalised to the length of the protein. 'To those that have, shall be given'
- Concern 1: There is no analogy of this grouping in the decoy database
- Concern 2: FDR is no longer a true peptide FDR and could be misinterpreted

Only the top ranking match is re-scored

- Never get re-ranking of peptide matches. Scores and expect values for other ranks are pro-rated

Unlikely to succeed if results contain very few good matches

MASCOT : Maximising Sensitivity with Percolator © *2010 Matrix Science*

We decided not to implement 'protein features' because of concerns that the results could be misleading. Essentially, there is only one protein feature: a count of the number of sequences assigned to the parent protein, normalised to the length of the protein. In biblical terms, 'To those that have, shall be given'.

There are some complications to this. For example, many peptides are found in multiple proteins, so which is the 'true' parent? The longest or the shortest or some average. Normalisation is critical if we want to avoid the 'titin effect', where the very largest proteins are promoted because they randomly match a huge number of peptides.

Another concern was that we may get artefacts because the whole concept of target-decoy validation is peptide-centric. Each peptide sequence match being independent of any other. If you increase the score of a weak match simply because it is found in a protein for which there is strong evidence, the FDR cannot be compared with a conventional, 'pure peptide' FDR

Only the top ranking match to each spectrum is used by Percolator. We tried to include all the significant matches, but couldn't get the stats to work properly. This is something Lukas and colleagues are working on, because there would be a real benefit from allowing Percolator to re-rank matches. For example, the features associated with the rank 1 match might indicate that it is unsafe and should be given a high PEP while the rank 2 match looks great and would get a very low PEP. At present, this change in order cannot happen. If the rank 1 match is given a high PEP then the rank 2 match can only be higher

Finally, you must have a population of good, strong matches to provide a positive training set for the SVM. The larger the data set, the more matches you need.

Limitations

	Filter	Siqnificance threshold p< 0.02	Max. number of families AUTO	[help]
		Ions score or expect cut-off 0		
		Show Percolator scores ☐ [help]		

▼*Decoy search summary*

Peptide matches	in SwissProt	in Decoy	FDR
– above identity threshold	13	0	0.00%
– above identity or homology threshold	45	2	4.44%

Decoy results are available in [the decoy report](#).

	Filter	Siqnificance threshold p< 0.02	Max. number of families AUTO	[help]
		Ions score or expect cut-off 0		
		Show Percolator scores ☑ [help]		

▼*Decoy search summary*

Peptide matches	in SwissProt	in Decoy	FDR
– above identity threshold	25	2	8.00%
– above identity or homology threshold	25	2	8.00%

Decoy results are available in [the decoy report](#).

MASCOT : **Maximising Sensitivity with Percolator** *© 2010 Matrix Science*

So, for example, if we take the famous T. Rex dataset, where there are only a tiny number of high confidence matches in 48,216 spectra, we don't see any sensitivity improvement. There simply aren't enough good matches for the SVM get traction. But, this is the exception. For a more typical search result, Percolator will give sensitivity a significant boost

To use retention time as a feature, the experimental RT values must be present in the MGF peak list. Some peak picking utilities simply embed the RT and scan information as free text in the scan title, which won't work. Percolator fits calculated values to the experimental retention times and then uses the deviations for individual matches as a predictive feature. This increases processing time for Percolator, so it is turned off by default. You can enable it as a global default in mascot.dat, or use a URL argument to enable it for an individual search

Retention Time

Original Mascot results

▼*Decoy search summary*

Peptide matches	in SwissProt	in Decoy	FDR
– above identity threshold	6198	61	0.98%
– above identity or homology threshold	6198	61	0.98%

Decoy results are available in the decoy report.

After Percolator, no RT

▼*Decoy search summary*

Peptide matches	in SwissProt	in Decoy	FDR
– above identity threshold	7207	74	1.03%
– above identity or homology threshold	7207	74	1.03%

Decoy results are available in the decoy report.

After Percolator, with RT

▼*Decoy search summary*

Peptide matches	in SwissProt	in Decoy	FDR
– above identity threshold	7473	78	1.04%
– above identity or homology threshold	7473	78	1.04%

Decoy results are available in the decoy report.

MASCOT : Maximising Sensitivity with Percolator © 2010 Matrix Science

MATRIX SCIENCE

Here is an example where enabling retention time as a feature gives a further useful improvement in sensitivity