J. Am. Soc. Mass Spectrom. (2011) 22:1111–1120
DOI: 10.1007/s13361-011-0139-3

CRITICAL INSIGHT

Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong

Nitin Gupta,[1] Nuno Bandeira,[2,4] Uri Keich,[3] Pavel A. Pevzner[1,2]

[1]Bioinformatics Program, University of California San Diego, La Jolla, CA, USA
[2]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
[3]School of Mathematics and Statistics, University of Sydney, Sydney, Australia
[4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

Abstract

The target-decoy approach (TDA) has done the field of proteomics a great service by filling in the need to estimate the false discovery rates (FDR) of peptide identifications. While TDA is often viewed as a universal solution to the problem of FDR evaluation, we argue that the time has come to critically re-examine TDA and to acknowledge not only its merits but also its demerits. We demonstrate that some popular MS/MS search tools are not TDA-compliant and that it is easy to develop a non-TDA compliant tool that outperforms all TDA-compliant tools. Since the distinction between TDA-compliant and non-TDA compliant tools remains elusive, we are concerned about a possible proliferation of non-TDA-compliant tools in the future (developed with the best intentions). We are also concerned that estimation of the FDR by TDA awkwardly depends on a virtual coin toss and argue that it is important to take the coin toss factor out of our estimation of the FDR. Since computing FDR via TDA suffers from various restrictions, we argue that TDA is not needed when accurate p-values of individual Peptide-Spectrum Matches are available.

Key words: Computational proteomics, Target-decoy approach, False discovery rate, False positive rate, Database search, Decoy database, P-value

Introduction

While tandem mass spectrometry (MS/MS) has emerged as a key technology in proteomics, the issue of the statistical significance of peptide identifications remained controversial, and the target-decoy approach (TDA) emerged as the standard for computing the false discovery rates (FDR) [1].

Elias and Gygi, 2007 concluded that TDA is "..accessible to any laboratory using any instrument platform and any database-searching algorithm" [1]. Such wide applicability makes TDA particularly attractive for comparing the performance of different search tools. Indeed, it appears that we can treat any search tool as a black box: to evaluate its utility, we need not look into how it works, but only check whether it finds more peptides compared to other search tools at the desired false discovery rate (FDR). Balgley et al., 2007 [2] compared the performances of four popular search tools using TDA, noting that "the target-decoy search strategy permits an impartial initial assessment of search results."

We argue that the assumption that TDA can be used with any MS/MS database search tool is incorrect and that some (useful) MS/MS database search tools are non-TDA-compliant.

While Elias and Gygi, 2007 [1] themselves noted that TDA imposes some restrictions on the MS/MS algorithms, the methods for checking whether a particular algorithm (treated as a black box) is TDA compliant were never developed. We demonstrate that some popular MS/MS search tools (e.g., a popular two-pass version of X!Tandem) are not TDA-compliant and argue that further studies are

Electronic supplementary material The online version of this article (doi:10.1007/s13361-011-0139-3) contains supplementary material, which is available to authorized users.

Correspondence to: Pavel Pevzner; *e-mail:* ppevzner@ucsd.edu

Received: 1 November 2010
Revised: 19 February 2011
Accepted: 22 February 2011
Published online: 5 May 2011

needed to check if some other tools (e.g., Sequest and Mascot) are TDA-compliant. Moreover, as we show, it is easy to develop a non-TDA compliant tool that offers a much better (TDA-estimated) performance than all TDA-compliant tools. Since the distinction between TDA-compliant and non-TDA compliant tools remains elusive, we are concerned about a possible proliferation of non-TDA-compliant tools in the future (developed with the best intentions).

TDA has done the field of proteomics a great service by filling in the need to estimate FDRs, but it must protect itself against possible (even if unintentional) exploitation by non-complying search tools. At the same time, the proteomics community must dispel the prevalent notion about its universal applicability.

The first line of defense against exploitation of TDA is to ensure that the search tool used is TDA-compliant (compliance test). However, since such tests are difficult to do in case of commercial software (without access to the source code), multi-stage search tools, or complex tools like InsPecT, should TDA be avoided in conjunction with such tools?

Methods

Can TDA be Exploited to Disguise Bogus Peptide Identifications as Good Ones?

We illustrate a flaw in the black box treatment of search tools in TDA using a *reductio ad absurdum* argument. Starting from any MS/MS search tool T, we can use it to design a new search tool T^+ by modifying T that finds twice as many peptide identifications as reported by T at the same FDR (Figure 1).

TDA assumes that the distribution of scores of incorrect identifications in the target database is the same as the distribution of scores in the decoy database [3]. In other words, if a bogus peptide is located in both the target or the decoy database, the expected score of a spectrum matching this peptide must be the same in both databases. When using T^+, such a bogus peptide may get a higher score if it happens to be located near a peptide scored highly by T. Since there are more high-scoring peptides in the target database, this bogus peptide is likely to get a higher score in the target database, thus violating the TDA assumption. Because TDA only looks at the scores (assigned by the search tool) but not at the false positive rate (FPR) [3] of individual identifications,[1] it becomes possible for such bogus identifications to be included in the results as long as the overall FDR among all identifications is acceptable. We are unaware of a test for checking whether a specific tool (treated as a black box, without access to its source code) is TDA-compliant.

Assigning a peptide from the same protein to an arbitrarily chosen unidentified spectrum, it assigns a peptide to the best matching unidentified spectrum. If the black box in Figure 1 is substituted by a black box with a two-pass X! Tandem inside, it will be difficult to detect that it is not TDA-compliant.

Can Tools That Do Not Comply with TDA be Useful?

While it is easy to discard T^+ as outrageous, judgment becomes difficult when the search tool is well-intentioned and produces useful results, though benefits unduly by exploiting TDA (see T^* search tool in Supplement A). Even proteomics leaders can step into this trap as illustrated by Percolator [4] that originally was published as a non-TDA-compliant tool (shortly after publishing [4], the authors of Percolator realized that it was not TDA-compliant and corrected it in [5], see Supplement J).

X!Tandem, a popular MS/MS search tool, uses a two-stage search approach [6]. When using TDA with X!Tandem, one starts with the combined database containing target and decoy sequences in equal proportions, but the filtering during the first-stage will often increase the proportion of the target sequences in the remaining database. As a result, the number of identifications in the decoy database in the second-stage will underestimate the number of incorrect identifications in the target database, resulting in unduly low estimates of FDR. Everett et al., 2010 [7], Tharakan et al., 2010 [8], Nesvizhskii, 2010 [9], Kim and Bandeira, 2010 [12], and Bern and Kil, 2011 [10] recently emphasized the concern that multi-stage searches may violate the key assumption about the rates of false positive identifications reported by TDA. These concerns are not limited to X!Tandem[2] since users of OMSSA and Mascot often use the second-pass search as a default option incorporated into these tools. Similar issue was encountered in [11] with respect to the two-stage search with ByOnic and discussed in [10].

Some peptide identification approaches proposed to use information about other peptides identified in a protein to adjust the peptide-level scores. Such protein-level feedback will lead to different score distributions for bogus peptides in the target and the decoy databases, thus violating the fundamental TDA assumption. We do not condemn these search tools; rather, only argue that using TDA to evaluate their FDR is illegitimate. Yet, X!Tandem (originally designed to compute FPR or E-values, rather than FDR by TDA) was one of the tools included in TDA-based benchmarking by Balgley et al, 2007 [2][3] and Kandasamy et al., 2009 [13]. However, if the protein-level features of

[1] FPR is defined as the probability that a spectrum matches a random peptide with a score exceeding a threshold. FPR can be used to compute the probability that a spectrum matches a random protein database with a score exceeding the threshold. See formal definitions of FPR and FDR below.

[2] We and [7–10] are only claiming that the most commonly used multi-stage option in X!Tandem is incompatible with TDA. If one runs X!Tandem with a single-pass option, it becomes TDA-compatible (but slow).

[3] The TDA tests of non-TDA-compliant X!Tandem in [2] revealed that X! Tandem (along with OMSSA) outperforms other tools with respect to FDR.

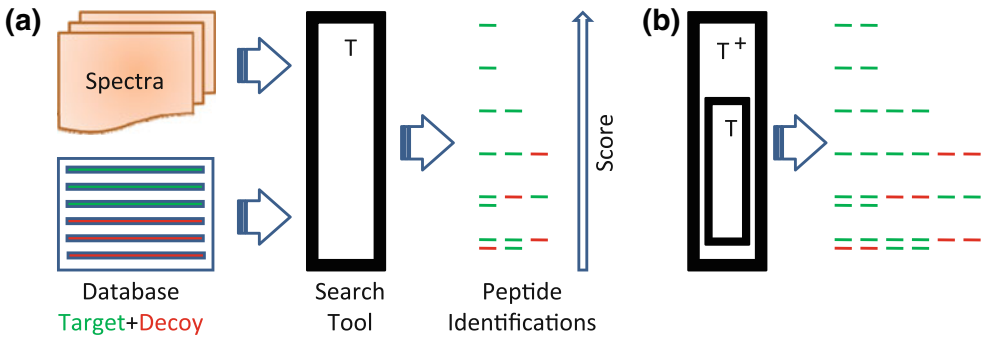

Figure 1. Black box representation of an MS/MS search tool in target-decoy approach (TDA). (**a**) The search tool T takes as input the set of all spectra and the database of amino acid sequences (combination of the target and the decoy databases, but T cannot distinguish between them). T reports the list of peptides along with scores. The peptides that come from the target and the decoy databases are marked in green and red, respectively. (**b**) For each peptide identified by T (with score x), T^+ simply adds an extra spurious peptide from the *same* protein (e.g., an overlapping peptide), assigning it to an arbitrarily chosen unidentified spectrum with the same parent mass and the same (fake) score x. Thus, T^+ will double the number of peptide identifications in both target and decoy databases, for any score threshold. T^+ is clearly a gimmick that we used for exposing vulnerability of TDA. We do not recommend using it in practice

Percolator or two-pass searches are useful (i.e., lead to new biological discoveries), should they be switched off? Does it mean that TDA-compliance limits our ability to identify more peptides and thwarts new discoveries? Should the new proteogenomics findings made with our non-TDA-compliant tool T^* (see Supplement A) be dismissed as artifacts?

What are the Disadvantages of TDA?

We argue that the time has come to critically re-examine TDA and to answer the following questions:

- Should X!Tandem (a useful tool that is not TDA-compliant) be excluded from TDA studies? If yes, how can one evaluate its performance and compare it with other tools?
- Should Mascot (a useful tool whose source code is not available for a test of TDA compliance) be excluded from TDA studies?
- In a typical TDA search, there are often thousands of peptides in the target database scoring higher than any peptide in the decoy database resulting in 0% FDR for these high-scoring peptides. Since there is no such thing as 100% accurate peptide identifications, something is wrong with TDA in this case, presumably the decoy database size is too small to accurately measure FDR. Should TDA be run with decoy databases that are much larger than target databases to accurately measure FDRs of highly reliable peptide identifications? This argument is further compounded in searches with highly accurate precursor masses since the number of database peptides with a given precursor mass becomes very small essentially reducing the search to Peptide Mass Fingerprinting.
- Larger decoy databases better sample the space of random peptides. Therefore, TDA should work even better if one uses a decoy database that is twice larger than the target database (in this case FDR should be normalized since the number of decoy hits is expected to double). The same is true if the size of the decoy database increases by a factor of 10 or 100. This experiment with Sequest reveals that the number of identified peptides changes significantly with varying the size of the decoy database (for the same FDR!), an artifact of using δ-scores. Does it mean that (i) Sequest is not TDA compliant, (ii) TDA only works correctly for a 50–50 split between the sizes of the target and decoy databases, or (iii) TDA does not provide a reliable estimate of FDR in the case of 50–50 split and should be practiced with decoy databases that are larger than the target database?
- TDA awkwardly depends on a virtual coin toss and clearly gives an inaccurate estimate of FDR in the case of small databases (e.g., decoy database consisting of a single protein like in studies of monoclonal antibodies). What are the limits of TDA applicability when it comes to lowering the size of the database? 1000 amino acids or 100,000 amino acids? Also, TDA approach gives an inaccurate estimate of FDR in the case of small spectral datasets. What are the limits of TDA applicability when it comes to lowering the size of the spectral dataset? 1000 spectra or 100,000 spectra?
- Imagine a sample containing exactly 1000 human peptides each producing a nearly perfect Peptide-Spectrum Match (PSM) with a low FPR. All other spectra in the sample match the database with very low scores corresponding to a high FPR (random hits). If one sets an FDR threshold of 3%, the TDA approach will return $\approx$ 1030 matches in the target database and $\approx$ 30 matches in the decoy database. Would it be better to setup an FPR (rather than FDR) threshold for individual PSMs (that would clearly separate reliable and unreliable identifications) and avoid contaminating

the output of the TDA approach with ≈ 30 bogus peptide identifications?

- The choice of a specific FDR cutoff in proteomics is somewhat arbitrary: it is hard to find a study with a biological conclusion that critically depends on a specific FDR cutoff (e.g., that holds at 0.5% FDR but does not hold at 5% FDR). Thus, a conclusion like "we found 1000 peptides at 1% FDR" can often be substituted by "we found 1347 peptides at 5% FDR." What a researcher really wants is to rank identified peptides in the order of their statistical significance and to evaluate FDR for top n peptides in the ranked list. It is important to realize that TDA does not offer such a ranked list, i.e., the top 1000 peptides output by Sequest or Mascot typically do not represent the 1000 most statistically significant peptides.
- There exists significant disagreement among leading proteomics researchers on how to apply TDA, e.g., using separate [1] or combined [15] TDA approaches. These approaches produce different FDRs and thus at most one of them is correct.

Can Computing FDR via FPR Substitute Approximating FDR via TDA?

TDA is not a universal tool: it does not allow one to evaluate FDR of some useful tools like a two-pass X!Tandem or Percolator with protein-level features. While computing rigorous FPRs of individual peptide identifications is an alternative way to compute FDR without TDA, efficient (polynomial-time) algorithms for computing FPRs are available for some but not all MS/MS tools. The authors of X!Tandem were the first to describe an algorithm for approximating (rather than computing) FPRs via a continuous approximation of a discrete random variable but such approaches are notoriously inaccurate when one attempts to approximate the extreme tail of the distribution, the most important region for FDR estimates in mass spectrometry.[4]

FPRs do not suffer from the shortcoming of TDA-based estimation of FDR and the concerns described above disappear if one switches to FPR-based estimation of FDR. For example, FDR for arbitrarily small spectral datasets and protein databases cannot be reliably approximated via TDA but can be computed precisely via FPRs. We therefore view the proliferation of TDA approaches in proteomics as a historical accident that only happened because the popular database search tools (Sequest and Mascot) failed to compute reliable FPRs of their individual identifications. Elias and Gygi, 2007 [1] wrote: "For any analytical tool to be truly useful there must be a convenient way to assess the validity of its results." By failing to rigorously assess the statistical significance of individual peptide identification, Sequest and Mascot failed to satisfy this test and triggered the proliferation of TDA approaches for assessing validity of peptide identifications at bulk (rather than validity of individual peptide identifications). TDA is not designed for evaluating the reliability of individual peptide identifications, a critical requirement for many applications like analyzing one-hit-wonders or in proteogenomics where a single spectrum may provide evidence for a new splicing variant.

However, it is not a good reason to make the entire field of proteomics hostage of TDA and it is unfair to marginalize some good tools (e.g., X!Tandem) because some other good tools (e.g., Sequest and Mascot) are unable to compute FPRs in polynomial time and thus have to settle for computing FDR via TDA. Moreover, it remains to be seen whether Mascot and Sequest are TDA-compliant (see below). Without rigorous statistical foundations, there is little trust in the score-based ranking of Sequest and Mascot identifications (i.e., their higher scoring PSMs may have higher FPRs than lower scoring PSMs) thus limiting their ability to analyze the reliability of individual peptide identifications. However, any PSM (identified by Sequest, Mascot, or any other tool) can be evaluated using other tools. For example, one can design a chimeric tool Sequest ⊕ Mascot that uses Sequest to generate a list of PSMs and uses Mascot to score them. Thus, one can generate Sequest or Mascot PSMs (or even combine them), rescore them with another statistically solid tool, and compute rigorous FPRs of their identifications (albeit, with a different scoring that allows computing FPR in polynomial time). If such rescoring leads to an increase in the number of peptide identifications, there is nothing wrong in using it to evaluate FPRs of tools that lack polynomial algorithms for computing FPRs (see Supplement B on benefits of turning scoring functions into FPRs).

Results

Estimating the False Discovery Rate using TDA

Below we use a probabilistic rather than a statistical hypothesis testing framework. While these two frameworks peacefully coexist and complement each other in genomics,[5] proteomics was dominated by the statistical hypothesis testing.

The common view is that the stochastic nature of spectra makes them more difficult to analyze than sequences (in a rigorous probabilistic framework). However, as soon as one defines a scoring function, matching a spectrum against a database becomes equivalent to exact matching of sets of peptides (spectral dictionary [17]) against a database. While this problem is well studied in probabilistic combinatorics [18], we are not aware of any studies of this problem in the statistical hypothesis testing framework.[6]

[4]Indeed, many correct PSMs attain scores that are very close to the maximum possible score for a given spectrum, the region where continuous approximations are particularly unreliable.

[5]E.g., BLAST analysis was fully developed in a probabilistic framework [16] and was cast as statistical hypothesis testing much later.

[6]We are not discarding the statistical hypothesis testing framework in proteomics but rather argue that it is not the only approach to analyzing spectra.

While in mass spectrometry literature, FPR and FDR are often mistakenly equated (and even interchanged) they represent very different notions. Below we precisely define these notions in the context of mass spectrometry and describe relationships between them.

Given a spectrum σ and a peptide π, a *scoring function* is a black box that outputs[7]

$$Score(\sigma, \pi)$$

Given a spectrum σ and a database DB, an MS/MS *database search algorithm* outputs a peptide $Peptide(\sigma, DB)$ from DB with maximum $Score(\sigma, \pi)$ among all peptides $\pi \in DB$:

$$Peptide(\sigma, DB) = \text{argmax}_{\pi \in \text{DB}} Score(\sigma, \pi)$$

If there are multiple peptides with the same maximal score the algorithm randomly selects one of them using a fair coin flip. We define

$$Score(\sigma, DB) = Score(\sigma, Peptide(\sigma, DB))$$

We assume the algorithm is given a threshold t and it declares the PSM $(\sigma, Peptide(\sigma, DB))$ as significant, or as a "*discovery*", if $Score(\sigma, Peptide(\sigma, DB)) > t$.[8] This PSM discovery can be either "true" or "false". It is customary in such a setup to define the FDR as the rate of false discoveries: what is the proportion of false discoveries among all discoveries.

Given a target database T, a decoy database R is a random database of the same size as T. A *combined* target-decoy database $T \oplus R$ is a concatenation of T and R.

Consider first the event: the spectrum σ generates a decoy discovery (necessarily a false one), or:

$$\{Score(\sigma, T \oplus R) > t\} \cap \{Peptide(\sigma, T \oplus R) \in R\}.$$

Recall that if $Score(\sigma, Peptide(\sigma, T)) = Score(\sigma, Peptide(\sigma, R))$, the ties in selecting $Peptide(\sigma, T \oplus R)$ are broken by a fair coin flip. Therefore, for the sake of simplifying the following analysis, we assume that there are no ties to begin with: $Score(\sigma, Peptide(\sigma, T)) \neq Score(\sigma, Peptide(\sigma, R))$. Under this assumption

$$\{Score(\sigma, T \oplus R) > t\} \cap \{Peptide(\sigma, T \oplus R) \in R\}$$
$$= \{Score(\sigma, R) > \max(t, Score(\sigma, T))\}.$$

Given the input set of spectra Σ, the databases R and T, and the threshold t, the total number of decoy discoveries, $DD(\Sigma, T \oplus R, t)$, is an observable RV (random variable) which is defined on the set of all random databases (the target T is a parameter here). We can calculate it as follows

$$DD(\Sigma, T \oplus R, t) = \sum_{\sigma \in \Sigma} 1_{Score(\sigma,R) > \max(t, Score(\sigma,T))},$$

where the indicator RV 1_A is equal to 1 iff the event A occurred. Another observable RV is the total number of discoveries

$$D(\Sigma, T \oplus R, t) = \sum_{\sigma \in \Sigma} 1_{Score(\sigma, T \oplus R) > t}$$

Elias and Gygi, 2007 [1] estimate the FDR as

$$\widehat{FDR}_{TDA} = \frac{2 \cdot DD(\Sigma, T \oplus R, t)}{D(\Sigma, T \oplus R, t)}. \quad (1)$$

It is important to note that being a ratio of two RVs, this estimate is itself an RV which varies with the draw of the random decoy set R. That is, this estimation of the FDR awkwardly depends on a virtual coin toss. Acknowledging this issue, many studies attempted to analyze the variability of TDA estimates [19].

Alternatively, we can account for the inherent variability by drawing many random decoy sets R and average over all the estimated FDRs using (1). Ignoring the computational costs, as we average over many such draws of R our average will converge to

$$E\left(\widehat{FDR}_{TDA}\right) = E\left[\frac{2 \cdot DD(\Sigma, T \oplus R, t)}{D(\Sigma, T \oplus R, t)}\right], \quad (2)$$

where the expectation is with respect to the random decoy R. Thus, if we know how to compute the above expectation we can save ourselves the costs of going through the averaging process. In doing so we would of course take the coin toss factor out of our estimation of the FDR.

While exactly computing the RHS (right hand side) of equation 2 can be challenging, if the size of the spectra set Σ is rather large we can rely on the common approximation of the mean of ratio as the ratio of means (e.g., Storey and Tibshirani 2003 [20]):

$$E\left(\widehat{FDR}_{TDA}\right) \approx \frac{E[2 \cdot DD(\Sigma, T \oplus R, t)]}{E[D(\Sigma, T \oplus R, t)]}. \quad (3)$$

We next describe how we can often compute the numerator and denominator on the RHS of (3).

False Positive Rate and eTDA:Is TDA Needed if Accurate FPRs are Available?

We denote by $FPR(\sigma, t)$ the probability that a spectrum σ matches a *random peptide*[9] with a score equal to or

[7] See Supplement C for a discussion about Δ-scores.

[8] The analysis below is streamlined if we use a strict inequality > with respect to the threshold t above (though a weak inequality ≥ works just as well).

[9] See Supplement E for the definition of a random peptide.

exceeding t. For an arbitrary scoring function, the exact value of FPR can always be computed in exponential time by simply generating and scoring all peptides with a given precursor mass. Starting from [6], many authors attempted to approximate FPRs of some scoring functions but there are concerns whether these approximations are accurate [14, 21].[10] Alves and Yu [21] and Kim et al., 2008 [14] described a polynomial algorithm for computing FPR for any scoring function that can be represented as a dot-product of vectors (see Supplement D).

We denote by $FPR(\sigma, N, t)$ the probability that a spectrum σ matches a peptide from a random database of size N with a score equal to or exceeding t.[11] As discussed in [23].

$$FPR(\sigma, N, t) \approx 1 - (1 - FPR(\sigma, t))^N.$$

If we can efficiently compute $FPR(\sigma, N, t)$ then we can use those to compute the expectations in (3). Indeed,

$$\begin{aligned} E[DD(\Sigma, T \oplus R, t)] &= E\left[\sum_{\sigma\in\Sigma} 1_{Score(\sigma,R)>\max(t,Score(\sigma,T))}\right] \\ &= \sum_{\sigma\in\Sigma} E\left[1_{Score(\sigma,R)>\max(t,Score(\sigma,T))}\right] \\ &= \sum_{\sigma\in\Sigma} P[Score(\sigma, R) > \max(t, Score(\sigma, T))] \\ &= \sum_{\sigma\in\Sigma} FPR(\sigma, |R|, \max(t, Score(\sigma, T))). \end{aligned} \quad (4)$$

Similarly, with $TD(\Sigma, T \oplus R, t)$ denoting the total number of target discoveries (which is also a RV that depends on the choice of R):

$$TD(\Sigma, T \oplus R, t) = \sum_{\sigma\in\Sigma} 1_{Score(\sigma,T)>\max(t,Score(\sigma,R))},$$

we have

$$\begin{aligned} E[TD(\Sigma, T \oplus R, t)] &= \sum_{\sigma\in\Sigma} P[Score(\sigma, T) > \max(t, Score(\sigma, R))] \\ &= \sum_{\sigma\in\Sigma: Score(\sigma,T)>t} P[Score(\sigma, T) > Score(\sigma, R)] \\ &= \sum_{\sigma\in\Sigma: Score(\sigma,T)>t} [1 - FPR(\sigma, |R|, Score(\sigma, T))], \end{aligned} \quad (5)$$

where we have again used the assumption that $P[Score(\sigma, T) = Score(\sigma, R)] = 0$.

Clearly,

$$D(\Sigma, T \oplus R, t) = TD(\Sigma, T \oplus R, t) + DD(\Sigma, T \oplus R, t),$$

allowing us to summarize our eTDA estimator:

$$\widehat{FDR}_{eTDA} := \frac{2 \cdot E[DD(\Sigma, T \oplus R, t)]}{E[TD(\Sigma, T \oplus R, t)] + E[DD(\Sigma, T \oplus R, t)]}, \quad (6)$$

where $E[DD(\Sigma, T \oplus R, t)]$ and $E[TD(\Sigma, T \oplus R, t)]$ are computed using equations 4 and 5.

The expectations of $DD(\Sigma, T \oplus R, t)$, $TD(\Sigma, T \oplus R, t)$, and $D(\Sigma, T \oplus R, t)$ give us more robust estimation than using these RVs directly. For example, for a spectral dataset consisting of a single spectrum, $DD(\Sigma, T \oplus R, t)$ is either 0 or 1 telling little about the statistical significance of matches between the spectrum and the database. $E(DD(\Sigma, T \oplus R, t))$, on the other hand, is an important characteristic of the statistical significance of these matches. In particular, it eliminates the dependency of the reported results on the choice of the random databases.[12]

Elias and Gygi's [1] TDA estimator of the FDR (1) has an advantage over our eTDA estimator (6) in that it does not require us to compute *FPR*s. On the other hand, the fact that eTDA is robust against the variability associated with the draw of the decoy set R makes it more reliable especially when the spectra set Σ or the target T are small.

Use of TDA in many mass spectrometry papers amounts to constructing the curves showing how $TD(\Sigma, T \oplus R, t)$ depends on $DD(\Sigma, T \oplus R, t)$ for various values of t. Afterwards, a (somewhat arbitrary) parameter t is chosen and all spectra contributing to $TD(\Sigma, T \oplus R, t)$ are reported as *peptide identifications* with an acknowledgment that some of them (represented by the value $DD(\Sigma, T \oplus R, t)$) may be incorrect. For example, in many cases, the ratio of $DD(\Sigma, T \oplus R, t)$ and $D(\Sigma, T \oplus R, t)$ is computed and an informal statement like "we identified 1000 spectra with 1% error rate" is made. It is important to realize that such conclusions also depend on the choice of the database R and not only on the inherent properties of the spectral dataset. We therefore argue that a better approach would be to construct curves showing how $E[TD(\Sigma, T \oplus R, t)]$ depends on $E[DD(\Sigma, T \oplus R, t)]$ thus eliminating the random effects on the reported results.

[10] Continuous approximations of discrete random variables are notoriously inaccurate when one attempts to approximate the extreme tail of the distribution [22].

[11] While we consider random iid database, the results below can be easily generalized to databases generated by arbitrary Markov chains.

[12] It also frees the researchers from the need to estimate the variance of the random variable $DD(\Sigma, T \oplus R, t)$ as is often done to demonstrate that a single observation of a random variable in TDA does not result in large deviations from the expected value of this variable.

See Supplement F on how eTDA can also be used in lieu of TDA applied to the *separate* databases rather than the combined.

Can the Decoy and Target Databases Contain Shared Peptides?

In our definition of eTDA (6), we glossed over a certain practical issue that needs to be addressed. Specifically, eTDA hinges on our ability to efficiently compute the FPRs. However, the polynomial algorithm from [14] is not compatible with the TDA procedure in [1]. The latter requires that the decoy database should not contain any "sufficiently long" peptides from the target database—a requirement that cannot be satisfied by the probabilistic model in [14]. Should we therefore declare eTDA as impractical? Or, as we next argue, is the non-intersection restriction statistically flawed and should be removed from the TDA procedure?

Elias and Gygi [1] claim that the non-intersection is required to ensure that any decoy discovery is indeed a false one. Otherwise, they argue, we would overestimate the number of false discoveries, that is we would err on the conservative side. However, Elias and Gygi [1] themselves argue that in most practical cases the relevant intersection between a random decoy database and the target database is negligible with very high probability.

While we agree with Elias AND Gygi, 2007 [1] that a non-empty intersection between the target and decoy databases might overestimate the number of false discoveries, we argue that the empty intersection condition does not yield an unbiased estimator. Indeed, the identical peptides are not the only source of FDR "inflation" in a random database, the homeometric peptides [24] may lead to an even larger inflation of FDR. Roughly speaking, peptides π and π' are homeometric if their theoretical spectra are nearly identical (see [24] for precise definition). Homeometric peptides inflate FDRs in exactly the same way as the peptides with identical sequences and, thus, according to Elias and Gygi, 2007 [1] logic, should be excluded. However, the number of pairs of homeometric peptides is orders of magnitude larger than the number of identical peptides making the problem of constructing the random database very difficult, if not intractable. Supplement G describes how to shuffle a target database to generate a transposed database that does not share long peptides with a target database (as suggested in [1]) yet contains many homeometric peptides resulting in a highly inflated estimate of FDR. According to [1], the transposed database is as legal as the commonly used reversed database.

So far we argued that allowing the intersection should not have a significant detrimental impact. But beyond that it would allow us to considerably simplify our random decoy model so, for example, it could include the iid model. Conversely, we are not aware of any method to rigorously generate random decoy databases that duly respects the non-intersection restriction from [1].

For all these reasons we recommend that TDA be carried out irrespective of the intersection of the random and decoy databases. Finally, in Supplement H we show that one can readily estimate the effect the intersection between the target and the decoy databases could possibly have on the estimated number of false discoveries.

FPR-Compliant Versus FPR-Noncompliant Scoring Functions

While FDR is a useful parameter, computing FDR via TDA suffers from restrictions on the minimal sizes of the spectral datasets and protein databases and requires an estimate of the variance of FDR that is often unavailable. In addition, all previously proposed methods for computing FPRs are faster than TDA [14]. We therefore argue that TDA is not needed when accurate FPRs are available. We are not claiming that FDR is not needed when FPRs are available, rather that it is better estimated using FPRs. We remark that decoy databases were also used in the early days of pre-BLAST genomics [25–27].

In particular, the FASTA tool (a predecessor of BLAST) included a program that generated a decoy database and computed z-values [28]. This approach was abandoned in favor of BLAST, mainly because BLAST enabled p-value (i.e., FPR) computations. We remark that BLAST employs a simple scoring function (that can be represented as a dot product of vectors encoding the amino acid sequences).

There exist two (often conflicting) criteria to scoring functions that are equally important in practice: an ability to accurately assess a match and the ability to efficiently compute its p-values. One can invent a more complex (and arguably better) scoring function than the one used in BLAST, e.g., by accounting for hydrophobicity of the amino acid sequences. While such changes may be potentially useful, the genomics community resisted them since they jeopardize the ability to compute p-values. We therefore raise the question whether some complex scoring functions in mass spectrometry provide a reasonable trade-off between the assessment of PSMs and the ability to efficiently compute p-value of PSMs.

We call a scoring function $Score(\sigma,\pi)$ FPR-compliant if there exist a polynomial algorithm for computing $FPR(\sigma,\pi)$ and we say that it is FPR-noncompliant if no polynomial algorithms for computing FPR are known.[13] For example, if one transforms a spectrum σ and a peptide π into vectors $\vec{\sigma}$ and $\vec{\pi}$ and defines $Score(\sigma,\pi) = \vec{\sigma}\cdot\vec{\pi}$ as a dot-product, then p-values can be easily computed [14] (thus all such scoring functions are FPR-compliant). These p-values represent a new (and better) scoring function –p-value(σ, π) that greatly

[13]We emphasize that computing precise FPR (as in [14]) is not a requirement for an FPR-compliant scoring function. Any tool that *accurately* approximates FPR (e.g., through the extreme value distribution approximation) should be viewed as FPR-compliant.

increases the number of identified PSMs as compared to the original scoring function $Score(\sigma, \pi)$ (for a fixed FDR). There is no convincing evidence yet that FPR-non-compliant scoring functions used in tools like Sequest, Mascot, and InsPecT result in better assessment of PSMs than p-values of a simple dot product (e.g., the MS-GF scoring from [14]). On the contrary, Kim et al., 2010 [29] recently demonstrated that MS-GF outperforms Sequest, Mascot, OMSSA, X!Tandem, and InsPecT for all CID, ETD, and CID+ETD spectral datasets they analyzed. Moreover, these scoring function have an inferior performance when applied to very large databases [14] or to spectra generated using new MS technologies.[14] It raises the question whether introducing complex components into the scoring functions employed in MS/MS searches is fully justified.

Are Individual FPRs Useful in the Context of High-Throughput Mass Spectrometry?

There is a common opinion that practicing mass-spectrometrists do not need FPRs since FDRs are sufficient to summarize the results of a high-throughput proteomics experiment. Indeed, if the goal of an experiment is to output a list of confidently identified PSMs (e.g., with 1% FDR), it appears that FDR is sufficient. However, generating a list of identified peptides is hardly ever the final goal of proteomics studies: biologists are usually interested in expressed proteins. Since TDA does not provide information on which PSMs are more statistically significant than others (as discussed, PSM scores are not very well correlated with their FPRs), biologists are left in the dark deciding which proteins are reliably expressed. It leads to awkward (statistically unsubstantiated) heuristics like the "two peptide rule" [23] that discard expressed proteins that are supported by only a single, even if reliable, PSM with low FPR.

As another example, in a proteogenomics study, a single PSM often provides evidence for a new translation start site or a new splicing variant. Knowing FPRs in such cases is paramount for verifying new biological discoveries and ranking them for further experimental verification. FPRs of individual PSMs may vary by 10 orders of magnitude in typical MS/MS searches and a PSM with FPR of 10^{-20} provides a much more reliable evidence for a new start site than an (otherwise respectable) PSM with FPR of 10^{-10}. For example, one can find evidence for a new gene supported by a single PSM with FPR 10^{-20} and yet another piece of evidence for a new gene supported by two PSMs with FPRs 10^{-10}. Knowing FPRs of individual PSMs is important since there may be hundreds of new PSM-supported genes [30] that have to be ranked for further experimental validation. TDA does not provide a way to generate such ranking while FPR does allow one to differentiate between reliable and unreliable proteogenomics evidence.

Another application that requires FPRs is identification of rare modifications, e.g., N-myristoylation in eukaryotes or N-acetylation in prokaryotes. While these modifications are crucially important for the cell, very few N-myristoylated proteins in eukaryotes or N-acetylated proteins in prokaryotes have been identified so far. Since each newly found N-myristoylated or N-acetylated peptide has to be individually evaluated (and possibly experimentally verified), FPR rather than FDR is needed in such studies. For example, currently only five N-acetylated proteins are documented in *E. coli*, an undoubtedly incomplete list [31]. If one analyzed a million spectra and identified a 6th N-acetylated protein in *E. coli* with 1% FDR, would it be a good reason to write a paper about a new N-acetylated protein and discuss its biological function? We are afraid that if the answer to this question is "yes" (as is often presumed in proteomics studies) we may have too many statistically unsubstantiated "discoveries." Indeed, 1% FDR tells us little about the statistical significance of a single PSM that led to this conclusion!

We emphasize that FPR represents the statistical significance of a single PSM and should not be used for assessing the statistical significance of observing a PSM with a given score in the context of comparing an entire spectral dataset against a protein database. Above we described how to correct FPR of a single PSM for the size of the protein database. However, an FPR that may appear significant while analyzing a dataset of 1 million spectra may become statistically insignificant in the context of a dataset with 100 million spectra [32, 33]. See Kall et al., 2008 [15] for a q-value based correction for both size of the protein database and the size/structure of the spectral dataset.

A Pandora Box of Database-Dependent Scoring Functions

We previously defined a scoring function as a black box that, given a spectrum σ and a peptide π, outputs

$$Score(\sigma, \pi)$$

which depends only on σ and π. However, the scoring function of many MS/MS tools (e.g., Sequest, Mascot, InsPecT, Percolator, etc.) are affected not only by σ and π but also by the database DB that is being searched and, in some cases, even by the spectral dataset Σ. For example, Sequest computes a δ-score which depends on the protein database that is being searched. Given σ, π, and DB, such tools compute database-dependent scoring function:

$$Score(\sigma, \pi, DB)$$

An MS/MS database search algorithm in this case is a black box that, given σ and DB, outputs a peptide $\pi(\sigma, DB)$ from the

[14] E.g., these scoring functions are inferior to the scoring functions employed in de novo peptide sequencing when searching the database of all possible peptides. We remark that most scoring functions used in de novo peptide sequencing can be represented as dot products.

database DB with maximum $Score(\sigma, \pi(\sigma, DB), DB)$ among all peptides in the database. We define

$$Score(\sigma, DB) = Score(\sigma, \pi(\sigma, DB), DB)$$

Although we illustrated that applying TDA to tools T^+ or to the original version of Percolator is ill-advised, the notion of non-TDA-compliant tool has not been defined yet. Moreover, we are not aware of a formal test that reveals whether or not a particular tool is TDA-compliant. In the absence of such a test, the only tools one can safely claim are TDA-compliant might be the single-stage tools with scoring functions $Score(\sigma, \pi)$ depending only on σ and π.

In Supplement I, we design a database-dependent scoring function that outputs unreliable identifications with excellent FDRs (evaluated by TDA). This result underscores the danger of using database-dependent scoring functions and raises the concern of whether some MS/MS tools feature respectable FDRs while generating some low-quality PSMs (see Supplement C). This question is far from being theoretical since the original version of Percolator and some other tools indeed fell into this trap [4, 5].

Presumably, there exists a division between TDA-compliant scoring functions and non-TDA-compliant ones but the question of how to decide whether a particular database-dependent scoring function is TDA-compliant remains open. We emphasize that the test for TDA-compliance has to be rigorous: i.e., one should be able to subject the original version of Percolator or the existing version of Sequest to a test that analyzes these tools as black boxes and automatically decides whether they are TDA-compliant.[15]

Some non-TDA-compliant tools (see Supplements A and J) were shown to produce peptide identifications leading to biologically important conclusions verified by other approaches (e.g., comparative genomics). We are not discarding these tools as useless but simply stating that they are not TDA-compliant and thus there is currently no way to evaluate their performance and compare to other tools. Similarly, it remains unclear whether database-dependent scoring functions employed in Sequest and other popular tools are TDA-compliant, yet these tools are useful. We therefore argue that the popular MS/MS tools with database-dependent scoring functions should be analyzed for TDA compliance before they are widely used in conjunction with TDA.

Conclusions

One of the reason TDA became so popular in proteomics (as opposed to genomics) is because many popular MS/MS database search tools use FPR-noncompliant scoring functions making it difficult to compute accurate p-values. For such scoring functions, TDA should still be considered; however, these scoring functions were designed at the time when the issue of statistical significance of peptide identifications was often ignored and the notion of PSM p-values was not even defined. We feel that the time has come to critically evaluate the drawbacks of FPR-noncompliant scoring functions and TDA and to ensure that mass spectrometrists have access to accurate p-values for all peptides they identify. We also stress that there are alternative methods for estimating the FDR, notably the q-value approach of Kall et al. 2008 [32], which are not discussed here.

[15]An "expert opinion" based on analyzing the source code (let alone an algorithmic description) of a particular tool does not qualify as a rigorous approach to deciding whether a particular tool is TDA-compliant.

Acknowledgments

The authors are indebted to Sangtae Kim, Lukas Kall, Michael MacCoss, and William Noble for many discussions and invaluable feedback without which this manuscript would not be feasible. This work was supported by US National Institutes of Health grant 1-P41-RR024851 from the National Center for Research Resources.

References

1. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207–214 (2007)
2. Balgley, B.M., Laudeman, T., Yang, L., Song, T., Lee, C.S.: Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol. Cell. Proteom.* **6**, 1599–1608 (2007)
3. Nesvizhskii, A.I., Vitek, O., Aebersold, R.: Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods* **4**, 787–797 (2007)
4. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., MacCoss, M.J.: Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4**, 923–926 (2007)
5. Spivak, M., Weston, J., Bottou, L., Kall, L., Noble, W.S.: Improvements to the Percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.* **8**, 3737–3745 (2009)
6. Craig, R., Beavis, R.C.: A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom.* **17**, 2310–2316 (2003)
7. Everett, L.J., Bierl, C., Master, S.R.: Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies. *Journal of Proteome Research* **9**, 700–707 (2010)
8. Tharakan, R., Edwards, N., Graham, D.R.M.: Data maximization by multipass analysis of protein mass spectra. *Proteomics* **10**, 1160–1171 (2010)
9. Nesvizhskii, A.I.: A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics* **73**, 2092–2123 (2010)
10. M.W. Bern and Y.J. Kil. Comment on "Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies". *Journal of Proteome Research*, (in press, available online), 2011.
11. Bern, M., Phinney, B.S., Goldberg, D.: Reanalysis of Tyrannosaurus rex Mass Spectra. *J. Proteome Res* **8**, 4328–4332 (2009)
12. S. Kim and N. Bandeira. False discovery rates in spectral identification. *Methods Mol Biol.*, (in press), 2011.
13. Kandasamy, K., Pandey, A., Molina, H.: Evaluation of Several MS/MS Search Algorithms for Analysis of Spectra Derived from Electron Transfer Dissociation Experiments. *Analytical Chemistry* **81**, 976–989 (2009)
14. Kim, S., Gupta, N., Pevzner, P.A.: Spectral Probabilities and Generating Functions of Tandem Mass Spectra: a Strike Against Decoy Databases. *J. Proteome Res.* **7**, 3354–3363 (2008)

15. Kall, L., Storey, J.D., MacCoss, M.J., Noble, W.S.: Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **7**, 29–34 (2008)
16. Karlin, S., Altschul, S.F.: Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 2264–2268 (1990)
17. Kim, S., Gupta, N., Banderia, N., Pevzner, P.A.: Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics* **8**, 53–69 (2009)
18. Guibas, L.J., Odlyzko, A.M.: String overlaps, pattern matching, and nontransitive games. *J. Comb. Theory, Ser. A* **30**, 183–208 (1981)
19. Huttlin, E.L., Hegeman, A.D., Harms, A.C., Sussman, M.R.: Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J. Proteome Res* **6**, 392–398 (2007)
20. Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–45 (2003)
21. Alves, G., Yu, Y.K.: Statistical Characterization of a 1D Random Potential Problem–with applications in score statistics of MS-based peptide sequencing. *Physica A: Statistical Mechanics and its Applications* **387**, 6538–6544 (2008)
22. Altschul, S.F., Gish, W.: Local alignment statistics. *Methods in enzymology* **266**, 460–480 (1996)
23. Gupta, N., Pevzner, P.A.: False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* **8**, 4173–81 (2009)
24. Frank, A.M., Savitski, M.M., Nielsen, M.L., Zubarev, R.A., Pevzner, P. A.: De novo peptide sequencing and identification with precision mass spectrometry. *J. of Proteome Research* **6**, 114–123 (2007)
25. Dayhoff, M.O., Barker, W.C., Hunt, L.T.: Establishing homologies in protein sequences. *Methods in enzymology* **91**, 524–45 (1983)
26. Feng, D.F., Johnson, M.S., Doolittle, R.F.: Aligning amino acid sequences: comparison of commonly used methods. *Journal of Molecular Evolution* **21**, 112–125 (1985)
27. Lipman, D.J., Wilbur, W.J., Smith, T.F., Waterman, M.S.: On the statistical significance of nucleic add similarities. *Nucleic acids research* **12**, 215–226 (1984)
28. Lipman, D.J., Pearson, W.R.: Rapid and sensitive protein similarity searches. *Science* **227**, 1435–41 (1985)
29. Kim, S., Mischerikow, N., Bandeira, N., Navarro, J.D., Wich, L., Mohammed, S., Heck, A.J.R., Pevzner, P.A.: The generating function of CID, ETD and CID/ETD pairs of tandem mass spectra: Applications to database search. *Molecular & Cellular Proteomics* **9**, 2840–52 (2010)
30. Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V., Briggs, S.P.: Discovery and revision of Arabidopsis genes by proteogenomics. *Proceedings of the National Academy of Sciences* **105**, 21034–21038 (2008)
31. Polevoda, B., Sherman, F.: The diversity of acetylated proteins. *Genome Biol* **3**, 1–0006 (2002)
32. Kall, L., Storey, J.D., Noble, W.S.: Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **24**, i42–8 (2008)
33. Choi, H., Nesvizhskii, A.I.: False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **7**, 47–50 (2008)