# Mass spectrometry and tandem MS/MS

### **Calculating False Discovery Rates (FDR)**

Many search engine attempts to provide an estimate of how likely a particular protein ID is to be the result of random matching, rather than a "true" ID. In the search engines that we primarily use (Mascot and ProteinPilot) the statistics on this are calculated by the strength of the match of a peptide mass fingerprint (Mascot score) or ms/ms spectrum match (both Mascot and ProteinPilot (Paragon Algorithm)), with the basic principle that the better the "score", the less likely the ID is to have arisen by a random match. In Mascot and other "probabilistic" search engines, the correlation between the Mascot score and the actual p-score is related to the size of the database searched (the number of possible sequence matches in the "Search Space", which is grows larger both as the number of database entries increases (e.g., searching against all Mammalian sequences vs. searching against only Mouse sequences), and as one increases the number of possible (variable) modifications considered (if each of 1000 tryptic fragments in a database can either HAVE or NOT HAVE a particular modification, then there are 2000 masses in the database rather than 1000). The larger the database searched, the stronger/higher the score for a match has to be in order to be considered "significant", i.e., not arising by chance. At a practical level, this means that no more than 2-4 variable modifications can be searched as possibilities in Mascot searches - attempting to include more possibilities makes it increasingly difficult to get significant matches for the peptides (modified or unmodified) which are actually in the sample.

Any search algorithm which assigns statistical values to the matches it identifies (PMF or MS/MS spectra to proteins or peptides) attempts essentially to calculate the probability of a random match, but no algorithm does this perfectly; therefore, most lists of ID'd proteins contain MORE false positives than the 5% one would expect by using a p<0.05 or 95% confidence cutoff. For this reason, many groups now advocate the use of Decoy Database searches (either Reversed or Randomized versions of the same Forward/Normal database used for searching), presumably containing NO real sequences, with the assumption that the number of IDs of Decoy (not real) peptides or proteins at a particular score threshold accurately estimates the number of FALSE identifications from the Forward/Normal database (see for example Elias, J. E., et al., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations, Nature Methods 2 667-675 (2005).), and these Decoy database methods generally provide a more stringent estimation of False Discovery Rates (FDR) than the individual Search Algorithms internal estimates.

There are at least two ways to use the "Decoy" hits to calculate the False Discovery Rate, however. The most commonly used one, called the "**Group False Discovery Rate**" (or sometimes "Aggregate False Discovery Rate"), uses the formula 100\*(2\*number of Decoy Database IDs)/Total IDs to define the FDR at any cutoff score chosen - for example, if one had a list of 200 identified proteins with Mascot scores ranging from 389 down to 57, and this list

contained 4 "IDs" from the Decoy database, then the FDR estimate using the Group method for the **set** of IDs with Mascot scores 57 or higher would be 100\*(2\*4)/200 = 4%.

There are important limitations to the group/aggregate calculation, however, most importantly that the estimated FDR of 4% applies to the whole **set** - it is intuitively obvious that the IDs with scores in the 57-75 range are MUCH more likely to be false positives than those with scores in the 300-357 range. Thus, if the AVERAGE probability of a false positive in the whole set of 200 proteins is 5% and the proteins at the top of the list have much LOWER than 5% probability of being false positives, the entries at the bottom of the list are likely to have much HIGHER than 5% probability of being false positives.

The Proteomics System Performance Evaluation Pipeline (PSPEP) algorithm (Tang, W.H., Shilov, I.V., and Seymour, S.L. A Non-linear Fitting Method for Determining Local False Discovery Rates from Decoy Database Searches, Journal of Proteome Research 2008

Sep;7(9):3661-7. Epub 2008 Aug 14.PMID: 18700793) that we use gets around this by using the slope of the accumulated Decoy database hits vs. total IDs to calculate the Local FDR (also called "Instantaneous FDR") for each individual protein ID. This means that with a PSPEP-analyzed list of ID'd proteins where there are 684 proteins with an instantaneous FDR of 5% or less, based on the rate of accumulation of Decoy database hits, the LAST protein on that list has an estimated 5% probability of being a false positive, and other proteins higher up that list will have decreasing estimated probabilities of being false positives. (Download PDF from ABI explaining more about this)

The tables below show a portion of the PSPEP analyses from a few recent large iTRAQ datasets (one human, one mouse, one rat dataset) from different groups at the College of Medicine, and one can easily see how much more stringent the Local/Instantaneous FDR is compared to the more frequently used Group/Aggregate FDR estimate, i.e., there are far fewer proteins (~60-80%) included at a 1% or 5% instantaneous FDR cutoff than there are at the same 1% and 5% FDR cutoffs using the aggregate calculation method. Similar analyses are also performed by PSPEP at the peptide level, with similar relative changes using the Local/Instantaneous vs. Group/Aggregate estimates (see last table below):

#### **Protein Level False Discovery Rate Analysis**

#### **Number of Proteins Detected at Critical False Discovery Rates**

Critical Value

Protein N Cutoff

| Accepted FDR | Instantaneous FDR | Aggregate FDR |
|--------------|-------------------|---------------|
| 1.0%         | 1055              | 1320          |
| 5.0%         | 1206              | 2034          |

| 10.0% | 1334 | 2931 |  |
|-------|------|------|--|
| 33.3% | 2931 | 2931 |  |
| 50.0% | 2931 | 2931 |  |

### **Protein Level False Discovery Rate Analysis**

### **Number of Proteins Detected at Critical False Discovery Rates**

Critical Value Protein N Cutoff

| Accepted FDR | Instantaneous FDR | Aggregate FDR |
|--------------|-------------------|---------------|
| 1.0%         | 1471              | 1747          |
| 5.0%         | 1605              | 2256          |
| 10.0%        | 1686              | 3419          |
| 33.3%        | 6196              | 6196          |
| 50.0%        | 6196              | 6196          |
|              |                   |               |

## **Protein Level False Discovery Rate Analysis**

## **Number of Proteins Detected at Critical False Discovery Rates**

Critical Value Protein N Cutoff

| Accepted FDR | Instantaneous FDR | Aggregate FDR |
|--------------|-------------------|---------------|
| 4.00/        | 500               | 700           |
| 1.0%         | 568               | 738           |
| 5.0%         | 684               | 954           |
| 10.0%        | 746               | 1248          |
| 33.3%        | 1449              | 1449          |
| 50.0%        | 1449              | 1449          |

## **PEPTIDE Level False Discovery Rates**

## **Number of Spectra Identified at Critical False Discovery Rates**

Critical Value

Number of Spectra Identifed

| Accepted FDR | Instantaneous FDR | Aggregate FDR |
|--------------|-------------------|---------------|
| 1.0%         | 8559              | 11084         |
| 5.0%         | 10293             | 13405         |
| 10.0%        | 11099             | 15010         |
| 33.3%        | 12889             | 25534         |
| 50.0%        | 14006             | 25818         |