Supplementary Materials

Supplement A: Paired-peptide approach for improving the performance of any given MS/MS search tool

In the main text of this manuscript, we described how one can design a tool T^+ that doubles the number of peptide identifications by introducing spurious identifications in a particular fashion that exploits TDA. T^+ is clearly undesirable in practice, since we know that many of its identifications are unacceptable.

Below we describe a more reasonable approach, the *paired-peptide* approach, for transforming a given search tool T into a new tool T^* that, unlike T^+, is well-intentioned and improves the level of peptide identifications and biological discovery from a given spectral dataset. We indeed designed this approach genuinely for improving peptide identifications from search tools, and only later realized that it is not TDA-compatible. It is described here to illustrate that precaution is needed when using TDA even with well-intentioned and biologically useful approaches.

Design of T^*

We first introduce the notion of *paired peptides*. Given a set X of peptide identifications (each peptide in X forms a PSM) and a parameter $Distance$, we call a peptide $P \in X$ paired if there is another peptide in X separated from P by less than $Distance$ amino acids within the same protein (i.e., the starting positions of these peptides are less than $Distance$ amino acids apart). We denote all peptides identified by a database search tool T with scores above $Score$ as $T(Score)$ and the set of all paired peptides as $T(Score, Distance)$. Supplementary Table 1 shows how the number of paired peptides $InsPecT(Score, Distance)$ in target and decoy databases (identified by InsPecT) varies depending on $Score$ and $Distance$. The table illustrates that for a fixed 1% FDR, the number of identified peptides remains stable for various values of $Distance$ (varying from maximum $\approx 42,000$ peptides at $Distance = 70$ to $\approx 40,000$ peptides at $Distance = 500$ and larger). It indicates that the method is robust with respect to parameter $Distance$ (we use the default value $Distance = 100$). Without the paired peptide concept (i.e., in the standard MS/MS database search setup), the number of the identified peptides at 1% FDR is significantly lower ($\approx 35,000$ peptides at 1% FDR).

Instead of fixing the parameter $Distance$, one can vary it for different proteins, for instance, setting it equal to the protein length (equivalent to setting $Distance = \infty$). This would, however, give an unfair advantage to long proteins and reduce the number of identified peptides (compared to a fixed $Distance = 100$) by $\approx 4\%$ (Supplementary Table 1). Also, in proteogenomics studies requiring search in the six-frame translation databases, the proteins are often not known in advance making the paired-peptide approach the only feasible option.

Given any tool T, we can build a new tool T^* that only reports paired-peptides, such that $T^*(Score) = T(Score, 100)$. Next, we compare the performance of T^* with T using various tools and datasets and show that T^* consistently performs better.

MS/MS datasets

The MS/MS datasets used in this study were obtained from *Shewanella oneidensis* MR-1 (generated in Dick Smith's lab at PNNL) and human (generated in Vivian Hook's lab at UCSD) samples. The *Shewanella* dataset has been previously studied in [1, 2], while the human dataset has been described in [3]. The datasets were generated on Thermo LCQ mass-spectrometer for *Shewanella* and Agilent XCT Ultra for human samples. The human dataset includes nearly 600 thousand spectra, while the *Shewanella* dataset includes 14.5 million spectra. The *Shewanella* dataset was searched against the *Shewanella* protein database (size $\approx$ 1.5 Mb, containing 4,928 sequences), and the human dataset was searched against the IPI database version 3.41 (size $\approx$ 40 MB, containing 72,155 sequences). Decoy databases were generated by randomly shuffling the sequence of each protein in the target database (preserving the background amino acid frequencies for each protein).

Database searches were carried out using InsPecT [4]. InsPecT searches were run with the default parameter settings (fragment ion tolerance of 0.5 Da and parent mass tolerance of 2.5 Da). MS-GF [5] was run on InsPecT results to evaluate the statistical significance of peptide identifications (*spectral probabilities*). We treat the spectral probability as a score, and since MS-GF was used to rescore all InsPecT identifications (including even the very low scoring ones that are typically never reported), $InsPecT \oplus MS-GF$ can be viewed as a hybrid peptide identification tool.

T^* improves peptide identifications and enables biological discoveries

In both the human and the *Shewanella* datasets, we find that the reduction in the number of peptide identifications by limiting attention to paired peptides is more pronounced in the decoy database compared to the target database. This results in a better sensitivity-specificity trade-off for paired peptides, as shown in Figure 1. The curves in the figure are plotted by computing the number of identifications in the target and the decoy database for varying score thresholds. For example, for the same number (14) of decoy database hits, InsPecT identifies 2,011 peptides in the target human database using traditional approach, and 2,577 peptides using the paired-peptide approach, a 28% increase. Similarly, the number of peptide identifications obtained from MS-GF [5] increases by 12%. These observations are replicated in the *Shewanella* dataset (Figure 1b), indicating that the paired-peptide approach is widely applicable.

Next, we illustrate with some examples how the paired-peptide approach improves proteogenomic analysis.

Improved detection of proteolytic cleavage sites

The increased number of peptide identifications is useful for various biological applications. For example, signal peptides are difficult to confirm experimentally, and computational predictions are used to fill the gap. There have been some concerns regarding the quality of signal peptide predictions [6] since these tools make predictions based on a rather small database of experimentally validated signal peptides. MS/MS serves as a high-throughput technique for identification of cleavage sites of signal peptides and can thus be used to increase the confidence of signal peptide predictions. Using the same approach as in [1] on the larger set of peptides identified by the paired-peptide approach at 1% FDR in the *Shewanella* dataset, we identified 180 putative signal peptides, significantly more than the previously reported number (117) [1].

In Gupta et al, 2008 [2], it was shown that comparative proteogenomic analysis can be used to identify a reliable set of proteolytic cleavages that are conserved between multiple species [2]. We find that the paired-peptide approach for *Shewanella oneidensis* MR-1 allows us to recover 41 conserved proteolytic sites among the three *Shewanella* species (Supplementary Table 2), significantly more than the previous study (31).

Identification of novel peptides in a six-frame translation database search

We further demonstrate that the paired-peptide approach allows improvements in the identification of peptides in previously unannotated regions of the genome in a six-frame translation database search, allowing for corrections in gene annotations and finding new genes. Six-frame translations of the whole genome are commonly used as the sequence database in proteogenomic studies for improving genome annotations such as correcting gene-start sites, identifying novel genes, programmed frameshifts, sequencing errors etc. [1]. These predictions are enabled by identification of *novel* peptides that fall outside the annotated protein-coding regions of the genome. Gupta et al., 2007 [1] used stringent FDR control to corroborate novel peptides (also validated using comparative genomics and DNA sequencing trace analysis) and demonstrated that many of these novel peptides reveal gene annotation errors. Therefore, increasing the number of novel peptides is helpful in obtaining better genome annotations. Below we show that the paired-peptide approach increases the number of novel *Shewanella* peptides by more than 50%.

Supplementary Table 3A provides the number of peptide identifications in the six-frame translation of *Shewanella* using InsPecT [4] with MS-GF scores [5]. The table shows that the paired-peptide approach leads to a significant increase in the number of identified novel peptides as compared to the traditional approach. For example, at a very stringent 0.1% FDR, 198 novel peptides are identified by the traditional approach, while 320 novel peptides are identified using the paired-peptide approach, a 62% increase in the number of identifications. It must be noted that the gene correction methods often limit attention to only those novel peptides that are located close to other peptides, to have confidence in the corrections [1]. This implies that a large fraction of novel peptides identified in the traditional approach are not usable for gene corrections. On the other hand, all novel peptides identified with the paired-peptide approach can be used for gene predictions/corrections.

At MS-GF score threshold of 12.3, the number of peptide hits in the decoy database becomes 0 with the paired-peptide approach, while 181 novel peptides are identified in the target database. Supplementary Table 3B lists these 181 virtually "error-free" peptides and shows their distances from the immediately downstream gene in the same frame (a small distance may indicate an N-terminal extension of the gene), and from the immediately upstream gene in any frame on the same strand (a small distance may reflect a programmed frameshift or a DNA sequencing error). We indeed find that these distances are small for many peptides, suggesting that these peptides can be very useful in correcting gene annotations. Moreover, many of these peptides either start from Methionine (Start codon) or start immediately after Methionine (N-terminal Methionine cleavage [1]) increasing the confidence that these novel peptides provide a solid basis for correcting gene annotations. A large distance from neighboring genes, on the other hand, may indicate that the peptides come from a novel gene that was missed in the original annotation. A detailed study of these novel peptides to analyze their implied gene corrections is beyond the scope of this paper.

T^* is not TDA-compliant

Even though using paired-peptide seems to be useful, the improvement in performance seen in Figure 1 is controversial, since T^* is not actually compatible with TDA as was the case with T^+. For example, if we insert a bogus peptide in the database, it will have a higher chance of getting identified in the target database (because of the presence of a larger number of neighboring peptides) as compared to the decoy database. Therefore, the score distribution of incorrect peptides in the target database is not similar to the score distribution of peptides in the decoy database. Thus, using TDA to establish the superiority of T^* is illegitimate.

Supplement B: TDA with inconsistent scoring functions robs researchers from statistically significant peptide identifications

For a spectrum σ and peptide π we define $FPR(\sigma, \pi)$ as $FPR(\sigma, Score(\sigma, \pi))$. We call a scoring function $Score(\sigma, \pi)$ *consistent* if

$$Score(\sigma_1, \pi_1) \geq Score(\sigma_2, \pi_2)$$

implies

$$FPR(\sigma_1, \pi_1) \leq FPR(\sigma_2, \pi_2).$$

The scoring functions of nearly all existing MS/MS database search tools (Sequest, Mascot, X!Tandem, InsPecT, etc.) are inconsistent [5, 7].[1] However, any scoring function $Score(\sigma, \pi)$ can be transformed into a new scoring function $-FPR(\sigma, \pi)$. While this transformation can always be accomplished in exponential time, efficient polynomial-time algorithms for converting scores into FPRs remain unknown for most existing tools. One can show that a TDA with an inconsistent scoring function $Score(\sigma, \pi)$ can be substituted by a TDA with a new scoring function (based on -$FPR(\sigma, \pi)$) that improves on the results of the original approach.

Supplement C: Are tools using Δ-scores TDA-compliant?

While some database-dependent scoring function are clearly non-TDA-compliant, we are not claiming that all database-dependent scoring functions (and particularly the ones utilizing δ-scores) are non-TDA-compliant. However, we are not aware of a study that proves that the specific scoring functions employing δ-scores are TDA-compliant. We further remark that computing δ-scores is an attempt to estimate expectation of a random variable $\max_{Peptide \in R} Score(\sigma, Peptide)$ defined on all random databases of the same size as T. If individual FPRs are available, this expectation can be computed precisely.

Supplement D: Dot-product scoring functions

One can transform spectra and peptides into vectors and further use their dot-product for scoring PSMs.

We represent each amino acid as an m-dimensional boolean string with $m - 1$ zeros and a single one in the end (m represents the integer mass of amino acid). A peptide is represented as a

[1] The use of δ-scores in some algorithms (e.g., Sequest) is merely an attempt to improve the consistency of their scoring functions. Indeed, these approaches are based on the observation that better PSMs with $FPR(\sigma_1, \pi_1) < FPR(\sigma_2, \pi_2)$ typically have larger δ-scores: $\delta(\sigma_1, \pi_1) > FPR(\sigma_2, \pi_2)$. Thus, boosting original scores with δ-scores (e.g., adding them) results in a more consistent scoring function.

string (vector) obtained by concatenation of its amino acids. Many MS/MS tools use PRM (prefix residue mass) representation of tandem mass spectra. $PRMspectrum$ is a vector $p_1 \dots p_t$ where p_i represents the likelihood that the peptide (that generated this spectrum) has prefix mass i [4, 8, 5]. Since both peptides and PRM spectra represent vectors, one can use dot-product to score PSMs.

Supplement E: What is a random peptide and a random database?

The iid random databases used in TDA [9] assume that the probability of finding a peptide P starting in an arbitrary position of the database equals to $\frac{1}{20^{|P|}}$, where $|P|$ is the length of the peptide.[2] Our notion of random peptides follows this model and assumes that the probability of a peptide P is $\frac{1}{20^{|P|}}$. However, in this case, the probabilities of all peptides do not sum up to 1 suggesting that they should be all normalized. Below we describe a different (but equivalent) approach to introducing the notion of random peptide that does not require normalization.

The notion of random peptide needs to be defined with caution since in mass spectrometry it is usually assumed that a peptide can match a spectrum iff its mass equals to the precursor mass of the spectrum (all other matches are scored as $-\infty$). We will find it convenient to extend this definition and to say that a (long) peptide *LongPeptide fits* a spectrum if a prefix of this peptide has a mass equal to the precursor mass of the spectrum (such prefix is denoted as $Peptide(LongPeptide, \sigma)$. We further consider a set of all 20^n long peptides (for a large n) and assign each peptide in this set a probability of $1/20^n$. The score between a spectrum σ and a $LongPeptide$ from this set is defined as $Score(\sigma, Peptide(LongPeptide, \sigma))$ if $LongPeptide$ fits σ. When we refer to a random peptide P, we mean all long peptides that have P as a prefix. The total probability of all such long peptides is indeed $\frac{1}{20^{|P|}}$.

When we refer to a random database, we mean an iid amino acid sequence.

Supplement F: eTDA applied to separate databases

Note that eTDA can also be used in lieu of TDA applied to the *separate* databases rather than the combined. Essentially, we only need to replace $DD(\Sigma, T \oplus R, t)$ with $DD(\Sigma, R, t)$ and $TD(\Sigma, T \oplus R, t)$ with $TD(\Sigma, T, t)$. Note that the latter no longer depends on R and is therefore a parameter of the model rather than an RV. In this case eTDA would be modified from equation (6) to

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

where instead of equation (4) we would use

$$E\left[DD(\Sigma, R, t)\right] = \sum_{\sigma \in \Sigma} FPR(\sigma,\ |R|,\ t).$$

[2] This approach to generating random databases can be easily generalized to account for various frequencies of amino acids or Markovian dependencies between consecutive amino acids.

Supplement G: Transposed database

Given a target database $p_1 \dots p_N$, we define a *transposed database* that swaps different amino acids $p_i \neq p_{i+1}$ for all $i = 0$ (modulo 7). If $p_i = p_{i+1}$, amino acid p_i is substituted by another randomly chosen amino acid. It is easy to see that the transposed database does not share peptides longer than 7 aa with the target databases (at the same positions) and thus satisfies the condition specified in [9] (target and decoy databases do not share long peptides). Nevertheless, one can see that the transposed database will result in a highly inflated estimate of FDR since it contains many peptides that are homeometric to peptides in the target database.

Supplement H: How the intersection between the target and the decoy databases affect FDR?

We can readily estimate the effect the intersection between the target and the decoy databases could possibly have on the estimated number of false discoveries. Indeed, let $SD(\Sigma, T \oplus R, t)$ denote the number of discoveries that are shared between the target and decoy that are counted as decoy discoveries:

$$SD(\Sigma, T \oplus R, t) = \sum_{\sigma \in \Sigma} 1_{\{\exists \pi \in T \cap R : Score(\sigma, \pi) = Score(\sigma, T \oplus R)\} \cap \{Peptide(\sigma, T \oplus R) \in R\}}.^{3} \tag{1}$$

If we follow the rationale of [9] then the number of decoy false discoveries is conservatively estimated between $DD(\Sigma, T \oplus R, t)$ and $DD(\Sigma, T \oplus R, t) - SD(\Sigma, T \oplus R, t)$[4]. Thus, examining the ratio of $SD(\Sigma, T \oplus R, t)/DD(\Sigma, T \oplus R, t)$ we can verify that the effect of shared peptides is indeed negligible before accepting the analysis. Similarly, we can examine the effect of the non-empty intersection on eTDA by comparing $E[SD(\Sigma, T \oplus R, t)]$ with $E[DD(\Sigma, T \oplus R, t)]$ where

$$\begin{aligned} E[SD(\Sigma, T \oplus R, t)] &= \frac{1}{2} \cdot \sum_{\sigma \in \Sigma : Score(\sigma, T) > t} P[Score(\sigma, R) = Score(\sigma, T)] \\ &\approx \frac{1}{2} \cdot \sum_{\sigma \in \Sigma : Score(\sigma, T) > t} [FPR(\sigma, |R|, Score(\sigma, T) - \delta) - FPR(\sigma, |R|, Score(\sigma, T))], \end{aligned}$$

and δ is some small number, say, 10^{-8}. Note that here we obviously do not assume that

$$P[Score(\sigma, R) = Score(\sigma, T)] = 0.$$

[3] We assume the function $Peptide(\sigma, T \oplus R)$ randomly selects R or T in this case of a shared discovery.

[4] A more sophisticated estimation can be used here, for example, by replacing SD with $(1 - \alpha) \cdot SD$ where α is the initial conservative estimate of the FDR (equation 1 in main text)

Supplement I: Database-dependent scoring functions may output unreliable identifications with excellent FDRs

We show that the use of database-dependent scoring functions opens a Pandora box that allows one to design algorithms that output unreliable identifications with excellent FDRs (evaluated by TDA). Indeed, given an arbitrary database-independent scoring function $Score(\sigma, \pi)$, one can design a database-dependent-scoring function $Score(\sigma, \pi, DB)$ as follows. Let π_1 and π_2 be the best scoring and the second-best scoring peptides in the database DB for a database-independent scoring function $Score(\sigma, \pi)$. As usual, the PSM (σ, π_1) is expected to be correct while PSM (σ, π_2) is expected to be bogus. We construct a database-dependent scoring function $Score(\sigma, \pi, DB)$ that is equal to the database-independent $Score(\sigma, \pi)$ for all peptide except for π_2 and define
$Score(\sigma, \pi_2, DB) = Score(\sigma, \pi_1, DB) + \epsilon$, where ϵ is a small constant. As a result, the second-best scoring peptide for the database-independent scoring function $Score(\sigma, \pi)$ becomes the best-scoring peptide for the database-dependent scoring function $Score(\sigma, \pi, DB)$.

Note that the constructed database-dependent scoring function produces bogus peptide identifications (since PSMs (σ, π_2) are bogus) yet results in virtually the same FDR (presumably good) as the database-independent scoring function. Moreover, it is easy to construct a (bogus) database-dependent scoring function that greatly improves on the FDR (evaluated by TDA) of the original database-independent scoring function!

Supplement J: Percolator

The Percolator algorithm introduced in Kall et al., 2007 [10], used the number of identified peptides in a protein to score individual PSMs and thus was not TDA-compatible. Recently, Spivak et al., 2009 [11]. released a new version of Percolator with an acknowledgment of this shortcoming and wrote: “In this work, we removed three features that exploit protein-level information because of the difficulty of accurately validating, via decoy database search, methods that use this type of information.”

References

[1] N. Gupta, S. Tanner, N. Jaitly, J. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. Smith, and P. Pevzner. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res*, 17:1362–1377, 2007.

[2] N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, N. Nguyen, N. Ollikainen, J. Rodriguez, J. Wang, M.S. Lipton, M. Romine, V. Bafna, R.D. Smith, and P.A. Pevzner.

Comparative Proteogenomics: Combining Mass Spectrometry and Comparative Genomics to Analyze Multiple Genomes. *Genome Res.*, 18:1133–1142, 2008.

[3] N. Gupta, S.J. Bark, W.D. Lu, L. Taupenot, D.T. O'Connor, P.A. Pevzner, and V. Hook. Mass Spectrometry-Based Neuropeptidomics of Human Secretory Vesicles from Adrenal Medullary Pheochromocytoma Reveals Novel Peptide Products of Prohormone Processing. *Submitted (in second revision).*

[4] S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafna. InsPecT: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77:4626–4639, 2005.

[5] S. Kim, N. Gupta, and P.A. Pevzner. Spectral Probabilities and Generating Functions of Tandem Mass Spectra: a Strike Against Decoy Databases. *J. Proteome Res.*, 7:3354–3363, 2008.

[6] H. Antelmann, H. Tjalsma, B. Voigt, S. Ohlmeier, S. Bron, J.M. van Dijl, and M. Hecker. A Proteomic View on Genome-Based Signal Peptide Predictions. *Genome Res.*, 11:1484–1502, 2001.

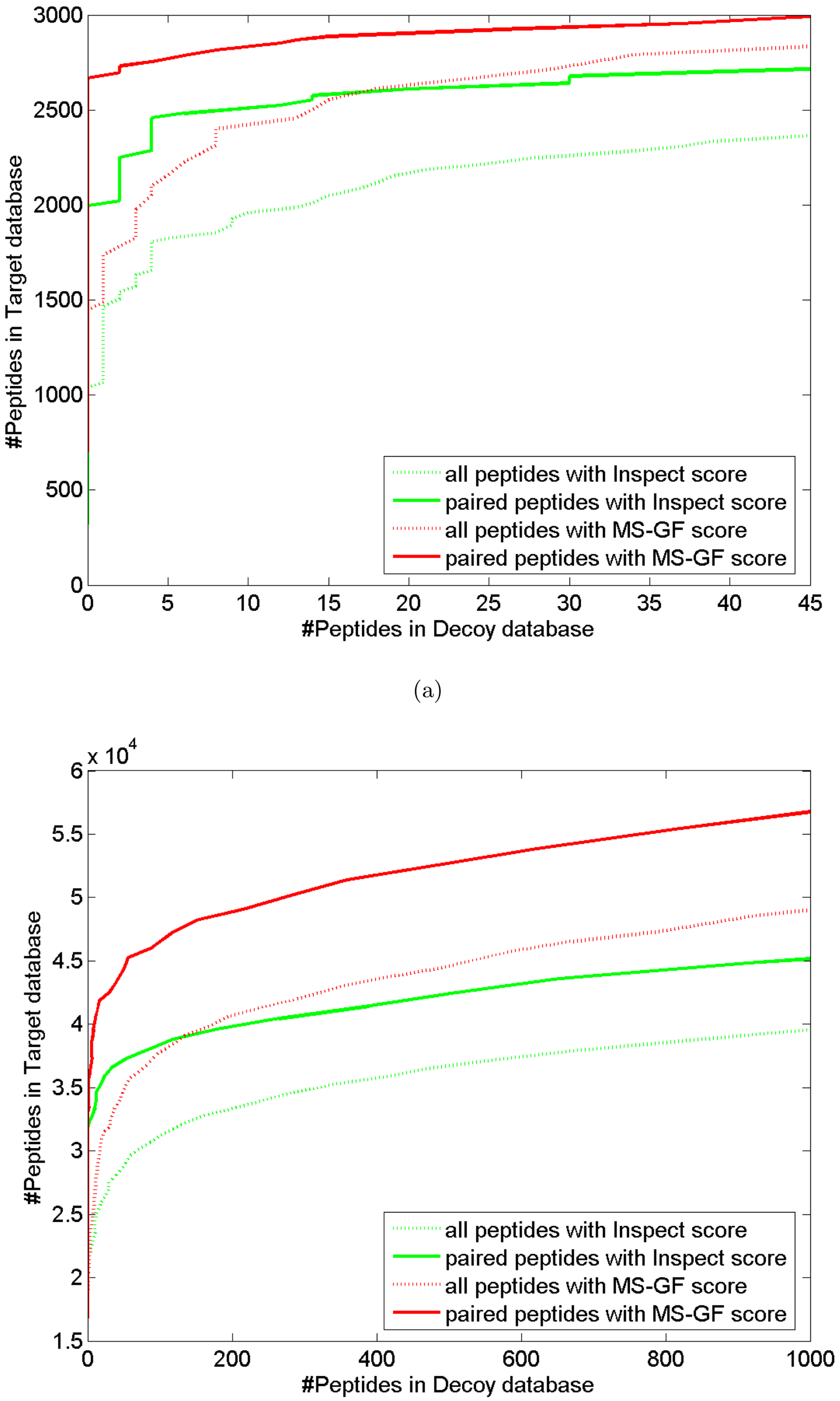
[7] G. Alves and Y.K. Yu. Statistical Characterization of a 1D Random Potential Problem–with applications in score statistics of MS-based peptide sequencing. *Physica A: Statistical Mechanics and its Applications*, 387:6538–6544, 2008.

[8] A. Frank, S.W. Tanner, V. Bafna, and P.A. Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *J. of Proteome Research*, 4(4):1287–1295, 2005.

[9] J.E. Elias and S.P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4:207–214, 2007.

[10] L. Kall, J.D. Canterbury, J. Weston, W.S. Noble, and M.J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–926, 2007.

[11] M. Spivak, J. Weston, L. Bottou, L. Kall, and W.S. Noble. Improvements to the Percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.*, 8:3737–3745, 2009.



(a)

(b)

Figure 1: (a) Identification of peptides in the human dataset using different approaches and scoring functions. For the same number (14) of decoy database hits, InsPecT identifies 2,011 peptides in the target database using traditional approach (T), and 2,577 peptides using the paired-peptide approach (T^*). The number of peptides identified by MS-GF increases from 2,503 to 2,869. (b) Similar plot as in (a) for *Shewanella* dataset.