Published in final edited form as:

J Proteome Res. 2013 February 1; 12(2): 1062–1064. doi:10.1021/pr301063v.

When Target:Decoy False Discovery Rate Estimations are Inaccurate and How to Spot Instances

Robert J. Chalkley

Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94158

Abstract

To address problems with estimating the reliability of proteomic search engine results from mass spectrometry fragmentation data the use of target-decoy database searching has become the *de facto* approach for estimating a false discovery rate. Several articles have been written about the effects of different ways of creating the decoy database, effects of the search engine scoring or effects of search parameters on whether this approach provides an accurate estimate; not all agreeing with each other's conclusions. Hence, there may be some confusion about how effective this approach is and how broadly it can be applied. Although it is generally very effective, in this article I will try to emphasize some of the pitfalls and dangers of using the target-decoy approach, and will indicate tell-tale signs that something may be amiss. This information will hopefully help researchers become more astute in their assessment of search results.

Keywords

proteomics; database searching; target-decoy database searching; false discovery rate; search parameters; post-translational modifications

Tandem mass spectrometry allows analysis of complex peptide mixtures derived from digestion of proteins and has become the enabling technology for proteomic analysis, whether it is discovery of protein composition in a sample or detection of post-translational modifications on these proteins. The identification of these peptides and modifications is through the use of database search engine software that compares observed fragmentation spectra to theoretical predictions of fragmentation spectra from peptides in a database to produce a peptide to spectrum match (PSM). There is a long list of software programs that can be used[1], all of which use very similar parameters for deciding which peptides should be considered and compared to, but they differ in the way they score the significance of a match between theoretical and observed fragmentation spectra. Some of these software report empirical, arbitrary scores, whereas others report statistical measures such as probabilities or expectation values. While statistical measures such as expectation values sound more useful measures, users must be careful not to read too much into their meaning, as they are not always statistically accurate: they often should be treated more like another

score. A significant reason why they are not necessarily accurate is a function of what is referred to as 'search space'. The software has to make an assumption of what peptides, or peptides with modifications, are possible to be present in the sample. This is determined by user-defined parameters such as the protein database to be queried, the specificity of the cleavage of peptides (whether both ends of the peptide need to be specific to the digestion enzyme used; whether missed cleavages should be allowed); what modifications should be considered to residues in the peptide and what mass accuracy the precursor ions were measured at. For example, as described in more detail later, considering phosphorylation increases the number of peptides being considered by an order of magnitude. Do unmodified peptide identifications become ten-fold less confident because a researcher tries to find phosphorylated peptides elsewhere in the data? A ten-fold change is generally the effect that occurs on expectation values due to this change in parameters. Hence, it is clear that the expectation value estimate can sometimes be somewhat crude.

When calculating a reliability measure for a PSM, as well as evaluating how well the observed spectrum correlates with the expected fragmentation of the peptide it is being matched to, software should also consider how many peptides there are in the database that could be derived from a precursor of the mass observed. This is easily understood if one looks at the expectation value calculation, which attempts to measure how many times a particular event is expected to happen at random:

$$\text{Expectation value} = \text{probability} * \text{number of trials}$$
$$= \text{probability of achieving a particular quality of match between an observed spectrum and a given theoretical spectrum} * \text{number of peptide entries in the database within the mass tolerance of the observed precursor mass}$$

If calculated accurately, the probability should be largely independent of the database being queried, whereas the number of trials is heavily influenced by search engine parameters and the database. Hence, it is important for users of these software to understand the effect of changing search parameters on how this will influence confidence measures. By narrowing the precursor mass tolerance, the number of peptides being considered per spectrum will be reduced, so expectation values will become more confident for a given quality of spectrum match. Conversely, by considering post-translational modifications the number of peptides being considered increases, so expectation values become more conservative. Concern has been expressed about this phenomenon leading to inaccurate FDR measurements for data with high precursor mass accuracy[2]. However, this is not a problem if the target-decoy (TD) database searching strategy is employed for FDR estimation, as the same effect will be observed for matches to target and decoy components of the database. Hence, although the score or expectation value threshold may change for a given FDR, the relative scores for target and decoy matches will be unaffected. This was misinterpreted as a problem[2], where the author noted that when narrowing the precursor mass tolerance, matches from the decoy database were lost at a faster rate than the target database. The error in his logic was forgetting that all matches to the target database were not random: if data is measured with higher mass accuracy than the precursor mass tolerance employed for the database search, then narrowing the precursor mass tolerance should lose practically none of the correct

answers; hence matches from the target portion of the database are lost at a lower relative rate (only random matches will be lost) and results become more reliable at a given score threshold. On the other hand, if data has not been measured with better mass accuracy than the search tolerance used, then losses to target and decoy databases will occur at a more similar relative rate; i.e. the effect of search parameters are not independent of the properties of the data being searched.

One of the better documented problems with TD estimation of FDR is its incompatibility with any scoring system that uses protein inference, such as weighting based on the presence of other peptides from the same protein. As more peptides are being matched to the target database, it is much more likely that a random PSM to the target database is to a protein already identified than a decoy PSM being to the same protein as another decoy database match. It was largely for this reason that the Percolator software removed protein inference parameters from its re-scoring [3].

Another less well publicized but common issue with the TD searching approach is when researchers, having employed this method to threshold results reported, then only focus on a subset of the answers and assume that these have the same reliability. This is a problem I have observed increasingly when reviewing articles where researchers are focusing on modified peptides and assume that these have the same reliability as the rest of their results, but it is not unique to modification analysis. The key facet to understand is that the incorrect results from the database search will closely reflect the composition of the peptides being considered by the search engine. For example, if a researcher considers phosphorylation of serines, threonines and tyrosines as variable modifications to their tryptic peptides in their database search and allow for one tryptic missed cleavage site, then around 90% of all the peptides being considered by the search engine are phosphorylated peptides; e.g. if there are four phosphorylatable residues in a peptide, then the search engine will consider four singly modified versions, six doubly modified, four triply modified and one quadruply modified. What this means is that roughly 90% of the incorrect results will be to phosphorylated peptides. It also means that a large percentage of the incorrect identifications will be to multiply phosphorylated peptides. Hence, if the sample did not contain 90% phosphorylated peptides, then the global FDR calculated by target-decoy searching will be inaccurate for the subset of peptides that are phosphorylated. A way to address this issue is to calculate the FDR based on the target-decoy results only for phosphopeptides; however, this is something that is rarely done[4].

So, how can one spot if there is a problem or bias in search results? Many of the signs are common sense observations based on the results. For example, if when looking for phosphorylated peptides the results contain identifications of doubly-phosphorylated peptides but the equivalent singly phosphorylated versions of the same peptide are not also found, then this is a highly suspicious result; similarly, if the majority of the identified phosphopeptides also contain another modification (e.g. the researcher also allowed for ubiquitination) then this is unlikely to be real; if the majority of the peptides identified have precursor mass errors of less than 10 ppm, but a significant number of the modified peptides have larger precursor mass errors, then this should be a concern.

An elegant example of where a researcher was able to flag results as unreliable without having access to the data followed this type of logic[5]. Foster noted that a set of published results contained an unusually large number of tryptic missed cleavage sites: most reported results contained two or more missed cleavage events. He showed that if he searched data with identical search parameters against a database that did not contain any correct answers (i.e. a decoy database) he observed exactly the same distribution of missed cleavages, whereas when the database contained correct sequences the distribution of tryptic missed cleavage sites differed significantly. When the raw data for this study was eventually made available, his concerns were proven to be well founded[6, 7]. This is also an interesting example due to the fact that Protein Prophet[8] was used for re-scoring the initial, incorrect results. Protein Prophet (and Percolator [3]) observe properties of the highest scoring answers, such as number of missed cleavage sites and number of tryptic termini, then assume that all correct answers will show a similar pattern, so re-score results accordingly. This is normally a sensible strategy and can significantly improve the ability to separate correct from random results, but it is reliant on there being more than a few correct answers in the initial database search engine results; in this example Protein Prophet probably trained its scoring on incorrect answers.

The large proteomic datasets produced on modern instruments make the viewing of a significant percentage of the data impossible. Recognizing this, journal publication guidelines require making available annotated spectra for those portions of the data that have the greatest potential for misinterpretation[9]. What I am advocating is that if researchers apply a little common sense when looking at their results, then this should save them (and journal reviewers) time in the long run, as it may reduce wild-goose chases based on unreliable data. Don't be fooled that if there is a statistical measure associated with a result, that this value was necessarily calculated properly.

Acknowledgements

I would like to acknowledge funding support from the Biomedical Technology Research Centers program of the NIH National Institute of General Medical Sciences NIH NIGMS 8P41GM103481.

References

1. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*. 2010; 73(11):2092–123. [PubMed: 20816881]
2. Cooper B. The Problem with Peptide Presumption and the Downfall of Target-Decoy False Discovery Rates. *Anal Chem*. 2012
3. Spivak M, Weston J, Bottou L, Kall L, Noble WS. Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. *J Proteome Res*. 2009; 8(7):3737–45. [PubMed: 19385687]
4. Baker PR, Medzihradsky KF, Chalkley RJ. Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state- and sequence-dependent scoring. *Mol Cell Proteomics*. 2010; 9(9):1795–803. [PubMed: 20513802]
5. Foster LJ. Interpretation of data underlying the link between colony collapse disorder (CCD) and an invertebrate iridescent virus. *Mol Cell Proteomics*. 2011; 10(3) M110 006387.
6. Knudsen GM, Chalkley RJ. The effect of using an inappropriate protein database for proteomic data analysis. *PLoS One*. 2011; 6(6):e20873. [PubMed: 21695130]

7. Foster LJ, Bromenshenk et al (PLoS One, 2011, 5(10):e13181) have claimed to have found peptides from an invertebrate iridovirus in bees. Mol Cell Proteomics. 2012; 11(1) A110 0063871.
8. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002; 74(20): 5383–92. [PubMed: 12403597]
9. Bradshaw RA, Burlingame AL, Carr S, Aebersold R. Reporting protein identification data: the next generation of guidelines. Mol Cell Proteomics. 2006; 5(5):787–8. [PubMed: 16670253]