

# False Discovery Rate

## ProteoRed Multicentre Study - 6

### Terminology

FDR: False Discovery Rate

PSM: Peptide-Spectrum Match (*a.k.a.*, a hit)

PIT: Percentage of Incorrect Targets

In the past years, a number of statistical tools and strategies have been developed to assess the error rate of peptide-spectrum matches (e.g., FDR calculations, Percolator, PeptideProphet, FDRAnalysis). In particular, the use of decoy databases is a simple yet powerful approach to estimate the false discovery rate (FDR) at the peptide identification level. In principle, FDR estimations should allow us to make a direct comparison of peptide identification results from different laboratories in this study. Despite its widespread use in proteomics experiments, PSM FDR has not been fully standardized yet. In what follows I will discuss the different strategies to calculate FDR and scoring systems commonly employed. I also provide some basic information of FDR calculation for newbies.

### Decoy sequence construction

The decoy strategy starts with a null hypothesis. The *decoy database* serves this purpose; if we search a set of spectra against the decoy database, we can be quite sure that the resulting PSMs are incorrect. A decoy database is a database of amino acid sequences that is derived from the original protein database (called the *target database*) by reversing the target sequences, shuffling the target sequences, or generating the decoy sequences at random using a Markov model with parameters derived from the target sequences.

In this study, we adopt the reversing strategy since reversing a given database sequence will always generate the same reversed sequence result, whereas shuffling/randomization generates sequences randomly resulting in different sequences every time you run the script/software.

Notably, if you use the automated 'Decoy' option from Mascot, you are in fact using a randomized strategy. Read the 'Mascot' section below for more information on Mascot FDR calculation.

### Measuring FDR

There are a number of approaches reported in the literature to calculate FDR and no clear consensus as to which method is best. Target-decoy searching is commonly performed using **concatenated databases**. In other words, the decoy database is combined with the target database and searched together. For each spectrum, the search engine must then choose between target and decoy sequences. It can be seen as a competition between the two databases for each spectrum. In this strategy, **the number of false positive peptide identifications is calculated for a given threshold by doubling the number of hits to the decoy database.**

Let's take a practical example.

Say that at a mascot identity score threshold of 27, we observe 1000 PSMs, and since we tagged all decoy sequences when the target-decoy database was built we are able to count 25 decoy PSMs among the 1000 PSMs total.

**FDR = 2x Decoy hits (FP) / All PSMs above threshold (FP + TP)**

$$\text{FDR} = 50/1000 = 5\%$$

Although searching a concatenated database is the most common strategy, some researchers advocate **searching target sequences separately from decoy sequences**. In this case, FDR is calculated by taking the ratio between the number of *decoy PSMs* above the threshold and the number of *target PSMs* above the threshold. In this case, using the same example previously mentioned, the FDR is calculated as follows:

**FDR = Decoy hits (FP) / target hits (FP + TP)**

$$\text{FDR} = 25/975 = 2.56\%$$

In the past years, slightly more sophisticated methods have been developed to calculate FDR:

- Percentage of Incorrect Target (PIT) by Kall et al.
- 'Refined' FDR method by Pedro Navarro and Jesus Vazquez.
- Percolator and MascotPercolator
- PeptideProphet (it can be thought as a *local FDR*)

If you decide to go with one of these strategies, please provide description of your parameter choices.

## Protein FDR

Often FDR is only calculated at the peptide level (PSM FDR). Deriving FDR for protein identification is not trivial. Because protein identifications are defined by assemblies of PSMs, errors determined at the PSM level propagate to the protein identification level. Therefore it comes with no surprise that in practice the protein identification FDR is often larger than the PSM FDR.

**We do not require protein FDR in this study.** Nevertheless, if you want to calculate it, feel free to do so. I advice you to take a look at MAYU software which is incorporated in the TransProteomic Pipeline (TPP). MAYU can also be downloaded as a standalone perl script.

## Calculating FDR (on the fly) with Mascot

Starting with version 2.2, Mascot offers an **automated 'decoy' search** option. To use this option you should install the target database (PME6.fasta) since the decoy search is performed on the fly. In other words, during the search, Mascot generates and tests a random version of each target database protein.

As far as I know Mascot is not straightforward when it comes to **setting a decoy threshold** (say FDR of 5%). Mascot will in fact compute whatever FDR for a given significance threshold (usually 0.05). For example, you may have a significance threshold of 0.05 and a FDR of 6.77%. In this case, you will have to tweak the significance threshold (to 0.0345, bottom figure) until your FDR is acceptable (I picked 5% here).

Filter Significance threshold p< 0.05 Max. number of families 20 [help]  
Ions score or expect cut-off 0  
Show Percolator scores  [help]

▼Decoy search summary

Peptide matches	in SProt	in Decoy	FDR
- above identity threshold	4683	317	6.77%
- above identity or homology threshold	5305	460	8.67%

Filter Significance threshold p< 0.0345 Max. number of families 20 [help]  
Ions score or expect cut-off 0  
Show Percolator scores  [help]

▼Decoy search summary

Peptide matches	in SProt	in Decoy	FDR
- above identity threshold	4428	223	5.04%
- above identity or homology threshold	4974	301	6.05%

If you prefer to perform a manual decoy search, then you should use the provided PME6\_decoy.fasta database. After searching against this database (do not check the 'decoy' box), you need to manually calculate the FDR. You may do this in a number of ways. One option is to use simple scripts to count decoy and target hits, then apply the appropriate FDR formula. The easiest way would be to use FDR\_table.pl (a perl script from MatrixScience) to tabulate the FDR values together with number of target hits and the associated threshold. You can find the script and usage instruction here:

[http://www.matrixscience.com/help/decoy\\_help.html](http://www.matrixscience.com/help/decoy_help.html)

***One of the most important things when reporting your FDR will be your choice of 'scoring threshold system'.*** When you perform a decoy search, Mascot will report two FDR values, one for '***identity threshold***' and the other for '***homology or identity threshold***'. Depending on the 'quality' of your data (spectra), the FDR calculated for these two threshold types may differ considerably. Although you may choose whatever satisfies you, ***we strongly 'recommend' you to report the 'identity threshold'.*** This request is for comparison purposes only since the 'identity threshold' may be also used with ProteomeDiscoverer's Mascot searched data. Don't forget to indicate your choice when reporting us.

### ***Mascot Percolator***

If you have Mascot v2.3 (or want to play with MascotPercolator program), you also can use 'Show Percolator Scores'. Percolator is a semi-supervised machine learning algorithm that improves the discrimination between correct and incorrect spectrum identifications. Percolator calculates a ***Posterior Error Probability*** (PEP) for each PSM. The original Mascot score will be replaced such that the score is equal to  $-10\log(\text{PEP})$ ; therefore, threshold greater or equal to 13 translate to an error rate of 5%.

<http://www.sanger.ac.uk/resources/software/mascotpercolator/>

[http://www.matrixscience.com/help/percolator\\_help.html](http://www.matrixscience.com/help/percolator_help.html)

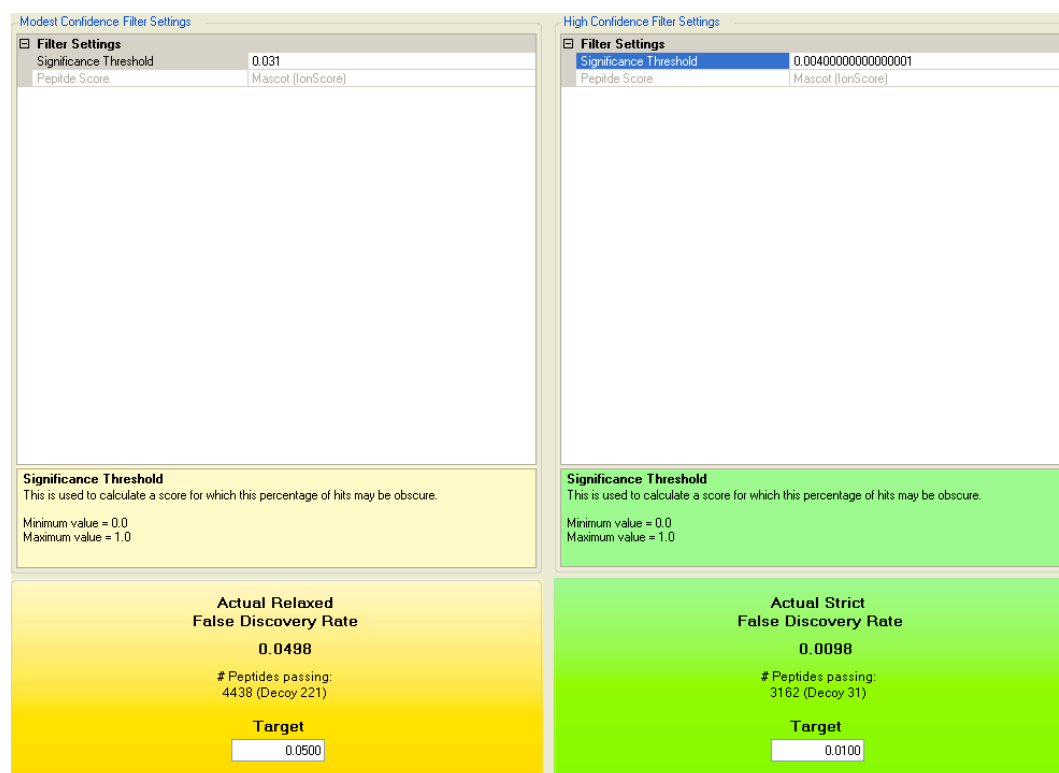
### ***Proteome Discoverer***

You can use Proteome Discoverer to search Sequest and/or Mascot applying the FDR approach (Mascot v2.2 or higher). If you set a FDR target value for a decoy database search, the application determines

and applies filter thresholds to the PSMs so that the resulting FDR is not higher than the set target value.

If you use Mascot as the search engine in ProteomeDiscoverer you should note that ***the default peptide score used to calculate the FDR is the Mascot Significance Threshold***. Make sure you have this option selected (figure below). As for Sequest, you should also stick with the default option; that is, the ***XCor versus charge state***.

In ProteomeDiscoverer, you must specify two target values for a decoy database search: a strict target FDR and a more relaxed FDR. The figure below shows the decoy search setting with target false discovery rates of one percent and five percent, respectively. After completing the search, the system automatically determines two sets of filter settings.



### ***FDRAnalysis (free software) for FDR calculation and filtering***

[http://www.ispider.manchester.ac.uk/FDRAnalysis/FDR\\_analysis\\_home.html](http://www.ispider.manchester.ac.uk/FDRAnalysis/FDR_analysis_home.html)

Recently, Andrew Jones et al developed this web-based application to analyze MS/MS identification data from different search engines (Mascot, X!Tandem, and OMSSA). One can mainly use the application for two purpose: (1) to combine results of different search engines

and (2) to calculate FDR from a given database search result. Two notes of caution about this application:

1) This algorithm uses the FDR calculation proposed by Kall et al which adjust FDR with PIT.

2) You have to search the concatenated target-decoy database. The reversed protein sequences in this database were tagged with the suffix "rev\_" (without quotation marks) in the "Tag Used in Decoy Search" field.

If you use this application, you need to input a "FDR Score Threshold" (e.g., 0.05). Finally, you can export the results to a table containing all peptides (and other information) with FDR below 0.05. In fact, **FDRAnalysis** will calculate a new score named 'FDR score' (a q-value-like score) that one can think about as *local FDR*. In other words, you will have the probability calculated for each peptide as opposed to *global FDR* calculations.

**If you have any question or do not agree with what is in this document, contact Alex Campos  
acampos@pcb.ub.es**