Published in final edited form as:

J Proteome Res. 2011 April 1; 10(4): 2123–2127. doi:10.1021/pr101143m.

Comment on “Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies”

Marshall Bern* and Yong J. Kil†

Marshall Bern: bern@parc.com; Yong J. Kil: yongjk@gmail.com

*Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304

†Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304

Abstract

Everett et al. recently reported on a statistical bias that arises in the target-decoy approach to false discovery rate estimation in two-pass proteomics search strategies as exemplified by X!Tandem. This bias can cause serious underestimation of the false discovery rate. We argue here that the “unbiased” solution proposed by Everett et al., however, is also biased, and under certain circumstances can also result in a serious underestimate of the FDR, especially at the protein level.

Keywords

Proteomics; Mass Spectrometry; Peptide Identification; Bioinformatics; X!Tandem; Paragon

1 Introduction

Shotgun proteomics identifies proteins in a digested sample, sometimes along with their modification states and abundances, by matching tandem mass spectra (MS/MS spectra) to peptides from a protein sequence database. A “two-pass” or “multi-stage” proteomics search strategy includes an initial search that is then used to alter the search space for one or more subsequent searches.

Multi-stage strategies can be implemented with any proteomics search engine [1, 2, 3], but two search engines, X!Tandem [4] and Paragon [5], build multi-stage strategies into the core software so that the programs cannot easily be run in any other mode. X!Tandem uses an initial protein search to compile a list of likely proteins, and then uses a refinement stage to find additional peptide-to-spectrum matches (PSMs) using the proteins on the list. The protein search is usually a narrow search over a large protein database, considering for example only fully tryptic peptides with a small number of common modifications. The refinement stage then extends a wider search over the relatively small protein list, typically allowing nonspecific digestion at one or both peptide termini and a larger number of modifications. Paragon follows a similar approach but at a different level of granularity by searching the most likely peptides, called “hot” peptides, for nonspecific cleavage and a large number of modification forms.

Everett, Bierl, and Master [6] made an important contribution to proteomics bioinformatics by pointing out that the popular target-decoy approach (TDA) to false discovery rate (FDR) estimation [7] can give biased results for multi-stage search strategies. Assume we supply X!Tandem with a protein database containing statistically balanced target and decoy proteins, for example, 50,000 human protein sequences along with reversals of each of these proteins [8]. The initial search may winnow this list to 1000 sequences, 900 target proteins and 100 reversed proteins, by filtering out all proteins that do not receive a PSM with E-

value 0.1 or less. Now in the refinement stage a false match to a target protein is about 9 times as likely as a false match to a reversed protein, simply due to the **quantity** of target and decoy sequence. This violates the key assumption of TDA: false matches are equally likely to hit target and decoy sequences. Here we remind the reader that a statistical estimator is unbiased if its expected value equals that of the quantity being estimated. Thus for unbiased FDR estimation, we require that the number of decoy PSMs is roughly equal to the number of false target PSMs for any given score or E-value threshold, and that the ratio of these two quantities converges to one as the number of PSMs grows large.

Everett et al. [6] proposed and implemented a fix to X!Tandem. For the scenario just sketched, this fix throws out the 100 reversed proteins surviving the initial search and reverses the 900 target proteins to build a balanced target/decoy database of 1800 proteins for use in the refinement stage. Unfortunately this proposed fix falls into a new trap, which we mentioned in passing in a previous paper [9] and which we now describe more thoroughly here.

2 Problem and Solution

The problem with the solution by Everett et al. for our example scenario is the following: the 100 reversed proteins surviving the first stage are “lucky” decoys that received PSMs of E-value 0.1 or less; these lucky decoys are the luckiest among the original pool of 50,000 decoys. Similarly we expect that among the 900 target proteins there are also around 100 lucky proteins that are not actually in the sample. The refinement stage usually generalizes the initial search, so that proteins lucky in the first stage are likely to remain lucky in the refinement stage. The Everett solution throws away the lucky decoys and replaces them with reversals of the target proteins, which are not known to be lucky. (We would expect only about 2 of them to be lucky by chance.) Thus the Everett solution produces a target/decoy database for the refinement step that is biased not by the quantity, but rather by the **quality**, of the target and decoy sequences, with false targets having a better fit to the MS/MS data set than do the decoys.

One mitigating factor is that X!Tandem sets aside the mass spectra that found PSMs with E-value better than 0.1 in the first stage and does not reuse them in the refinement stage. Craig and Beavis [10] describe this step as optional, so we study multi-stage search strategies with and without this **matched spectrum removal** (MSR) step. With the MSR step, a protein lucky in the first stage will not be lucky in the refinement stage due to exactly the same spectrum, but it is quite possible that another MS/MS spectrum of the same or related precursor ion will find a PSM to the same peptide (possibly modified) with E-value below 1.0, which is the default cutoff for the refinement stage.

What we need is a target/decoy database for the second stage in which the targets and decoys are balanced in both quantity and quality. There are several natural ways to do this. For example, we might choose the target proteins based on a criterion of at least one PSM with E-value 0.1 or better, and then choose a looser E-value threshold for the decoy proteins, one that balanced the quantities of target and decoy sequence. For example, an E-value threshold of 100 may admit sufficient decoy sequence. Alternatively we might choose targets and decoys based on the same criterion, and then supplement the decoys with reversals of the target proteins until the quantity of target and decoy sequence is balanced. We chose the latter approach for the computational experiments reported below, because in some searches it may not be possible to find a criterion loose enough to balance the quantities of target and decoy sequence. For simplicity, we did not actually balance the total number of residues or tryptic peptides in the target and decoy proteins, but just the numbers of target and decoy proteins. We also did not attempt to balance fine details of the target and

decoy proteins, for example, individual protein lengths, residue content, distribution of masses of tryptic peptides, amount of redundancy, and so forth. In our previous paper [9], we built in a conservative bias by choosing the lucky decoys along with the reversals of all the lucky targets. This strategy uses an unnecessarily large amount of decoy sequence, but does a better job matching the fine details.

3 Computational Experiments Using ByOnic

We first studied the bias of various TDAs for multi-stage search with our own search engine, ByOnic [11]. For all our experiments, we used the publicly available Aurum data set of 9987 MALDI-TOF-TOF spectra from a sample containing 248 purified reference proteins, along with a few contaminants [12]. A realistic, in fact, a well-optimized, multi-stage search strategy for this data set uses an initial fully tryptic search with no modifications enabled except for the fixed modification of carbamidomethylated cysteine (C[+57]), followed by a semitryptic refinement search with the following variable modifications enabled: oxidized methionine (M[+16]), oxidized and doubly oxidized tryptophan (W[+16] and W[+32]), pyro-glu cyclization (N-terminal Q[−17], C[+57] [−17], and E[−18]), deamidated asparagine and glutamine (N[+1] and Q[+1]), methyl esterification (E[+14]), and cysteine propionamide (C[+71]). Precursor and fragment mass tolerances were set at 250 and 300 ppm respectively.

For our first experiment, we studied two-stage ByOnic search of the Aurum data set against a **non-sense database** containing 49,450 random target proteins and their reversals. This experiment removes the confounding effect of true target proteins so that we can directly compare the false target and decoy PSM rates. The random target proteins were made by starting with the IPI human protein database and then applying a randomly chosen SNP mutation independently at each residue. In Figure 1 the plotted lines show the numbers of decoy and target identifications as the x- and y-coordinates of curves, parameterized by score rank, running from lower left to upper right. The four curves show the following search strategies: (XT) the X!Tandem approach, which lets lucky decoys from the first stage serve as decoys for the second stage; (XMSR) the X!Tandem approach with MSR; (EV) the Everett approach, which throws away the decoys from the first stage and reverses the targets; (EMSR) the Everett approach with MSR; and (BK) the approach proposed here, which keeps the lucky decoys from the first stage but augments them with enough extra decoys to balance the total number of target and decoy proteins. The criterion for acceptance of a target protein into the second stage database was at least one PSM with ByOnic score 250, which is approximately equal to a Mascot score of 25 and lets 709 target proteins into the protein list for the second stage. For XT and XMSR the same criterion allowed 714 decoy proteins into the second stage. The criterion of ByOnic score 250 is looser than the criterion of X!Tandem E-value of 0.1 or better used by X!Tandem's default multi-stage search, and we shall say more about this criterion in the Conclusions. (Note that a PSM to an ambiguous peptide, that is, a peptide that appears in more than one protein, brings all the proteins that contain it into the second stage.) As stated above, strategies EV and EMSR used reversals of the 709 target proteins as decoys, and discarded the 714 decoys that made it through the first stage. For XMSR and EMSR, the plot shows the union of identifications from the first and second stages, as in X!Tandem. Since neither the target nor the decoy proteins match the spectra, persistent deviations away from the $x = y$ diagonal line are evidence of bias. We see that the Everett solution is badly biased without the MSR step and mildly biased with the MSR step. When searching against a nonsense database, X!Tandem's strategy is unbiased with or without MSR, because the second-stage protein list contains roughly equal numbers of target and decoy proteins. BK is the approach proposed here, which explicitly balances the number of target and decoy and proteins.

Usually PSMs are not the final product of a proteomics experiment. The next stage in the bioinformatics pipeline integrates PSMs into a ranked list of protein identifications, and target/decoy approaches are frequently used to estimate FDR at the protein level as well [1, 13]. We applied our own protein-ranking program, ComByne [1], to the outputs from the nonsense database searches. Strategies XT, XMSR, and BK give roughly equal numbers of target and reverse proteins in the list of top protein groups, no matter where we cut the list. The highly biased strategy (EV) gives no decoys among its top 10 protein groups and only 15 decoy proteins among its top 200 protein groups. The mildly biased strategy EMSR gives 2 decoys among its top 10 protein groups and 19 decoys among its top 60 protein groups.

For our second experiment, we studied two-stage ByOnic search of the Aurum data set against an uncorrupted database containing 57,355 target proteins and their reversals. The target proteins included the IPI human protein database, an *E. coli* database (since the reference proteins were expressed in *E. coli*), and a “crap” database of likely contaminants. We performed the same five two-pass search strategies as described for the first experiment. We also included (IPASS) a slow one-pass search, which searched the full database allowing semitryptic cleavage and modifications. This search used reversals of the full database as decoys, the most widely accepted method for unbiased FDR estimation.

Figure 2 shows that the biased strategies XT and EV give more target matches than the unbiased strategies BK and IPASS. For strategy EV, the FDR underestimation is even more apparent at the protein level, as this strategy gives 318 target proteins, including a number of human proteins that were not put into the sample, ranked above the top-ranked decoy protein. Strategy XT, which uses decoys of adequate quality but inadequate quantity, does not give a corresponding bias at the protein level, as it produces a protein list in almost exact agreement (identities and ranks) with the one produced by the unbiased search, with 280 target proteins ranked above the first reverse, almost all of them either reference proteins or plausible contaminants. Strategy BK outperformed IPASS, as two-pass search focuses the search on likely peptides [1]. Even though their FDR estimations are biased in the optimistic direction, strategies XMSR and EMSR, which use the MSR step, actually give worse sensitivity/specificity than the unbiased searches. The first stage accepted about 300 decoys among its 3419 identifications with ByOnic score 250 or better (again slightly more identifications than would have been accepted by an X!Tandem E-value of 0.1 better), and the MSR step prevents second-stage correction of these identifications. Along with giving measurably worse FDR, the MSR step may also give more undetectable “half-right” identifications, for example a PSM to an unmodified peptide with E-value 0.05 found in the first match will prevent a match to a deamidated version of the same peptide with E-value 0.0001.

4 Computational Experiments Using X!Tandem

We obtained a modified version of X!Tandem from Logan Everett in order to evaluate TDA estimation of FDR using the original two-pass search engine. We used the default X!Tandem strategy (XMSR) and Everett et al.'s implementation of strategy EMSR. We implemented a new strategy (BKMSR), which makes the second-pass database as in strategy BK but uses the matched spectrum removal step. We did not change the criterion for acceptance (E-value 0.1 or better) for a first-pass match. For X!Tandem we removed W[+16] as a variable modification, but retained the more common W[+32], because X!Tandem allows only one variable (“potential”) modification per residue type. We also opened the mass tolerances to 400 ppm for both precursors and fragments, because X!Tandem gives poor results with the narrower tolerances we used for ByOnic.

We repeated the first experiment, searching Aurum against a nonsense protein database, for the three strategies, XMSR, EMSR, and BKMSR, and to our surprise we found that none of the three strategies showed significant bias. Recall that with ByOnic we saw that EMSR showed a bias due to an excess of lucky targets relative to lucky decoys.

We also used X!Tandem to search Aurum against the uncorrupted database using XMSR, EMSR, and BKMSR, with results shown in Figure 3. We see small differences in performance, consistent with both XMSR and EMSR producing slightly more false target PSMs than decoy PSMs. Figures 2 and 3 do not give an exact comparison of ByOnic and X!Tandem, due to X!Tandem's loss of a small number of W[+16] identifications, and apparent differences in the interpretation of mass tolerances. With 800 ppm mass tolerances X!Tandem gives performance closer to ByOnic's.

5 Conclusions and Discussion

In this comment, we have repeated Everett et al.'s warning about bias in the target/decoy approach to FDR estimation as applied to multi-stage proteomics search strategies. We also point out and correct a new bias introduced by Everett et al.'s proposed solution to the problem. As we saw with ByOnic and X!Tandem in the nonsense-database experiments, these biases may differ in severity depending upon the search engine. We believe that X!Tandem behaved differently because it uses an E-value score that evaluates the top PSM relative to all the other PSMs for the same spectrum, whereas ByOnic uses an absolute score, akin to a dot product of predicted and observed spectra. A peptide lucky in the first pass may still achieve top scores for spectra in the second pass and hence be lucky for ByOnic, but these top scores may not stand out from the bulk of the second-pass scores, which in this case might include many modified forms of the same peptide, and hence not be very lucky for X!Tandem.

Biases tend to be mild when the multi-stage search includes the matched spectrum removal step, but severe without the matched spectrum removal step, because this step retains high-scoring identifications from the unbiased first stage and thereby dilutes the bias. The MSR step, however, may lose sensitivity and specificity, because the second-stage search space is better tailored to the data set than is the first stage, so we recommend against MSR. Finally, biases may be more severe at the protein level than at the individual spectrum level, because correct PSMs concentrate on a small number of proteins but false target PSMs hit many proteins.

The solution proposed here works for X!Tandem-like multi-stage strategies, but other multi-stage search strategies may require slightly different solutions. For example, Paragon is currently incompatible with TDA, because it expands the search space around “hot” peptides, which are more likely to be targets than decoys, and hence noise spectra are compared against more targets than decoys. This approach could be fixed by using reversed proteins as decoys and locking the temperature of a reversed peptide to the temperature of its target counterpart, so that the target and decoy search spaces expand in parallel. This solution introduces a new subtlety: modifications that depend upon both position and residue identity, for example, N-terminal Q[−17], may not apply equally to target and decoy. We leave it to the interested reader to find a correction.

Finally we take this opportunity to remark on the role of multi-stage search. In its original conception [10], multi-stage search serves both as a speed-up and as a means to find new PSMs for proteins already known or strongly suspected to be in the sample. This role suggests the use of a stringent criterion (such as E-value 0.1) to assemble a high-quality protein list with a modest false positive rate. A stringent criterion, however, runs the risk of false negatives, proteins with low confidence from the first search that if allowed into the

refinement search would find high-scoring modified or semitryptic PSMs to confirm their presence in the sample. With this in mind, we usually set a loose criterion (such as ByOnic score 150, roughly equivalent to Mascot 15) for inclusion in the second stage, so that the protein list typically includes many more false proteins than true proteins. We have found [1] that multistage search can improve protein sensitivity as well as peptide and modification sensitivity, especially in samples with a small number of frequent modifications (for example, pyro-glu N-termini and oxidized methionine) and frequent nonspecific cleavage (for example, ragged N-termini due to endogenous proteases).

Acknowledgments

Marshall Bern was supported in part by NIH grant R21GM094557 and Yong Kil by an NSF Computing Innovations Fellowship. We thank Logan Everett for sharing his code and for help in modifying X!Tandem.

References

1. Bern M, Goldberg D. Improved ranking functions for protein and modification-site identifications. *J Comp Biology*. 2008; 15:705–719.
2. Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*. 2002; 2(10):1426–1434. [PubMed: 12422359]
3. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of posttranslational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics*. 2006; 5(4):652–670. [PubMed: 16352522]
4. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 2004; 20(9):1466–1467. [PubMed: 14976030]
5. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol Cell Proteomics*. 2007; 6:1638–1655. [PubMed: 17533153]
6. Everett LJ, Bierl C, Master SR. Unbiased statistical analysis for multi-stage proteomic search strategies. *J Proteome Research*. 2010; 9(2):700–707. [PubMed: 19947654]
7. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007; 4(3):207–214. [PubMed: 17327847]
8. Moore RE, Young MK, Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom*. 2002; 13(4):378–386. [PubMed: 11951976]
9. Bern M, Phinney BS, Goldberg D. Reanalysis of *Tyrannosaurus rex* mass spectra. *J Proteome Res*. 2009; 8(9):4328–4332. [PubMed: 19603827]
10. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom*. 2003; 17(20):2310–2316. [PubMed: 14558131]
11. Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal Chem*. 2007; 79:1393–1400. [PubMed: 17243770]
12. Falkner JA, Kachman M, Veine DM, Walker A, Strahler JR, Andrews PC. Validated MALDI-TOF/TOF mass spectra for protein standards. *J Am Soc Mass Spectrometry*. 2007; 18:850–855.
13. Weatherly DB, Atwood JA, Minning TA, Cavola C, Tarleton RL, Orlando R. A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics*. 2005; 4(6):762–772. [PubMed: 15703444]

Figure 1.

Results of five two-stage ByOnic searches of the Aurum spectra (MALDI-TOF-TOF spectra of reference proteins) against a nonsense protein database containing no true proteins. EV shows the substantial bias of Everett et al.'s proposed solution when the multi-stage search does not use the matched spectrum removal step. EMSR shows the mild bias of the proposed solution when the MSR step is used. XT, XMSR, and BK do not show any bias in this nonsense-database experiment.

Figure 2. Results of six ByOnic searches of the Aurum spectra against a good protein database. EV and XT are two-stage searches biased in quality and quantity respectively. BK is a two-stage search using decoys matched in both quality and quantity. 1PASS is a one-stage search using the most widely accepted target-decoy approach. XMSR and EMSR are two-stage searches with the matched spectrum removal step.

Figure 3. Results of three X!Tandem searches of the Aurum spectra against a good protein database. All searches used X!Tandem's matched spectrum removal step with acceptance criterion E-value 0.1 or better. EMSR and XMSR are two-stage searches biased in quality and quantity respectively. BKMSR is a two-stage search using decoys matched in both quality and quantity.