Group-walk, a rigorous approach to separate FDR analysis by TDC

Jack Freestone[1], Temana Short[1], William Stafford Noble[2], Uri Keich[1]
[1]School of Mathematics and Statistics F07
University of Sydney
[2]Departments of Genome Sciences and of Computer Science and Engineering
University of Washington

January 30, 2022

Abstract

Target-decoy competition (TDC) has been commonly used for a while now in the analysis of tandem mass spectrometry data [9]. Recently, this approach of competition-based false discovery rate (FDR) control has gained significant popularity in other fields after [2] laid its theoretical foundation in a more general setting that included the feature selection problem. In both cases, the competition is based on a head-to-head comparison between an (observed) target score and a corresponding decoy (knockoff) score. The effectiveness of TDC depends on whether the data is homogeneous, which is often not the case: the data might consists of groups with different score profiles or different proportion of true nulls. In such cases, ignoring the group structure TDC might report imbalanced lists of discoveries where some groups might include too many false discoveries. An alternative approach of applying TDC separately to each group does not rigorously control the FDR. We developed *Group-walk*, a procedure that controls the FDR in the target-decoy / knockoff setting while taking into account a given group structure. We show using simulated and real datasets that Group-walk can deliver substantial power gains when the data naturally divides into groups with different characteristics. Group-walk will be made available at publication.

1 Introduction

Tandem mass spectrometry (MS/MS) currently provides the most efficient means of studying proteins in a high-throughput fashion. In a "shotgun proteomics" MS/MS experiment, proteins in a complex biological sample are extracted and digested into peptides, each with an associated charge. These charged peptides, called "precursors," are measured by the mass spectrometer, and a subset of the precursors are then selected for further fragmentation into charged ions, which are detected and recorded by a second round of mass spectrometry [16, 25]. The recorded tandem fragmentation spectra, or spectra for short (A typical 30-minute MS/MS experiment will generate approximately 18,000 such spectra.), are then subjected to computational analysis. Canonically,

each observed spectrum is generated by a single peptide. Thus, the first goals of the downstream analysis are to identify which peptide generated each of the observed spectra (the spectrum-ID problem) and to determine which peptides and which proteins were present in the sample (the peptide/protein detection problems).

In each of those three problems the canonical approach to determine the list of discoveries is by controlling the false discovery rate (FDR) through some form of target-decoy competition (TDC). TDC has been widely practiced by the computational mass spectrometry community since it was first proposed by Elias and Gygi [9, 6, 17, 10, 14, 30].

Consider for example the process of spectrum-ID. It is typically initiated by scanning each input spectrum against a *target* peptide database for its best matching peptide (in practice the spectrum is matched only against a subset of candidate peptides whose mass is within the measurement tolerance of the so-called precursor mass associated with the spectrum). Pioneered by SEQUEST [12], the search engine uses an elaborate score function to quantify the quality of the match between each of the database peptides and the observed spectrum, recording the optimal peptide-spectrum match (PSM) for the given spectrum along with its score Z_i [24]. In practice, many expected fragment ions will fail to be observed for any given spectrum, and the spectrum is also likely to contain a variety of additional, unexplained peaks [25]. Hence, sometimes the reported PSM is correct — the peptide assigned to the spectrum was present in the mass spectrometer when the spectrum was generated — and sometimes the PSM is incorrect. Ideally, we would report only the correct PSMs, but obviously we do not know which PSMs are correct and which are incorrect; all we have is the score of the PSM, indicating its quality. TDC allows us to report a thresholded list of top-scoring PSMs while controlling the list's FDR as explained next.

First, the same search engine is used to assign each input spectrum a *decoy* PSM score, $\tilde{Z}_i$, by searching for the spectrum's best match in a decoy database of peptides obtained from the original database by randomly shuffling or reversing each peptide in the database. Each decoy score $\tilde{Z}_i$ then directly competes with its corresponding target score Z_i to determine the reported list of discoveries, i.e., we only report target PSMs that win their competition: $Z_i > \tilde{Z}_i$. Additionally, the number of decoy wins ($\tilde{Z}_i > Z_i$) in the top k scoring PSMs is used to estimate the number of false discoveries in the target wins among the same top k PSMs. Thus, the ratio between the number of decoy wins and the number of target wins yields an estimate of the FDR among the target wins in the top k PSMs. To control the FDR at level α, the TDC procedure chooses the largest k for which the estimated FDR is still $\leq \alpha$, and it reports all target wins among those top k PSMs. It was recently shown that, assuming that *incorrect PSMs are independently equally likely to come from a target or a decoy match*, and provided we add 1 to the number of decoy wins before dividing by the number of target wins, this procedure rigorously controls the FDR [15].

As first noted by Efron in the context of the canonical p-value-based FDR analysis, the implicit assumption that FDR control relies on is that the data is essentially homogeneous. When, as often happens, the data is made of groups of hypotheses with distinct group-wise characteristics (e.g., different proportion of true nulls), performing a global FDR analysis can be problematic [8].

Efron's observation is even more valid in the context of TDC where the issue of score calibration is known to play a significant role [17, 20]. For example, Baker, Medzihradszky, and Chalkley noted that spectra characteristics vary with the precursor's charge-state [1]. Specifically, they noted that by separately applying TDC to each charge-state they get more discoveries at the 1% FDR threshold then they get when applying TDC to the combined dataset [1, Table V]. Arguably even more disturbing is the fact that because their group-specific 1% FDR thresholds vary

significantly [1, Table IV], inevitably when a single cutoff is used for the entire set, the discoveries from some groups will be overly conservative while for others they will be overly liberal.

Baker et al. tried to address the specific charge-state structure by redesigning their score function so it is better calibrated with respect to the charge state. However, they also noted a similar problem of group-wise imbalance when considering phosphopeptides vs. unmodified ones: ignoring the distinction between the two groups meant that, in their case, the FDR in the phosphopeptides group was significantly underestimated [1, Table VI]. Their suggestion in this case coincided with Efron's solution which is to apply a separate FDR analysis to each group. Clearly, the latter is more attractive as it does not require us to redesign our score function to address any potential group structure. However, as Efron noted that, applying a separate FDR analysis to each group can fail to control the FDR in the final sample setting.

Indeed, we show here that separately applying TDC to each group can significantly underestimate the true FDR. Instead, we offer Group-walk a new procedure that shares the essence of separate TDC analyses but rigorously controls the FDR even for finite samples. Group-walk derives its theoretical guarantees from a recently published paper that developed a general framework for the problem of multiple hypothesis testing with generic side information, i.e., associated with each hypothesis is a p-value, as well as a predictor encoding contextual information about the hypothesis [22]. Of course, a group ID, as in our case, is a special case of such side information but to the best of our knowledge we are the first to adapt this general framework of Lei and Fithian to our context.

We demonstrate the potential power of Group-walk using simulated, as well as real datasets. The simulated data allows to examine how Group-walk takes advantage of the different ways in which groups may vary. We then continue to analyze real spectra sets, where we investigate multiple, naturally-defined, group structures in randomly selected datasets. Thereby demonstrating the significant potential power gains one can have when the data naturally partitions into groups.

2 Background

TDC is a procedure that uses competition to control the FDR which can be phrased in more general terms as follows. Let H_i ($i = 1, \ldots, m$) denote our m null hypotheses, e.g., in the spectrum-ID problem H_i is "the ith PSM is incorrect" and in the peptide detection problem H_i is "the ith peptide is present in the sample." Associated with each H_i are two competing scores: a target score Z_i (the higher the score the less likely H_i is) and a decoy score $\tilde{Z}_i$. For example, in spectrum-ID Z_i ($\tilde{Z}_i$) is the score of the optimal target (decoy) peptide match to the ith spectrum, whereas in peptide detection Z_i corresponds to the maximal score of all PSMs that are optimally matched to the ith target peptide (and $\tilde{Z}_i$ is similarly defined for the ith decoy peptide).

Adopting the notation of [11] we associate with each hypothesis a ("winning") score W_i and a target/decoy-win label L_i. For the problems we consider here $W_i = Z_i \vee \tilde{Z}_i$ (where $x \vee y = \max\{x, y\}$) but other functions can be considered as well [2]. As for L_i, we define it as 1 if H_i corresponds to a target win and as -1 if H_i corresponds to a decoy win. Here we assume that ties are broken randomly but ignoring hypotheses with ties is also valid.

Let $N = \{i : H_i \text{ is a true null}\}$ and note that while typically in the context of hypotheses testing N is a constant, albeit unknown set, here we allow N to be a random set. The fundamental assumption that TDC relies on is:

Assumption 1. *Conditional on all the scores $\{W_i\}_i$ and all the false null labels $\{L_i : i \notin N\}$, the true nulls are independently equally likely to be a target or a decoy win, i.e., the random variables (RVs) $\{L_i : i \in N\}$ are conditionally independent uniform ± 1 RVs.*

Clearly, if the assumption holds then $\{L_i : i \in N\}$ are still independent uniform ± 1 RVs after ordering the hypotheses in increasing order so that $W_1 \leq W_2 \leq \cdots \leq W_m$. Hence, without loss of generality the scores are assumed to be in increasing order.

Let $D(k) = \sum_{i=k}^{m} 1_{\{L_i=-1\}}$ (the number of decoy wins with score $W_i \geq W_k$) and $R(k) = \sum_{i=k}^{m} 1_{\{L_i=1\}}$ (the corresponding number of target wins). TDC defines its discovery list as $\{i \geq K : L_i = 1\}$, i.e., all the target wins whose score is $\geq W_K$ where $K = K(\alpha) = \min\{k : (D(k)+1)/(R(k) \vee 1) \leq \alpha\}$ is TDC's rejection threshold. It was proved by He et al. that TDC controls the FDR in the finite sample setting [15]. That is, with $V(k) := \sum_{j=k}^{m} 1_{\{L_j=1, j \in N\}}$ (the number of true null target wins with score $W_j \geq W_k$), the FDP among the corresponding target wins is $Q_k = V(k)/(R(k) \vee 1)$, and $E(Q_K) \leq \alpha$. The expectation is taken with respect to the labels of the true null hypotheses conditional on the false null labels and on all the given scores.

At essentially the same time Barber and Candés established the same result through a sweeping generalization that deals with sequential hypothesis testing. In that setup the order of the hypotheses is predetermined and a p-value is associated with each hypothesis so that given the order, as well as the p-values of the false nulls, the p-values of the true null hypotheses are iid with a distribution that stochastically dominates the uniform distribution (so they are valid p-values). In our case we order the hypotheses by the scores W_i and the p-values are defined as 1 for a decoy win and 1/2 for a target win. It is easy to see that Assumption 1 guarantees that the conditions of the sequential hypothesis testing are satisfied so it indeed generalizes TDC's setup.

One reason TDC was adopted in the mass spec context, rather than relying on standard methods for control of the FDR such as the procedures by Benjamini and Hochberg [3] or Storey [28], is that the latter require sufficiently informative p-values and, initially, no such p-values were computed in this context (as noted above, using the decoys we can always assign a "1-bit p-value" to the hypotheses but those are not informative enough to obtain effective results using the latter procedures). Moreover, the proteomics dataset typically consist of both "native" spectra (those for which their generating peptide is in the target database) and "foreign" spectra (those for which it is not). These two types of spectra create different types of false positives, implying that we typically cannot apply the standard FDR controlling procedures to the spectrum-ID problem even if we are able to compute p-values [18].

3 Results

3.1 Separate TDC analysis does not generally control the FDR

It is not difficult to find examples showing separate TDC analyses fail to control the FDR in a finite sample setting. For example, consider the following setup for two groups. For $g \in \{1, 2\}$, let m_g be the number of hypotheses in group g and π_g be the fraction of false nulls. We set $m_1 = 100$, $m_2 = 1000$ and $\pi_1 = \pi_2 = 0.1$. With Z_i^g and $\tilde{Z}_i^g$ denoting the target and decoy scores of the ith hypothesis in group g, respectively, we draw all scores independently as follows: $Z_i^g \sim N(0, 1)$ if Z_i^g comes from a null hypothesis, and $Z_i^g \sim N(1.5, 1)$ if Z_i^g comes from a false null hypothesis. The decoy score $\tilde{Z}_i^g$ is distributed $N(0, 1)$. For each hypothesis, we determine the

winning scores and labels as defined in Section 2 and apply the TDC procedure separately, resulting in two discovery lists. Repeating the above simulation for 100,000 runs at an FDR threshold of $\alpha = 0.25$, we found the *combined* empirical FDR was 0.258 with an estimated standard error of 0.0005; 16 standard errors above the threshold.

While the above failure to control the FDR is statistically significant, it is rather negligible in terms of the magnitude of the excess FDR. However, by increasing the number of groups we can readily construct examples where the actual FDR of separate TDC analyses is substantially higher than the nominal threshold.

Specifically, we consider n_g independent and identically distributed groups where each group is made of 22 hypotheses, 6 of which are true null hypotheses and 16 are false nulls. Moreover, we assume the range of possible scores for each hypothesis is such that the order of the hypotheses is deterministic. Specifically, when sorted in increasing order first come 3 of the null hypotheses followed by 8 of the false nulls, then the remaining 3 nulls followed by the remaining 8 false nulls. For each false null hypothesis we assume the target always wins, whereas the labels of all true null hypotheses are independently and uniformly drawn.

We repeatedly and independently drew the sequence of labels for each of the n_g groups, applied TDC separately to each group at level $\alpha = 0.1$ and noted the actual FDP in the combined list of discoveries. We estimated the FDR in the combined list by averaging the actual FDP over 100,000 sets of draws for each value of n_g, as we increased it from 1 to 10. These estimated FDR values given in Figure 1 demonstrate how rapidly the excess FDR grows in this case, indicating that even for a small number of groups, separate analyses can provide a poor method of FDR control. For example, with just 3 groups at 0.143 the actual FDR is already over 40% higher than the nominal $\alpha = 0.1$ and with 8 groups at 0.202 it is more than 100% higher.

Let V_g denote the number of true null target wins discovered by TDC in group g, and let R_g be the corresponding number of all discoveries. It is not hard to show that as $n_g \to \infty$ the FDR of the separate TDC analyses will converge to the mFDR of each group which is defined as $E(V_g)/E(R_g)$. Given the size of the problem, it is straightforward to evaluate the ratio of expectations and find that in this case mFDR $\approx 1.156/5.406 \approx 0.214$; significantly over the nominal 0.1.

3.2 Group-walk

Group-walk is our novel procedure designed to analyze the same target-decoy data that TDC does while taking advantage of the data's group structure. Specifically, in addition to the label L_i and the score W_i, each hypothesis has a group identifier $G_i \in \{1, 2, \ldots, g\}$ where $g \geq 1$ is the number of groups. Accordingly, we revise Assumption 1 to account for the group information:

Assumption 2. *Conditional on all the scores $\{W_i\}_i$, the group IDs $\{G_i\}_i$ and all the false null labels $\{L_i : i \notin N\}$, the true nulls are independently equally likely to be a target or a decoy win.*

To understand how Group-walk works it is instructive to revisit TDC. Assuming the scores W_i are in increasing order, we define the "front" as the index k that we are currently considering as the rejection threshold. Starting with the front set to $k = 1$, TDC iteratively checks if the FDR estimated from all the hypotheses that the front is yet to pass over is below the threshold α (i.e., if $(D(k) + 1)/(R(k) \vee 1) \leq \alpha$). If it is, TDC stops and reports the location of the front (implying

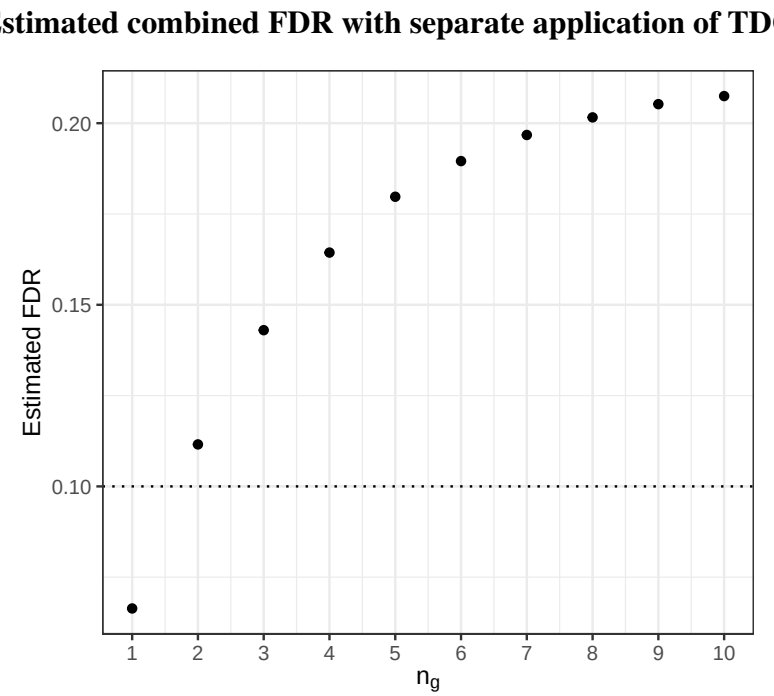

Figure 1: The figure plots the estimated FDR averaged over $100,000$ runs as a function of n_g, the number of groups. The setup for each run are described in the previous paragraph. We computed the estimated standard errors for each estimated FDR, though they were too small to be seen in the figure.

all the target-winning hypotheses that the front is yet to pass over are rejected); otherwise, TDC advances the front by one step $k := k + 1$.

TDC considers a single group, and therefore a 1-dimensional front, whereas Group-walk's front is an n_g-dimensional vector $\mathbf{k} = (k_1, \ldots, k_{n_g})$, where each $k_g \in \{1, 2, \ldots, m_g\}$ and m_g is the size of the gth group. Assuming the scores $\{W_i^g\}_{i=1}^{m_g}$ in each group are sorted in increasing order, Group-walk starts with the trivial front $\mathbf{k} = (1, \ldots, 1)$, i.e., the first index (lowest scoring hypothesis) in each of the groups. Just like TDC, it then iteratively checks if the FDR estimated from all the hypotheses that the front is yet to pass over is below the threshold α. For Group-walk this translates to checking whether $(1 + \sum_{g=1}^{n_g} D_g(k_g)/(1 \vee \sum_{g=1}^{n_g} R_g(k_g)) \leq \alpha$, where $D_g(k) = \sum_{i=k}^{m_g} 1_{\{L_i^g = -1\}}$ (the number of decoy wins in group g with score $\geq W_k^g$) and $R_g(k) = \sum_{i=k}^{m_g} 1_{\{L_i^g = 1\}}$ (the corresponding number of target wins).

If that estimated FDR is $\leq \alpha$ Group-walk stops and returns its current front (implying all the target-winning hypotheses that the front is yet to pass over are rejected), otherwise Group-walk advances the front by one step along a single coordinate. This is where Group-walk fundamentally differs from TDC: its front is n_g-dimensional so it needs to choose which of the g groups should be advanced.

Group-walk employs an initial trivial strategy of iteratively advancing one group at a time until the front reached a pre-determined index K in each group: $\mathbf{k} = (K, K, \ldots, K)$ (here we used $K = 40$). Note that in some rare cases Group-walk might stop before the front reaches this point, and if the front needs to advance beyond the length of a group then that group is tossed away with no discoveries.

When Group-walk reaches the uniform front of $\mathbf{k} = (K, K, \ldots, K)$ it switches its strategy to

try and predict which of the groups is most likely to have a decoy at the current location of the front. That is, considering $L^g_{k_g}$ for $g = 1, \ldots, n_g$, which one is most likely $= -1$? The motivation for this is that any hypothesis that the front passes over is "lost": it can no longer be discovered. Of course, hypotheses for which $L^g_i = -1$ cannot be discoveries as they are decoy wins so we might as well sacrifice those. Another way of looking at it is that where the chances of $L^g_i = -1$ are higher it is less likely be a region of high quality target wins.

Group-walk attempts to solve the question of which $L^g_{k_g}$ is most likely $= -1$ by looking at the last K labels of each group and choosing the one that has the most decoy wins. The full description of Group-walk is provided in Algorithm 1.

Algorithm 1: Group-walk

Input: the FDR threshold $\alpha \in (0, 1)$;
a window size $K \in \mathbb{Z}^+$;
a list of m_g labels $L^g_i = \pm 1$ for each group $g \in \{1, \ldots, n_g\}$ where the indices are sorted so that the winning scores W^g_i in each group are increasing: $W^g_1 \leq W^g_2 \leq \cdots \leq W^g_{m_g}$;
Output: A discovery list $\bigcup_g \{i \geq k_g \mid L^g_i = 1\}$
Initialize the threshold indices $k_g := 1$ for $g \in \{1, \ldots, n_g\}$;
while *there exists* $g \in \{1, \ldots, n_g\}$ *such that* $k_g \leq m_g$ **do**
  $F\hat{D}P := \frac{1+\sum_g D_g(k_g)}{1 \vee \sum_g R_g(k_g)}$;
  if $F\hat{D}P \leq \alpha$ **then**
    break;
  else
    if *there exists* $g \in \{1, \ldots, n_g\}$ *such that* $k_g \leq K$ **then**
      Randomly select l from $\{l \mid k_l = \min_g k_g\}$;
      and set $k_l := k_l + 1$;
    else
      for $g \in \{1, \ldots, n_g\}$ **do**
        $w_g := D_g(k_g) - D_g(k_g - K)$
      end
      Randomly select l from $\{l \mid w_l = \max_g w_g\}$; and set $k_l := k_l + 1$;
    end
  end
end
return $\bigcup_g \{i \geq k_g \mid L^g_i = 1\}$

3.3 Group-walk controls the FDR in the finite-sample setting

We establish Group-walk's FDR control by showing it is a special case of (a slightly modified version of) the Adaptive p-value Thresholding (AdaPT) framework of Lei and Fithian. AdaPT was developed to address the problem where associated with each hypothesis is a p-value p_i, as well as a predictor x_i encoding contextual information about the hypothesis [22]. In our case the p-value is the 1-bit p-value (1/2 for a target win and 1 for a decoy win), and $x_i = (W_i, G_i)$.

AdaPT progresses through iteratively defined rejection thresholds, $s_t(x)$, that are analogous to Group-walk's front. Indeed, s_t describes the front position at step t of the procedure and for each fixed x the rejection thresholds are monotone decreasing in t. To show that Group-walk is an adaptation of AdaPT we introduce t as the same kind of step counter to Group-walk, and we define $s_t(x)$ in that case as 0 for all the hypotheses that Group-walk's front has passed over by step t, and as 1/2 for all the other hypotheses: with $\mathbf{k}_t = (k_1, \ldots, k_{n_g})$ denoting Group-walk's front at step t, $s_t(W_i^g, g) = 0$ for all $1 \leq i < k_g$ and $s_t(W_i^g, g) = 1/2$ for all $k_g \leq i \leq m_g$.

With this definition of s_t, and provided we slightly modify the definition of AdaPT's A_t to use a strict inequality:

$$A_t = |\{i \,:\, p_i > 1 - s_t(x_i)\}|, \tag{1}$$

A_t coincides with Group-walk's $\sum_{g=1}^{n_g} D_g(k_g)$. Similarly AdaPT's $R_t = |\{i \,:\, p_i \leq s_t(x_i)\}|$ coincides with Group-walk's $\sum_{g=1}^{n_g} R_g(k_g)$. It follows that with this modification their estimated FDR, $(1 + A_t)/(1 \vee R_t)$, coincides with Group-walk's and hence AdaPT's rejection criterion coincides with Group-walk's in this case.

AdaPT does not specify how to adjust its front s_t except to note that the update needs to be a function of A_t, R_t, s_t as well as of all the contextual information ((W_i, G_i) in our case), and all its $\tilde{p}_{t,i}$ which are defined in (2) of [22]. Here is where we need to make our second change to AdaPT's original definitions. Specifically we change the second inequality in their definition (2) to a weak one:

$$\tilde{p}_{t,i} = \begin{cases} p_i & s_t(x_i) < p_i \leq 1 - s_t(x_i) \\ \{p_i, 1 - p_i\} & \text{otherwise} \end{cases}. \tag{2}$$

With the latter change it is easy to see $\tilde{p}_{t,i}$ resolves the target/decoy label, L_i, for any hypothesis that the front has passed over by step t and the label remains unresolved otherwise. It follows that Group-walk's update procedure, which moves the front by moving one of the indices forward, $k_g := k_g + 1$, based on the information gathered from the labels of hypotheses the front already passed over, as well as on the scores and the group assignments is consistent with AdaPT's generic update protocol.

Finally, we verified that Theorem 1 of [22], which proves that AdaPT controls the FDR in the finite sample setting, carries through with our two modifications above: (1) and (2). It follows that under Assumption 2 Group-walk also controls the FDR in the finite sample setting.

4 Applications

4.1 Simulations

Roughly, a group structure can be characterized by its position along three axes: the overall distribution of the scores within groups ("group score-effect"), the separation between true/false null scores within groups ("group separation-effect") and the proportion of false nulls within groups ("group π-effect"). We use simulated data to look at how Group-walk exploits the group structure along each of those three axes.

A score is calibrated if all true null scores are generated according to the same distribution. Therefore we could only examine the group score-effect by simulating non-calibrated data but we used both calibrated and non-calibrated simulated data to study the group separation- and π-effects.

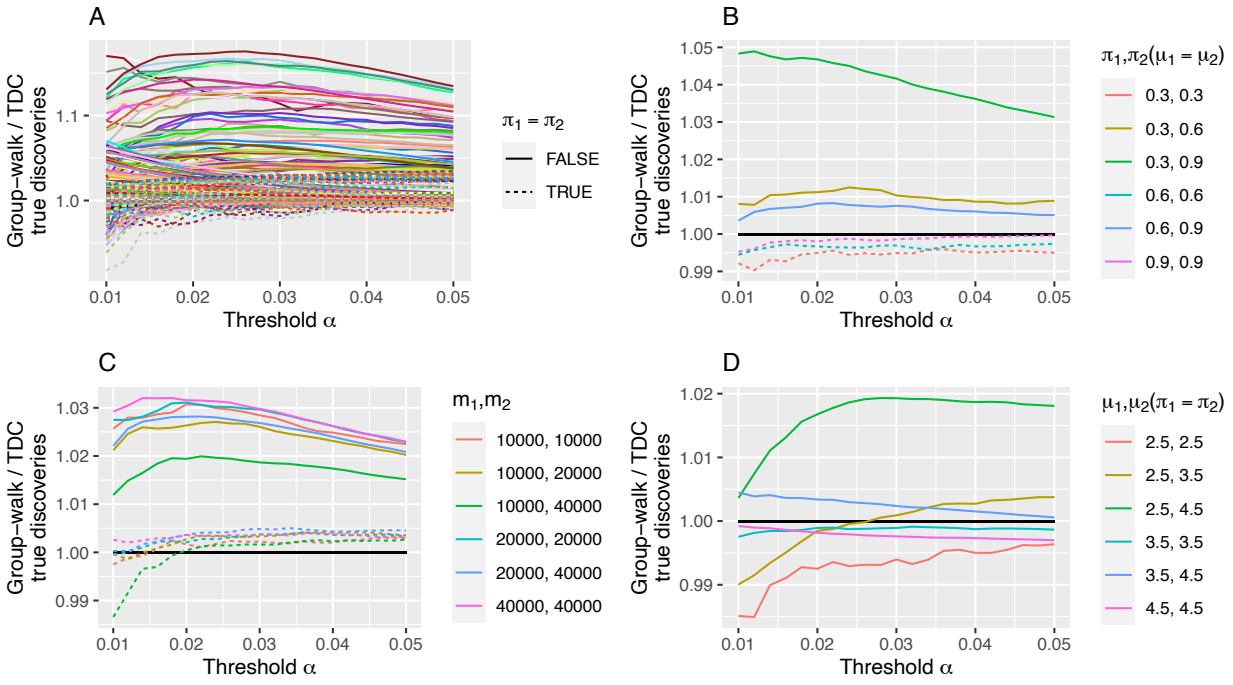

Figure 2: Panel A plots the average ratio between the true discoveries from Group-walk and TDC for all unique configurations of the parameters as outlined in Section 4.1.1. The solid lines indicate when the two groups have a different proportion of false nulls, and the dotted lines when they are the same. Panels B, C and D are the averages of the curves in Panel A, stratified by the parameters (π_1, π_2) when $\mu_1 = \mu_2$, (m_1, m_2) and μ when $\pi_1 = \pi_2$ respectively.

4.1.1 Calibrated scores with group differences in correct PSMs

We simulated generic competition / target-decoy datasets by drawing null scores from the $N(0, 1)$ distribution and drawing false null scores from the alternative, $N(\mu, 1)$, distribution, where $\mu > 0$ is the separation parameter.

Each randomly drawn dataset consists of $n_g = 2$ groups, where for $g \in \{1, 2\}$ each group has m_g null hypotheses ("the observed target score was generated by the true null distribution"). A fraction of π_g of the target scores (Z_i^g) within each group are sampled from the alternative distribution, and the remaining target scores, as well as all the decoy scores ($\tilde{Z}_i^g$), are sampled from the null distribution. All the draws are made independently.

Varying $m_1, m_2 \in \{1e4, 2e4, 4e4\}$, $\pi_1, \pi_2 \in \{0.3, 0.6, 0.9\}$, and $\mu_1, \mu_2 \in \{2.5, 3.5, 4.5\}$ we drew for each parameter combination 40 random datasets ("runs") as described above. Accounting for symmetric combinations the total number of distinct combinations is 379. We then applied Group-walk to each drawn dataset (with $K = 40$), as well as TDC, but the latter is applied to the combined list of target-decoy scores, i.e., ignoring the group structure.

Figure 2A depicts for each parameter combination the average ratio (across the 40 runs) of Group-walk to TDC true discoveries number. Notably, the majority of the curves lie above 1, indicating that Group-walk yields improvement in power under the majority of these parameter settings.

Figure 2B examines the group π-effect by looking for each configuration of π_1 and π_2, at the average of the curves in panel A for which $\mu_1 = \mu_2$ (removing the separation-effect). As expected, the discovery-ratio is highest for groups with the most significant differences between π_1 and π_2.

When $\pi_1 = \pi_2$ (and $\mu_1 = \mu_2$), we see a decrease in the number of discoveries in Group-walk relative to TDC. This outlines the cost of grouping hypotheses when no group structure is to be found. However, that loss in the number of discoveries is marginal when the fraction of alternatives for both groups is high enough, which is typically the case in the analysis of mass spec data.

Figure 2C shows how the size of the groups can affect the relative power gain of Group-walk by displaying for each configuration of m_1 and m_2 the average of the curves in panel A. The solid curves in the figure ($\pi_1 \neq \pi_2$) suggest that, in the presence of group structure, on average Group-walk performs relatively better when $m_1 = m_2$ and its advantage decreases as the difference between the group sizes increases. This makes sense intuitively if you think about a limiting case where, say $m_1 \ll m_2$, in which case there will be little difference between applying TDC to the combined list and to group 2 by itself.

Finally, Figure 2D examines the group separation-effect by showing for each configuration of μ_1 and μ_2 the average of the curves in panel A for which $\pi_1 = \pi_2$ (removing the π-effect). While the separation-effect is overall weaker than the π-effect for this data, the figure suggests that Group-walk can typically take advantage of this group effect: the curves for which $\mu_1 \neq \mu_2$ are mostly above 1.

4.1.2 Using a non-calibrated score function

The "spectrum-ID model", described in [19, 21], provides a natural setup for simulating non-calibrated datasets. Specifically, with each (virtual) "spectrum", σ_i, we associate a score X_i as follows. If σ_i is native (cf. Section 2) then X_i models the match between σ_i and the peptide that generated it, so we draw $X_i \sim 1 - \text{Beta}(a, b)$, where we fix $a = 0.05$ and we consider b as a separation parameter. Otherwise, σ_i is foreign and we model X_i as a random match against a target database candidate peptide by drawing $X_i \sim \text{Unif}(0, 1)$. All the draws described here are done independently.

In addition, assuming that associated with each spectrum are n candidate peptides (those within the measurement tolerance of the precursor mass, here we used $n = 100$), we associate with each spectrum a score Y_i corresponding to the best match to σ_i among the remaining $n - 1$ candidate target peptides. Here we model Y_i as the maximum of $n - 1$ $\text{Unif}(0, 1)$ scores, i.e., we sample $Y_i \sim \text{Beta}(n - 1, 1)$. The target PSM score is then $Z_i = X_i \vee Y_i$.

We also associate with each spectrum a decoy PSM score, $\tilde{Z}_i$, modeling the best match to σ_i among n candidate decoy peptides. Consistent with our $\text{Unif}(0, 1)$ model of a random match we therefore draw $\tilde{Z}_i \sim \text{Beta}(n, 1)$. It follows that Assumption 2 is only approximately valid: for native spectra there is a slightly larger chance a true null will be a decoy win (which creates a slightly conservative — and hence not overly concerning — bias). That said, if σ_i is foreign then $Z_i \sim \text{Beta}(n, 1)$ so for a foreign spectrum target and decoy wins are equally likely.

To make the scores uncalibrated, we define the winning scores as $W_i^g := Q\left(\max\left(Z_i^g, \tilde{Z}_i^g\right), \mu_i^g, \beta_i^g\right)$, where $Q(\cdot, \mu, \beta)$ is the quantile function of the Gumbel distribution with parameters μ, β. The latter location and scale parameters are sampled from a population of such parameters generated by fitting Gumbel distributions to PSM scores generated from real yeast data, with a presumed charge state of 2 or 3. Further details can be found in [21].

As in the calibrated simulations we consider two groups while varying $m_1, m_2 \in \{1\text{e}4, 2\text{e}4, 4\text{e}4\}$, and the separation parameters $b_1, b_2 \in \{5, 50\}$. We also varied $\pi_1, \pi_2 \in \{0.3, 0.6, 0.9\}$ but note

that here they correspond to the proportion of native spectra in each group, which is related to but not identical to the proportion of false nulls (Section 2).

In addition, to study the group score-effect we create two different types of datasets. In the first, "homogeneous" type, the Gumbel parameters that induce the non-calibrated feature of the data, $\{\mu_i^g, \beta_i^g\}_{i,g}$ are all randomly drawn with replacement from the same population of location and scale parameters. Any such homogeneous dataset is uncalibrated but there is no group score-effect. The second, "non-homogeneous" data type is generated where the Gumbel parameters of the first group are drawn from the population of parameters that were originally fitted to PSMs with a presumed charge state of 2, while the Gumbel parameters of the second group are drawn from the set fitted to a presumed charge of 3. As noted in the Introduction, different charge states are known to be associated with a group score-effect.

Again, we generated 40 samples for each combination of parameters and examined the average ratio of Group-walk ($K = 40$) to TDC true discoveries number. Figure 3A shows all these average ratios and, similarly to the calibrated analysis, most of the curves indicate Group-walk is more powerful. Notably this applies to datasets which include the group score-effect (solid lines) as well as to those which do not (dashed lines).

Figure 3B examines the group π-effect (with and without the score-effect) by looking for each configuration of π_1 and π_2, at the average of the curves in panel A for which $b_1 = b_2$ (removing the separation-effect). While on average Group-walk takes advantage of the π-effect in all parameter combinations (all curves with $\pi_1 \neq \pi_2$ are above 1), Group-walk's power advantage is particularly pronounced when the π-effect is combined with the score-effect (solid curves). Notably, Group-walk benefits substantially from the score-effect even in the absence of the π-effect (solid lines corresponding to $\pi_1 = \pi_2$).

Figure 3C similarly shows the average of the curves in panel A for each configuration of m_1 and m_2. As in the calibrated analysis we note that a large difference in m_i reduces Group-walk's advantage (solid green curve). Again, we see the substantial impact of the score-effect on Group-walk's performance (solid vs. dashed curves).

Finally, Figure 3D examines the group separation-effect (with or without the score-effect) by showing for each configuration of b_1 and b_2 the average of the curves in panel A for which $\pi_1 = \pi_2$ (removing the π-effect). As expected, the separation-effect increases Group-walk's power (green vs. red and blue curves), but only when combined with the score-effect is this increase in power non-negligible and the power is overall higher than TDC. Overall, the consistent picture emerging from all the panels is that the group score-effect can significantly increase Group-walk's power.

4.2 Applications to real data

In this section, we demonstrate the effectiveness of Group-walk when analysing real data endowed with a group structure.

The following data sets used were sampled from projects available from the Proteomics Identifications Database (PRIDE) [27]. 14 spectra files were downloaded from 14 randomly selected projects such that 7 were high-resolution spectra files and 7 were low-resolution spectra files. If a project included multiple high (low) resolution spectra files, then one was downloaded at random. Each project was published to the database no earlier than 2018. The 7 low-resolution spectrum files came from PXD012522, PXD013482, PXD014774, PXD015485, PXD017622, PXD022092

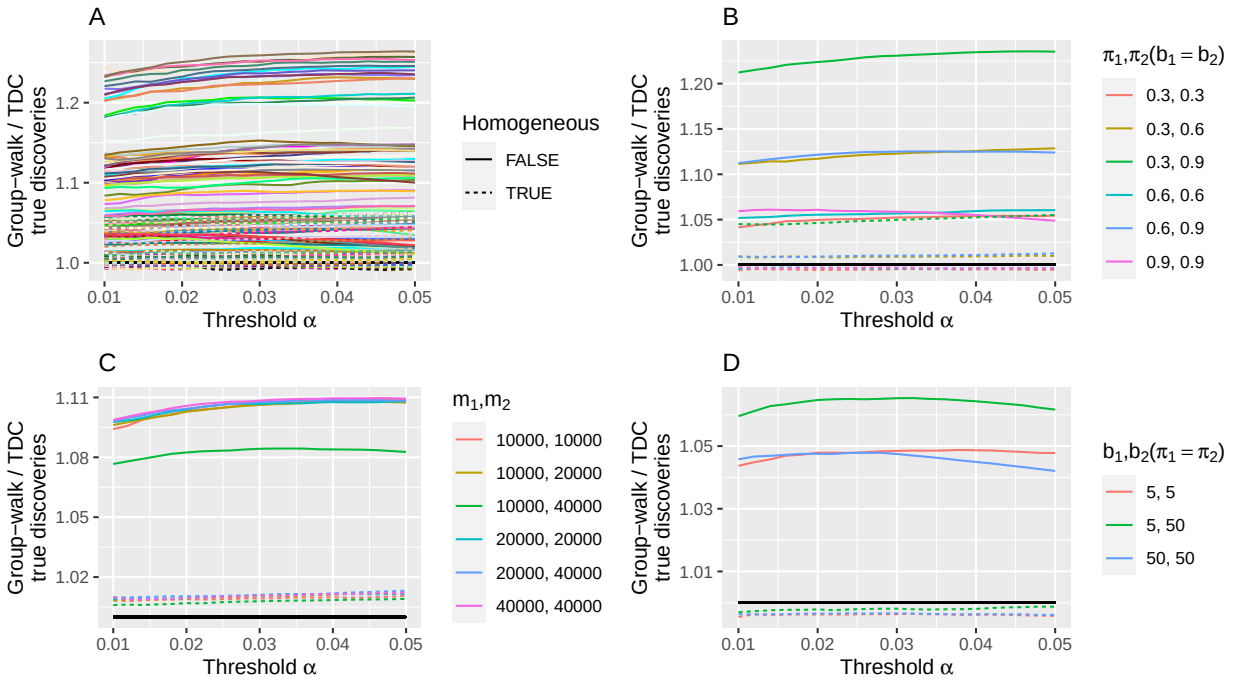

Figure 3: Panel A plots the average ratio between the true discoveries from Group-walk and TDC for all configurations of the parameters as outlined in Section 4.1.2. The solid lines indicate when the two groups include group score-effect, and the dotted lines when they do not. Panels B, C and D are the averages of the curves in Panel A, stratified by the parameters (π_1, π_2) when $b_1 = b_2$, (m_1, m_2) and (b_1, b_2) when $\pi_1 = \pi_2$ respectively.

and PXD027867. The 7 high-resolution spectrum files came from PXD006856, PXD008996, PXD012611, PXD013274, PXD019186, PXD022257 and PXD025130.

We used the Crux toolkit [26] to determine the PSMs and their scores from the spectra files. Tide-index in Crux [7] was used to pre-process the peptide databases used for each spectra file with the option `--auto-modifications-spectra` set to `TRUE`, to account for variable modifications and to optimise parameter settings for the spectrum file [23]. We repeated this process 20 times with respect to different seeds to create 20 random decoy peptide databases which our results were averaged over. Tide-search [7] was used to generate the PSM scores, where the exact options are specified in each of the following subsections.

Spectrum-level analysis of the PSMs can suffer from independence issues that potentially violate FDR control [15], thus, we chose to focus on the peptide-level analysis. We used the following peptide-level analysis of the PSMs: that is, the target score Z_i (or decoy score $\tilde{Z}_i$) is equal to the maximal score of all PSMs matched to the ith target peptide (decoy peptide).

4.2.1 Charge and Peptide length in low-resolution MS/MS Data

It is well understood that charge contributes to the lack of calibration for some score functions, like SEQUEST's Xcorr score. Additionally, we observe a similar behaviour relating to the length of matched peptides (the number of amino acids) in the database. Here we demonstrate how the usual TDC procedure can be improved upon when applying Group-walk to such uncalibrated scores.

We partition the peptides into groups via charge by considering the charge state of the spectrum that was maximally matched with the peptide. Note that spectra with charge states 2 or 3 include the majority of the spectra, while spectra with charge states 1, 4 and 5 are less frequent. For this

reason, Group-walk will have likely exhausted through the small groups before the estimated FDR reaches the threshold α. To get around this, we merged the smaller groups, say if they comprise of less than 10% of the total number of targets, to the previous charge state group (or in the case of the lowest charge state group, we merged with the subsequent group). In this way, we avoid missing out on discovering high scoring peptides which improve our power, even if it is only a small number of discoveries.

Grouping peptides via the length of matched peptides suffers from some ambiguity – it is not clear how to form the groups given the varying lengths of peptides. However given the results in Figure 3C, it seems optimal to partition the peptides into equally sized groups. Also, we expect that peptides with similar lengths are likely to be more calibrated than peptides with arbitrary lengths. Hence a natural partition of the peptides is to form g similarly sized groups such that the peptide lengths fall into g adjacent intervals. For example if, $g = 3$, then we have 3 groups such that the length of the matching peptides in each group fall into one of the intervals, say, $[2, 11]$, $[12, 17]$, and $[18, \infty)$, where the breaks between each interval are chosen such that the size of each group is approximately the same.

To generate the PSMs and their scores, we used the following options in tide-search: `--auto-precursor-window warn`, `--auto-mz-bin -width warn`, `--concat T`. We focused on using the 7 low-resolution ms/ms data, as the differences in calibration were empirically pronounced there and thus served as a viable application of Group-walk. These data sets are still relevant despite the fact that they are low-resolution as the oldest one was published in 2019.

The results are given in Figure 4. Here we can see a positive increase in the number of discoveries for all FDR thresholds in $[0.01, 0.05]$ and for all groupings. A majority of the improvements of Group-walk are concentrated at lower thresholds of α. At the 1% FDR level of the charge-based grouping, Group-walk does 10.5% better than TDC, while for the length-based grouping, Group-walk offer 6.7% for 2 groups and 5.8% for 3 groups over TDC.

4.2.2 Open search

Canonically each sample spectrum is not searched against the complete target-decoy database but instead against a subset of those peptides in the database whose mass approximates the precursor mass, up to some narrow window of error. We can extend this window as large as we like creating what is called an *open search*. This may be beneficial if we suspect that there exist correct peptide-identifications where the peptide-mass in the database differs from the peptide in our sample by more than the narrow tolerance. This is natural to expect as modifications of the peptide mass may occur through several mechanisms such as methylation, phosphorylation etc. [4].

Once the peptides in the database have been identified via open search, one could perform ordinary TDC. Instead, we propose to group the identified peptides in the following manner and apply Group-walk. For each identified peptide in the database, determine the difference in mass between the spectrum's precursor mass, and the identified peptide. Bin the mass differences according to some regular intervals. Then assign the identified peptides whose corresponding mass difference falls into the top b most frequent intervals into one group, and the remaining into a second group. As an example, the bins might be intervals of the form $\ldots [-1, 0), [0, 1), [1, 2) \ldots$ and each identified peptide gets assigned to an interval according to its corresponding mass difference. Then we look at the identified peptides in the top, say $b = 10$, most assigned intervals into group one, and the rest into group two. We expect that the identified peptides in the first group are indica-

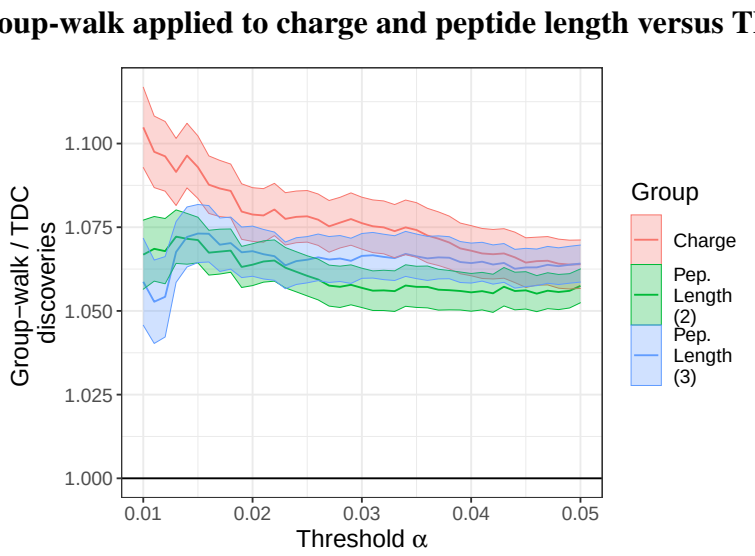

Figure 4: This figure plots the average ratio between the discoveries from Group-walk and TDC. Group-walk was applied to partitions of the peptides via the charge state of the spectra (red) or the length of the matching peptide into either 2 (green) or 3 (blue) groups. The shaded region corresponds to 1SE.

tive of the peptide modifications described in the previous paragraph and signal correct peptide identifications.

It is important to note that this grouping satisfies Assumption 2 required for theoretical FDR control. More specifically, knowing the mass differences between the matched peptides and the sample peptides (along with the winning scores and labels of correctly matched peptides), we still expect the incorrect matches to be independent and equally likely to be a target or a decoy win.

To generate PSMs and their scores in the open search setting, we used the following options in tide-search: `--auto-mz-bin-width warn`, `--precursor-window-type mass`, `--precursor-window 100`, `--use-tailor-calibration T`, `--concat T`. We used both the Xcorr score, and a calibrated version of Xcorr score called *Tailor* score [29] (Note, the option `--use-tailor-calibration T` will return both scores). The data sets used contained the 7 low resolution data sets and 7 high resolution data sets. We used multiples of the canonical value $\lambda = 1.0005079$ Da, to construct breaks in the mass differences at $\{\ldots, -2\lambda, -\lambda, 0, \lambda, 2\lambda, \ldots\}$ (where λ is the difference in mass between the centers of clusters that form naturally from all possible peptides).

Figure 5 demonstrate a substantial increase in the average number of discoveries for all FDR thresholds in the interval $[0.01, 0.05]$ relative to TDC in the open-search setting. As Tailor scores are relatively calibrated, this suggests Group-walk is at least leveraging the fact that different groups exhibit a different proportion of correct peptide identifications as predicted, and not just a difference in calibration between groups. For instance, at the 1% FDR Threshold, Group-walk using Tailor scores returns 18% more discoveries than TDC. Impressively, we see more than a 40% increase in discoveries using Xcorr score at the same threshold.

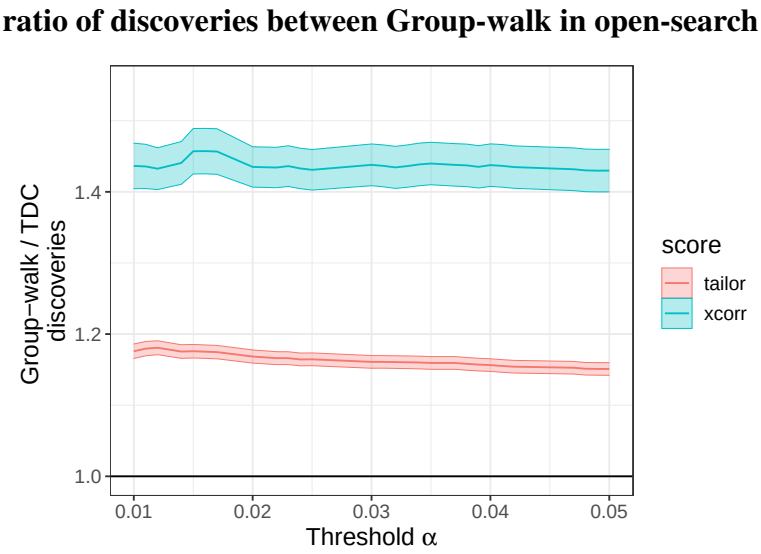

Figure 5: This figure plots the average ratio between the discoveries from Group-walk and TDC in open-search using either Xcorr score (blue) or Tailor score (red). The peptides are partitioned into two groups: the peptides whose mass difference to the spectrum's precursor mass falls into the top $b = 10$ must abundant bins in one group, and the rest in the second group. The shaded region corresponds to 1SE.

5 Discussion

TDC is the common approach to FDR control in the analysis of tandem mass spectrometry data. The implicit assumption that TDC relies on is that the data is essentially homogeneous. However, in practice the data is often made of groups of hypotheses (PSMs, peptides) with distinct group-wise characteristics. Applying TDC while ignoring this structure creates imbalanced list of discoveries when viewed with the correct group structure in mind: the discoveries from some groups will be overly conservative while for others they will be overly liberal.

While applying a separate TDC analysis to each group addresses this problem it fails to control the FDR in the combined list of discoveries and the examples we provide show that this failure can be substantial. Instead, we offer Group-walk, a novel procedure that shares some of the characteristics of separate analyses while still controlling the combined FDR. It is worth noting that, particularly when using the common 1% FDR threshold, Group-walk can possibly deliver even more discoveries than the, non-rigorous, separate analyses. Indeed, the +1 "penalty" in the estimated FDR is applied to each group in the latter case but Group-walk uses a combined FDR estimate so in its case there is a single penalty.

Group-walk can be considered as a specialization of a slightly modified version of AdaPT, which is a recently published general framework for the problem of multiple hypothesis testing with generic side information [22]. In particular, this establishes that Group-walk controls the FDR in the finite sample setting.

To be effective, Group-walk requires typically at least K hypotheses in each group (throughout this work we used $K = 40$). Some other approaches in the spirit of separate TDC analyses have been offered for cases where some groups are small but no finite-sample FDR control was established (e.g. [13]).

We demonstrated using simulated data how Group-walk can benefit from the varying types of group structures: score-effect, separation-effect, and π-effect. It should be noted that both the

separation- and π-effects have general trends that affect how well Group-walk can take advantage of them. For example, when μ is large enough so the alternative distribution becomes sufficiently separated from the null distribution, the problem will become trivial, and TDC will do just as well as Group-walk even if the data exhibits a significant group π-effect.

This raises the important question of how do we know when should we use Group-walk? For now we leave this for future research, noting that our results show that the penalty of using Group-walk in the absence of group structure seems rather marginal, particularly when compared with the significant gains it potentially offers.

A particularly appealing property of Group-walk is that it frees us from worrying about whether our score is sufficiently calibrated with respect to some property of the data, such as the precursor's charge state, put differently, whether our data exhibits a group score-effect. Indeed, as we saw in the impressive discovery gains Group-walk offered when grouping the low-resolution PRIDE datasets by charge, all we have to do is partition the data into groups according to the considered feature and apply Group-walk which will then take advantage of the score-effect.

While we also saw gains when analyzing the PRIDE data sets according to the peptide length and to the mass difference in open search, we also noted that it is not always obvious how to partition the hypotheses into groups. Even more so if the feature we consider is not discrete. This suggests another promising direction for future research, which is to go back to AdaPT's view of having a more general side information rather than just a group ID. Note that it is not obvious what needs to be done as it would still require specializing AdaPT's generic UPDATE function.

While TDC has been in use for some time now in computational mass spectrometry, Barber and Candès recently used the same principle in their knockoff+ procedure to control the FDR in feature selection in a classical linear regression model [2]. Moreover, following their work and the introduction of a more flexible formulation of the variable selection problem in the model-X framework of Candès et al. [5], competition-based FDR control has gained a lot of interest in the statistical and machine learning communities.

While Barber and Candès' knockoff construction is significantly more elaborate than that of the analogous decoys in, say the spectrum-ID problem, knockoffs and decoys serve the same purpose in competition-based FDR control. In particular, Group-walk is applicable to the knockoff context just as well as its is to TDC. That said, in the kind of feature selection problems considered by Barber and Candès the fraction of false nulls (relevant features) is often too low to effectively employ Group-walk.

References

[1] Peter R Baker, Katalin F Medzihradszky, and Robert J Chalkley. "Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state-and sequence-dependent scoring". In: *Molecular & Cellular Proteomics* 9.9 (2010), pp. 1795–1803.

[2] Rina Foygel Barber and Emmanuel J Candès. "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[3] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[4] Boris Bogdanow, Henrik Zauber, and Matthias Selbach. "Systematic errors in peptide and protein identification and quantification by modified peptides". In: *Molecular & Cellular Proteomics* 15.8 (2016), pp. 2791–2801.

[5] Emmanuel Candes et al. "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3 (2018), pp. 551–577.

[6] Fabio R Cerqueira et al. "MUDE: a new approach for optimizing sensitivity in the target-decoy search strategy for large-scale peptide/protein identification". In: *Journal of proteome research* 9.5 (2010), pp. 2265–2277.

[7] Benjamin J Diament and William Stafford Noble. "Faster SEQUEST searching for peptide identification from tandem mass spectra". In: *Journal of proteome research* 10.9 (2011), pp. 3871–3879.

[8] Bradley Efron. "Simultaneous inference: When should hypothesis testing problems be combined?" In: *The annals of applied statistics* 2.1 (2008), pp. 197–223.

[9] Joshua E Elias and Steven P Gygi. "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry". In: *Nature methods* 4.3 (2007), pp. 207–214.

[10] Joshua E Elias and Steven P Gygi. "Target-decoy search strategy for mass spectrometry-based proteomics". In: *Proteome bioinformatics*. Springer, 2010, pp. 55–71.

[11] Kristen Emery et al. "Multiple Competition-Based FDR Control and Its Application to Peptide Detection." In: *RECOMB*. 2020, pp. 54–71.

[12] Jimmy K Eng, Ashley L McCormack, and John R Yates. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database". In: *Journal of the american society for mass spectrometry* 5.11 (1994), pp. 976–989.

[13] Yan Fu and Xiaohong Qian. "Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry". In: *Molecular & Cellular Proteomics* 13.5 (2014), pp. 1359–1368.

[14] Viktor Granholm et al. "Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics". In: *Journal of proteomics* 80 (2013), pp. 123–131.

[15] Kun He et al. "A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics". In: *arXiv preprint arXiv:1501.00537* (2015).

[16] Patricia Hernandez, Markus Müller, and Ron D Appel. "Automated protein identification by tandem mass spectrometry: issues and strategies". In: *Mass spectrometry reviews* 25.2 (2006), pp. 235–254.

[17] Kyowon Jeong, Sangtae Kim, and Nuno Bandeira. "False discovery rates in spectral identification". In: *BMC bioinformatics* 13.16 (2012), pp. 1–15.

[18] U Keich and WS Noble. "Controlling the FDR in imperfect database matches applied to tandem mass spectrum identification". In: *Journal of the American Statistical Association* 10.01621459.2017 (2017), p. 1375931.

[19] Uri Keich, Attila Kertesz-Farkas, and William Stafford Noble. "Improved false discovery rate estimation procedure for shotgun proteomics". In: *Journal of proteome research* 14.8 (2015), pp. 3148–3161.

[20] Uri Keich and William Stafford Noble. "On the importance of well-calibrated scores for identifying shotgun proteomics spectra". In: *Journal of proteome research* 14.2 (2015), pp. 1147–1160.

[21] Uri Keich and William Stafford Noble. "Progressive calibration and averaging for tandem mass spectrometry statistical confidence estimation: Why settle for a single decoy?" In: *International Conference on Research in Computational Molecular Biology*. Springer. 2017, pp. 99–116.

[22] Lihua Lei and William Fithian. "AdaPT: an interactive procedure for multiple testing with side information". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4 (2018), pp. 649–679.

[23] Damon H May, Kaipo Tamura, and William S Noble. "Detecting modifications in proteomics experiments with Param-Medic". In: *Journal of proteome research* 18.4 (2019), pp. 1902–1906.

[24] Alexey I Nesvizhskii. "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics". In: *Journal of proteomics* 73.11 (2010), pp. 2092–2123.

[25] William Stafford Noble and Michael J MacCoss. "Computational and statistical analysis of protein mass spectrometry data". In: *PLoS computational biology* 8.1 (2012), e1002296.

[26] Christopher Y Park et al. "Rapid and accurate peptide identification from tandem mass spectra". In: *Journal of proteome research* 7.7 (2008), pp. 3022–3027.

[27] Yasset Perez-Riverol et al. "The PRIDE database and related tools and resources in 2019: improving support for quantification data". In: *Nucleic acids research* 47.D1 (2019), pp. D442–D450.

[28] John D Storey. "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[29] Pavel Sulimov and Attila Kertész-Farkas. "Tailor: A nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics". In: *Journal of proteome research* 19.4 (2020), pp. 1481–1490.

[30] Matthew The, Ayesha Tasnim, and Lukas Käll. "How to talk about protein-level false discovery rates in shotgun proteomics". In: *Proteomics* 16.18 (2016), pp. 2461–2469.