Does protein FDR have any meaning?

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

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

- 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.
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.
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.
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.
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.