# Table of Contents

Abstract .................................................................................................................................................. 3  
1) Introduction .................................................................................................................................... 3  
2) Methods .......................................................................................................................................... 5  
   2.1) The reference datasets ............................................................................................................ 5  
   2.2) The Spectral library search ...................................................................................................... 6  
   2.3) Calculating the FDR ................................................................................................................ 6  
   2.4) Our decoy algorithm - DeLiberator .......................................................................................... 7  
   2.5) Benchmarking with different decoy library creation algorithms ........................................... 11  
3) Results and Discussion ................................................................................................................... 12  
   3.1) Characteristics of top-ranked incorrect SSM ............................................................................ 12  
   3.2) Comparing the DeLiberator algorithm to alternative approaches .......................................... 25  
4) Concluding Remarks ..................................................................................................................... 31  
5) Acknowledgment ............................................................................................................................ 32  
6) References ..................................................................................................................................... 32
Abstract

As more proteomic data are gathered, libraries of annotated MS2 spectra become increasingly important, since they provide a fast and accurate way to identify newly measured MS2 spectra. In any type of search that involves multiple testing the control of the number of false positive identifications expected in the final result list by means of the false discovery rate (FDR) is of crucial importance. In classical sequence searches, where the experimental MS2 spectra are compared to theoretical peptide spectra calculated from a sequence database the FDR is estimated by searching randomized or decoy sequence databases. Even though there is some discussion on how exactly the FDR has to be calculated, this method is widely accepted in the proteomic community. A similar approach to control the FDR of spectrum library searches was recently presented [1]. In this paper we provide a detailed analysis of the random similarity of distinct peptide spectra and present our own solution for decoy library creation (DeLiberator), which differs from the previously published results in some important details, mainly in implementing new methods to avoid that decoy spectra are too similar to the original library spectra. Using different proteomic datasets and library creation methods, we evaluate our approach and show that it performs favorably.

Abbreviations: SSM - Spectrum Spectrum Match, PSM - Spectrum Spectrum Match, FDR - False Discovery Rate, CID – Collision Induced Dissociation, ETD - Electron Transfer Dissociation,

1) Introduction

A spectrum library is a collection of high quality tandem mass spectra and their peptide identifications. Each spectrum library entry represents a peptide sequence, the peptide ion charge state and mass over charge ratio (m/z) and includes a fragment peak-list compiled from one or more experimental spectra confidently identified to that peptide. Typically m/z and intensity pairs of the peak-list are labeled with one or more ion-types that match the peak within a given mass tolerance. The analysis results of several large proteomics datasets have been converted into spectrum libraries and made available through public repositories. The National Institute of Standards and Technology, NIST, provides a continuously updated list of libraries for samples from several different organisms analyzed using different fragmentation techniques (http://peptide.nist.gov/). Open source software tools have been developed which allow automatic creation of spectrum libraries [2, 3, 4, 5].

Library spectra are rich in reproducible information describing peptide fragmentation compared to a theoretical spectrum modeled by a sequence search tool such as Sequest, Mascot and Phenyx [6, 7, 8]. Several spectrum library search tools have been developed so far [2, 4, 5]. Screening a dataset of unidentified experimental spectra against a spectrum library, produced from a similar sample to the origin of the query dataset provides rapid and sensitive identification of MS2 spectra, and has proven to be an effective complement to sequence search analysis [3, 9]. Ahrné et al. presented a simple workflow, where MS2 spectra were first identified with the Phenyx sequence search tool. A library of the identified spectra was then compiled and searched using SpectraST. The reduced search space and
the high sensitivity of the library search made it possible to substantially increase the number of identified spectra. It has also been proven meaningful to make use of the information describing the fragmentation pattern of the unmodified peptide species available in a spectrum library, to recover modified versions of the same peptide by performing an open modification spectral library search \[10, 3, 11, 12\].

When matching a large set of unidentified MS2 spectra against a large spectral library, most of the spectrum-spectrum matches (SSM) that pass an initial parent mass and charge filter will produce a score larger than the minimal value. Low scoring SSM's are usually associated with random matches and are discarded. This does not mean that they are necessarily wrong, but that the same match score could be obtained with a high likelihood by searching an equal number of unrelated or random spectra. These low scoring matches mainly consist of low quality spectra, but also of high quality spectra of peptides not found in the search database such as peptides carrying unexpected modifications, unexpected cleavage products or peptides not yet present in the spectrum library.

In sequence searches the target-decoy strategy, where a randomized protein database is compiled from shuffled or reversed protein sequences, is an easily implemented and widely accepted way to estimate the FDR associated with the similarity score of a peptide spectrum match (PSM). This approach has been thoroughly discussed in the literature \[13, 14, 15, 16\] while methods to estimate the FDR in spectral library search results have only very recently gained some attention.

Finding an appropriate decoy counter-part of a spectrum library is less straight-forward than generating a decoy sequence database as a spectrum library is a more complex dataset than a list of protein or peptide sequences. A spectrum library contains the actual intensities of a wide range of fragment ions including precursor ions, neutral loss ions, immonium ions and various uncommon or unknown fragments, here referred to as non-annotated peaks. The correlation between ion-type and intensity is determined by the properties of the fragmentation technique used to dissociate the peptide. Furthermore spectrum peaks are implicitly linked not only by peptide cleavage at consecutive fragmentation sites but also as groups of peaks originating from the same cleavage site.

In this work we present a deliberate analysis of frequently shared spectral features of high scoring incorrect SSM produced in a spectral library search of iontrap-CID, QTOF-CID and iontrap-ETD tandem mass-spectrometry data. These initial studies allow us to derive a generic method to compile appropriate decoy spectral libraries, suitable for estimating the accuracy in the results output returned by a spectrum library search tool. Our methods are similar to the recently published decoy creation feature included in the SpectraST software tool \[1\], but we also point out important differences especially with regards to the similarity between the original spectrum and its decoy version. The resulting decoy spectral library generation algorithm is implemented in the Java based software tool, DeLiberator. Our decoy library creation tool is compared to simple spectrum randomization algorithms as well as the decoy library creation algorithm in SpectraST.
2) Methods

2.1) The reference datasets

Fragmentation of peptides using Collision Induced Dissociation (CID) [17] on iontrap (IT) or Quadrupole Time-of-Flight (QTOF) mass spectrometers are mature proteomics technologies which populate several publicly available libraries for different organisms or sample types. In recent years ETD has grown in popularity and it has been shown to be particularly useful for the identification of certain type peptides, such as those that are large, highly basic and/or contain post-translational modifications [18, 19, 20]. In contrast to CID spectra, where a,b,y type sequence ions dominate, ETD spectra mainly contain c and z type fragment ions. While CID spectra commonly contain fragment ions with neutral losses of immonium ions or water molecules, a list of other neutral loss ions are commonly seen in ETD spectra. Precursor ions that received one or more electrons but did not dissociate often constitute a significant fraction of the total ion current in ETD spectra [21].

Our decoy creation tool is benchmarked on publicly available data sets of the three data types described above; IT-CID, QTOF-CID and IT-ETD. For the analysis of decoy libraries created for each data-type, we used one experimental dataset of non-identified spectra and one spectrum library of the same species origin as well as an additional experimental dataset, of a different species origin. Detailed properties of datasets used in this study are listed in Table 1. Throughout this manuscript we refer to the different dataset by their names listed in the table.

Table 1. Description of spectral test datasets.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Organism</th>
<th>Fragmentation</th>
<th>Analyzer</th>
<th>Enzyme</th>
<th>Distributor</th>
<th>Intensity of NA peaks</th>
<th>No. of spectra</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibY-CID-IT</td>
<td>lib⁹</td>
<td>yeast</td>
<td>CID</td>
<td>IT</td>
<td>Trypsin</td>
<td>NIST</td>
<td>0.17</td>
<td>84659</td>
</tr>
<tr>
<td>ExpY-CID-IT</td>
<td>exp⁹</td>
<td>yeast</td>
<td>CID</td>
<td>IT</td>
<td>Trypsin</td>
<td>PeptideAtlas</td>
<td>n/a</td>
<td>230073</td>
</tr>
<tr>
<td>ExpH-CID-IT</td>
<td>exp⁹</td>
<td>human</td>
<td>CID</td>
<td>IT</td>
<td>Trypsin</td>
<td>PeptideAtlas</td>
<td>n/a</td>
<td>50977</td>
</tr>
<tr>
<td>LibY-CID-QTOF</td>
<td>lib⁹</td>
<td>yeast</td>
<td>CID</td>
<td>QTOF</td>
<td>Trypsin</td>
<td>NIST</td>
<td>0.07</td>
<td>3076</td>
</tr>
<tr>
<td>ExpY-CID-QTOF</td>
<td>exp⁹</td>
<td>yeast</td>
<td>CID</td>
<td>QTOF</td>
<td>Trypsin</td>
<td>Chalkley et al.</td>
<td>n/a</td>
<td>3269</td>
</tr>
<tr>
<td>ExpH-CID-QTOF</td>
<td>exp⁹</td>
<td>human</td>
<td>CID</td>
<td>QTOF</td>
<td>Trypsin</td>
<td>PeptideAtlas</td>
<td>n/a</td>
<td>24966</td>
</tr>
<tr>
<td>LibY-ETD</td>
<td>lib⁹</td>
<td>yeast</td>
<td>ETD</td>
<td>LTQ-Orbitrap</td>
<td>Lys-C</td>
<td>PeptideAtlas</td>
<td>0.25</td>
<td>9207</td>
</tr>
<tr>
<td>ExpY-ETD</td>
<td>exp⁹</td>
<td>yeast</td>
<td>ETD</td>
<td>LTQ</td>
<td>Lys-C</td>
<td>PeptideAtlas</td>
<td>n/a</td>
<td>48189</td>
</tr>
<tr>
<td>ExpH-ETD</td>
<td>exp⁹</td>
<td>human</td>
<td>ETD</td>
<td>LTQ-Orbitrap</td>
<td>Lys-C</td>
<td>Scherl et al.</td>
<td>n/a</td>
<td>9089</td>
</tr>
</tbody>
</table>
a) The intensity fraction of all spectrum library peaks explained by non-annotated peaks.
b) Spectral library
c) Experimental dataset of non-identified spectra.

2.2) **The Spectral library search**

Spectrum filtering is a meaningful data-preprocessing step in order to reduce the complexity and noise in MS2 spectra. A long list of different spectrum filtering methods have been presented in the literature and evaluated for different search tools [22]. For all MS2 data in this study we keep the two most intensive peaks in a spectral window of 10 m/z. The removal of, often high intensity, precursor mass ions in the ETD data has previously been proposed to increase spectrum identification rates [21], in order to give more weight to the matching of sequence ions. For ETD data we remove all library peaks annotated as precursor mass peaks as well as all peaks at the regions which could contain potential neutral losses of precursor ion peaks. Furthermore we filter out all peaks within a spectral window of -60 Da to +5 Da surrounding all feasible precursor ion charge states. We also consider interfering high intensity precursor mass peaks from co-eluting peptides of a neighboring precursor charge states (the spectrum precursor charge state +/- 1) and filter them out applying the precursor ion filtering criteria described above. In this paper we evaluate SSMs using a common scoring scheme which calculates the normalized dot-product of the library spectrum and the query spectrum, where the square root transformation is employed to scale peak intensities [23].

Throughout this study each experimental spectrum was screened against all library spectra sharing the same precursor charge state and within a precursor mass window of +/- 0.6 Da. To calculate the normalized dot-product score of a SSM the spectrum peaks of both spectra were binned in m/z intervals of 0.4 Da. Each spectrum bin was assigned the cumulative intensity of all peaks within its defined m/z interval.

2.3) **Calculating the FDR**

There has been some debate about which decoy search strategy is appropriate for MS2 sequence searches, e.g. whether the decoy database should be concatenated to the target database when screening the query dataset or whether the two databases should be searched separately. [14, 15]. Wang *et al.* [14] did an extensive study of different FDR calculation methods for peptide sequence searches. They found they yield essentially the same results if FDR’s are carefully calculated. The way the decoy entries are created is important [14, 16]. The decoy database should reflect the main properties of the target database such as mass distribution, amino acid frequencies and correlations as well as peptide redundancy. On the other hand it should be sufficiently different from the target database to provide random matches. This is of special importance for the creation of decoy spectral libraries in order to provide realistic estimates of the FDR. For sequence searches the differences between these decoy creation methods are fairly small for large databases and small FDR’s [14]. When comparing different datasets and tools, it is crucial that the same methods are applied allover in order to provide a fair
comparison. In this paper we search concatenated target-decoy libraries and calculate the FDR as suggested in Elias et al. [16] (FDR = 2 * # Decoy SSM / (# Decoy SSM + # Target SSM)). Only the highest scoring SSM (rank 1 SSM) is considered for each query spectrum unless stated differently.

2.4) Our decoy algorithm - DeLiberator

The results of our initial investigations studying some characteristics of a typical score distribution of incorrect SSM, presented in section 3.1, led us to generate decoy libraries where:

1) Each query spectrum has an equal number of candidate spectra in decoy and target libraries.
2) For each target spectrum a decoy spectrum is created with the same number of peaks and intensity sum.
3) Decoy spectrum peaks are positioned on realistic m/z values guided by a protease specific peptide sequence.
4) A realistic intensity distribution of the decoy spectrum is ensured where the correlation between ion-type and intensity is withheld.
5) Each decoy spectrum is ensured to be substantially different from its target version.

The first 4 of the above criteria are ensured by a similar algorithm to the one presented in Lam et. al. [1], where the peptide sequence of each target library spectrum is shuffled and all annotated fragment ions are repositioned according to the new sequence. When shuffling the peptide sequence the C-terminal amino acid is left untouched while the other amino acids are reordered (Figure 1).
The upper spectrum represents the doubly charged peptide DFEYEINGNEGK (precursor ion = 707.8) in the spectrum library LibY-CID-IT. The lower spectrum is a decoy version of this library entry created by the DeLiberator algorithm (decoy peptide EFEDNIGNGYEK). Matched sequence ions are drawn as solid black lines, unmatched sequence ions drawn as solid gray lines, matched non-annotated ions as bold dotted lines and unmatched non-annotated ions as dotted lines.

Figure 1: Target spectrum as template and spectrum created by DeLiberator

The upper spectrum represents the doubly charged peptide DFEYEINGNEGK (precursor ion = 707.8) in the spectrum library LibY-CID-IT. The lower spectrum is a decoy version of this library entry created by the DeLiberator algorithm (decoy peptide EFEDNIGNGYEK). Matched sequence ions are drawn as solid black lines, unmatched sequence ions drawn as solid gray lines, matched non-annotated ions as bold dotted lines and unmatched non-annotated ions as dotted lines.
The 5th criterion marks the main difference between SpectraST and DeLiberator. One clearly wants to avoid creating a decoy database where decoy spectra are highly similar to their target versions, since this may produce high scoring matches to the decoy database and consequently the calculated FDR will be overestimated.

The level of similarity between a decoy and target spectrum can be managed at different steps of the shuffling and repositioning algorithm. When randomizing the peptide sequence to generate a decoy sequence some sequence permutations will produce decoy spectra that are more similar to the target version than others. Furthermore any spectrum library entry is expected to contain non-annotated peaks, which cannot be displaced in accordance with a peptide sequence, and the positioning of these peaks will affect the overall similarity between a target and decoy spectrum.

We take two measures to reduce this similarity. We calculate the normalized dot-product of the annotated peaks of the target spectrum and its decoy spectrum version. This contrasts to the SpectraST approach, where the similarity is evaluated on the peptide sequence level. If this dot-product score, which ranges from 0 - 1 (0 no similarity, 1 identical spectra), is larger than 0.7 the peptide sequence is iteratively re-shuffled until a decoy variant with a spectral similarity to the target library spectrum lower than the set threshold is found. If no reorganization of the peptide sequence produces a dot-product score lower than the threshold, the spectrum version with the lowest similarity score is selected.

Most often spectrum libraries are made up of consensus spectra, meaning that each peptide entry is represented by an averaged spectrum compiled from multiple spectra identified to the same peptide and precursor charge state. A number of different algorithms have been developed to create a peptide representative consensus spectrum from a spectrum cluster [23, 24, 25]. Typically the resulting spectrum generated by any of these algorithms emphasizes peptide specific peaks detected in multiple spectra and discards or penalizes non-reproducible peaks. Consequently a significant portion of non-annotated consensus peaks are likely to be peptide specific ions such as sequence ions of an ion-type not considered by the library creation tool or internal fragments. Ideally when generating a decoy version of a target library spectrum the sequence related portion of the non-annotated peaks should be positioned in accordance with the new decoy peptide sequence, but other peaks (such as immonium ions and precursor ions) should be kept where they are and peaks not related to the sequence should be randomized. Keeping all non-annotated peaks at the same m/z values in both target and decoy spectra, as proposed by the SpectraST decoy creation algorithm, may produce spectral pairs with high similarity in cases where non-annotated peaks make up a large fraction of the spectrum peak intensity. The DeLiberator algorithm keeps all non-annotated peaks below 200 m/z in the decoy version. Non-annotated peaks within a window of 60 Da below all possible precursor ions charge states are shared between a target spectrum and its decoy version (a spectrum of a triply charged precursor includes three such windows below the singly, doubly and triply charged precursor ion). The decoy m/z values of all other non-annotated peaks are sampled from the overall distribution of non-annotated library peaks, grouped by precursor charge state and m/z bins of 100 m/z.

The Pseudo code of DeLiberator algorithm is shown below.
DeLiberator algorithm - Pseudo code

FOR i = 1 to number of spectra in TargetLibrary

    TargetSpectrum = TargetLibrary[i]
    TargetSequence = getPeptide(TargetSpectrum)

    Charge = getCharge(TargetSpectrum)
    ParentMass = getParentMass(TargetSpectrum)
    MassWeight = ParentMass*Charge-MassProton*Charge
    WindowSize = 60 // 60Da for precursor filtering

    PrecursorWindow // set window for the precursor peaks

    // precursor window
    FOR z = 1 to value of Charge
        WindowStart = (MassWeight+z*MassProton-WindowSize)/z
        WindowEnd = (MassWeight+z*MassProton)/z
        PrecursorWindow.setWindow(WindowStart,WindowEnd) // specify the window from start to end
    END FOR

    IF fragmentType of TargetSpectrum is ETD

        CoelutionWindow // set window for the precursor peaks of potential coeluted peptide

        // consider potential charge state of coeluted peptide
        FOR ChargeCoelu = Charge-1 to Charge+1
            FOR z = 1 to value of ChargeCoelu
                WindowStart = (MassWeight+z*MassProton-WindowSize)/z
                WindowEnd = (MassWeight+z*MassProton)/z
                CoelutionWindow.setWindow(WindowStart,WindowEnd) // specify the window from start to end
            END FOR
        END FOR
        CoelutionWindow.filterPeaks(TargetSpectrum) // recreate new TargetSpectrum
    END IF

    DotProductScore = 1
    Count = 0
    DecoySequence

    WHILE DotProductScore > ScoreThreshold or Count < MaximumShuffling

        DecoyCandidateSpectrum
        DecoySequence = ShuffleSequence(TargetSequence)

        FOR j = 1 to number of peaks in TargetSpectrum

            TargetPeak = TargetSpectrum[j]
            TargetMz = getMz(TargetPeak)
            TargetAnnotation = getAnnotation(TargetPeak)

            // decoy and target have common annotations
            DecoyAnnotation = TargetAnnotation
            IF TargetAnnotation is fragment ion

            END IF

            END FOR

        END WHILE

    END
// calculate the mz of decoy using the annotation
// ie) if the annotation is "y4", calculate new m/z from decoy sequence
DecoyMz = RepositionPeak(DecoyAnnotation,DecoySequence)
ELSE IF TargetAnnotation is precursor or immonium ion
    DecoyMz = TargetMz
ELSE IF TargetAnnotation is non-annotated
    IF TargetMz < 200 or PrecursorWindow.isIncluded(TargetMz)
        DecoyMz = TargetMz
    ELSE
        // ie) if TargetMz is 430, one value is randomly sampled
        // from the m/z distribution between 400 and 500 m/z
        DecoyMz = SamplePeak(TargetMz)
    END IF
END IF
END IF
DecoyPeak = createPeak(DecoyMz,DecoyAnnotation)
ADD DecoyPeak into DecoyCandidateSpectrum
END FOR

DotProductScore = DotProduct(TargetSpectrum,DecoyCandidateSpectrum)

IF DotProductScore < DotProductScore of DecoySpectrum
    DecoySpectrum = DecoyCandidateSpectrum
END IF

INCREMENT Count
END WHILE
ADD DecoySpectrum into DecoyLibrary
END FOR

2.5) Benchmarking with different decoy library creation algorithms

Screening a dataset of experimental spectra against a spectral library of peptide spectra, produced from a sample of a species sufficiently distant from the species of the query dataset, is expected to return a minimum number of correct SSM. Although some sequence homology is present this search serves as an appropriate way of generating a realistic distribution of mainly incorrect SSM scores, if the datasets were produced from samples digested with the same protease and analyzed using the same fragmentation technique. A library matching these three criteria is from here on referred to as a different species library (DSL). DSLs are useful to investigate the properties of randomly matching spectra. However, they are not necessarily suitable for estimating the FDR in the score distribution of SSM to the target library. Target and decoy libraries should not only contain the same number of spectra but ideally every query spectrum should have an equal number of candidate peptide spectra in the target and decoy library. Finding such a DSL for any queried target library would be difficult. It would be more convenient and robust to have a method that creates decoys from existing target spectra.
Furthermore, in order to ensure an accurate modeling of the score distribution of incorrect target library matches, target and decoy libraries need to share similar spectral structures, as will be detailed in this study.

To evaluate the shuffling and repositioning algorithm we implemented two additional simple decoy creation algorithms: 1) A sampling algorithm where peaks are randomly sampled from the target spectrum library. In order to preserve a realistic spectrum intensity distribution the overall distribution of peaks are grouped by precursor charge state and sampling intervals of 100 m/z. For every peptide in the target spectrum library a sampled version is generated containing an equal number of peaks and the same peak annotations. Precursor mass peaks are kept at their original m/z values. 2) The second decoy creation algorithm simply shifts all peaks in the target spectra by 3 m/z (precursor ion peaks are not shifted).

We benchmark the DeLiberator algorithm against the two simple decoy algorithms described above as well as the SpectraST (version 4.0) decoy creation algorithm.

3) Results and Discussion

3.1) Characteristics of top-ranked incorrect SSM.

To define generic criteria stating the properties of appropriate decoy spectra we investigated the characteristics of commonly shared spectral features of distinct peptides for the different datasets. We started out by trying to understand the defaults of simple peak sampling and shifting decoy creation algorithms.

Figure 2 shows that searching an experimental dataset against a DSL consistently produces higher scoring SSMs than when screening the same dataset against simple sampled and shifted decoy spectral libraries (created as detailed in section 2.5). We can thereby conclude that spectra created by the simple decoy creation algorithms are not sufficiently realistic. Searching these libraries is expected to lead to an underestimation of the FDR of SSMs to the target library. Figure 2 also shows that the difference in dot-product score between DSL and decoy SSMs is substantially higher than the difference in number of matched peaks, indicating that real spectra of distinct peptides share some structural similarity that leads to the matching of higher intensity spectrum features, compared to the peaks matched between experimental and sampled or shifted spectra. In the following figures of this section we try to illuminate this spectral structure and show how it differs between different data-types.
We calculated the total number of matched peaks (cumulative Shared Peak Count, SPC) and the cumulative dot-product score of top-ranking SSMs when searching an experimental spectrum against a DSL. These two properties were also calculated when screening the same experimental dataset against sampled and shifted decoy libraries (created as detailed in section 2.5). The box-plot shows the ratio of decoy to DSL values of these two properties (Decoy cumul. spc / DSL cumul. Spc, Decoy cumul. dp / DSL cumul. dp). A ratio of approximately 1, for both properties, would indicate that the simple decoy libraries contain sufficiently realistic spectra. The datasets used for this analysis were; (IT) ExpH-CID-IT searched against LibY-CID-IT, (QTOF) ExpH-CID-QTOF searched against LibY-CID-QTOF and (ETD) ExpH-ETD searched against LibY-ETD.

Figure 3 (MS2-MS1 plot) gives a global overview of all spectra in the LibY-ETD spectrum library revealing the rich structure of ETD tandem mass spectra. Vertical and diagonal bands of spectral features are clearly visible, which correspond to ion types shared between spectra of distinct peptides. Figure 3A highlights both parent (blue) and non-annotated (red) ions for triply charged ETD spectra, and indicates that the parent ions can be detected with high intensity and in different charge states (the parent mass charge states 3, 2, 1 correspond to the 3 diagonal blue bands with decreasing slope). These precursor ions received 0, 1, or 2 electrons but did not dissociate [26]. Some non-annotated ions are
found on bands parallel to parents ion bands and probably correspond to non-annotated neutral losses or isotopes. However, others lie on diagonal bands with different slopes than the parent ions and most likely correspond to parent ions of co-eluting peptides carrying 2 or 4 charges. Vertical lines are explained mainly by low mass c or z sequence ions, whereas large c and z ions form bands parallel to precursor ions of charge 2 and 1 (Figure 3B). Supplementary Figure 4 shows all the c (z) ions, which have intensities in the highest 20% of all intensities of their spectrum and a complementary high intensity z (c) ion from the same fragmentation site. These ions seem to occur quite frequently over the entire mass range (78% of all high intensity c and z ions have a complement, 59% have a high intensity complement) indicating their importance when modeling real or decoy spectra.

a)

![Image](image1)

b)

![Image](image2)
Gray-color coding indicates the quintile of the ion intensity in the spectrum (0-20%, 20-40%, 40-60%, 60-80%, 80-100%). Light gray correspond to low intensity and dark gray to high intensity. Only annotations with a mass difference less than 0.2 Da were considered. A) Scatterplot of the entire mass range. Blue dots identify high intensity ions (80-100%) annotated as parent mass peaks and red dots are high intensity non-annotated peaks. B) Zoom into lower m/z region. Blue dots indicate c-ion and red dots z-ions (all intensities). C) Zoom into intermediate m/z range. Blue dots indicate c-ion and red dots z-ions (all intensities). D) Red dots represent all high intensity (80-100%) c (z) ions that have a complementary high intensity z (c) ion.

The MS2-MS1 plots for the triply charged ion trap spectra (LibY-CID-IT) (Figure 4) disclose less structure. Doubly and singly charged parent ions are largely absent and the presence of parent ions of
co-eluting peptides is not obvious. Vertical bands are formed mainly by low mass a, b and y ions, but diagonal bands of higher order ions are not visible. The bulk of high intensity fragment ions lies within a 500 Da interval centered at the parent m/z value and complementary b and y ions occur with a similar frequency as complementary c- and z ions in ETD. Doubly charged spectra from the QTOF library (LibY-CID-QTOF) show an intermediate pattern (since the QTOF library was fairly small we took doubly charged ions, which are the most frequent ones). Intense low order a, b and y ions form vertical bands in the lower mass range, whereas bands of large, singly charged a, b and y ions parallel to the singly charged parent mass are clearly visible. Complementary high intensity a and b ions are formed predominantly in the low m/z range of the spectra, whereas complementary b and y ions can be found over the entire mass range (Figure 5).

a)

![Graph a)](image)

b)

![Graph b)](image)
Gray-color coding indicates the quintile of the ion intensity in the spectrum (0-20%, 20-40%, 40-60%, 60-80%, 80-100%). Light gray correspond to low intensity and dark gray to high intensity. Only annotations with a mass difference less than 0.2 Da were considered. A) Scatterplot of the entire mass range. Blue dots identify high intensity ions (80-100%) annotated as parent mass peaks and red dots are high intensity non-annotation peaks. B) Zoom into lower m/z region. Blue dots indicate y ions and red dots a- or b ions (all intensities). C) Red dots represent all high intensity (80-100%) b (y) ions that have a complementary high intensity y (b) ion, whereas blue dots show all high intensity (80-100%) b (a) ions that have a complementary high intensity a (b) ion.
Gray-color coding indicates the quintile of the ion intensity in the spectrum (0-20%, 20-40%, 40-60%, 60-80%, 80-100%). Light gray correspond to low intensity and dark gray to high intensity. For the annotations, only annotations with a mass difference less than 0.2 Da were considered. A) Scatterplot of the entire mass range. Blue dots identify high intensity ions (80-100%) annotated as parent mass peaks and red dots are high intensity non-annotation peaks. B) Zoom into lower m/z region. Blue dots indicate y ions and red dots a- or b ions (all intensities). C) Zoom into intermediate m/z range. Blue dots indicate y ions and red dots a- or b ions (all intensities). D) Red dots represent all high intensity (80-100%) b (y) ions that have a complementary high intensity y (b) ion, whereas blue dots show all high intensity (80-100%) b (a) ions that have a complementary high intensity a (b) ion.

In order to obtain more quantitative information about shared spectral features of distinct peptides, we looked at the most commonly matched ion-types of top-ranked incorrect SSM. When screening dataset ExpH-ETD against the DSL LibY-ETD, precursor mass ions clearly dominate (Figure 6a). Commercial and open source sequence search based identification tools are adapting their identification algorithms to better accommodate analysis of ETD datasets and it has been shown that removing precursor mass ions, which typically have high intensities but contain little peptide sequence information, should be a mandatory filtering step of ETD data [26]. Spectral library search tools using correlation scoring give important weight to intense spectral features thus the removal of high intensity precursor mass features improve the separation of scores attributed to correct and incorrect SSMs.

Figure 6b describes the ion-type annotations of matched peaks in top-ranking SSM when screening dataset ExpH-ETD against the DSL LibY-ETD where precursor mass ions have been discarded from the spectral data. The results show a very clear bias towards matching high-order ions, in particular $c_{n-1}$ and $c_{n-2}$ ions.
Figure 6: Frequency of matched peaks in the spectra including sequence ion peaks (ExpH-ETD vs library LibY-ETD)
The bar-plot describes the ion-type annotations (y, z and c-ions) of matched peaks in top-ranking SSM when screening dataset ExpH-ETD against library LibY-ETD where precursor mass ions have been kept from the spectral data (a) and precursor mass ions have been discarded from the spectral data (b). The bars indicate the frequency of the different ion-types in the library spectra representing peptides of length 14 (precursor charge state 3+), and the overlaid points indicate the frequency at which these ion-types were matched in top-ranked SSM. Stars indicate ion-types that are significantly “over-matched” (binomial pval < 0.05).

A similar pattern is observed when analyzing the QTOF reference data (searching the experimental dataset ExpH-CID-QTOF against the DSL LibY-CID-QTOF) showing a strong bias towards matching light C- and N-terminal ions, e.g. $y_1$, $y_2$, $a_2$, $b_2$, fragment ions (Figure 7A). Creating the same bar plot representing the results of the ion trap data (searching the experimental dataset ExpH-CID-IT against the DSL LibY-CID-IT) shows that matches to ions in the densely populated m/z region around the parent mass are preferred (Figure 7B).
Figure 7: Frequency of matched peaks in the spectra including sequence ion peaks (QTOF and IT)

The bar-plot a) describes the ion-type annotations (a, b, y-ions and precursor mass ions) of matched peaks in top-ranking SSM when screening dataset ExpH-CID-QTOF against library LibY-CID-QTOF. The bars indicate the frequency of the different ion-types in the library spectra representing peptides of length 9 (precursor charge state 2+), and the overlaid points indicate the frequency at which these ion-types were matched in top-ranked SSM. The bar-plot b) describes the ion-type annotations of matched peaks in top-ranking SSM when screening dataset ExpH-CID-IT against library LibY-CID-IT. The bars indicate the frequency of the different ion-types in the library spectra representing peptides of length 11 (precursor charge state 2+), and the overlaid points indicate the frequency at which these ion-types were matched in top-ranked SSM. Stars indicate ion-types that are significantly “over-matched” (binomial pval < 0.05).

Figure 8a shows the frequencies of the simultaneous matching of implicitly linked peak pairs when separately screening dataset ExpH-ETD against the DSL LibY-ETD and a sampled version of LibY-ETD created as detailed in section 2.5. The two libraries globally share similar m/z distribution and correlation between m/z values and their intensity, but the sampled spectra do not capture the correlative occurrence of spectrum peaks as part of an ion-series or as part of a group of peaks originating from the same fragmentation site. We observe that correlated ions are matched at significantly higher frequencies in the DSL than in the sampled version.
We compared the frequencies of the simultaneous matching of implicitly linked peak pairs when separately screening the experimental dataset ExpH-ETD against the DSL LibY-ETD (Ref. library) and a sampled version of this library (Sampled Library) (a). All spectra are precursor mass filtered. The bar-plot display the matching frequencies of consecutive ions pairs (e.g. $c_n$ and $c_{n+1}$, CC,ZZ,BB,YY ), charge state pairs (e.g. $c_{n+}$ and $c_{n++}$, C.S.), same sequence fragment pairs (e.g. $y_n$ and $z_n$, YZ,BC) and complimentary sequence pairs (e.g. $c_1$ and $z_{n-1}$, CZ), in top ranked SSM for all triply charged precursor query spectra. b) shows the analysis results when searching the experimental dataset ExpH-CID-IT against the DLS LibY-CID-IT (Ref. library) and a sampled version of this library (Sampled Library). c) shows the analysis results when searching the experimental dataset ExpH-CID-QTOF against the DLS LibY-CID-QTOF (Ref. library) and a sampled version of this library (Sampled Library). In b) and c), the bar-plot display the matching frequencies of consecutive ions pairs (e.g. $y_n$ and $y_{n+1}$, BB,YY ), charge state pairs (e.g. $b_{n+}$ and $b_{n++}$, C.S.), same sequence fragment pairs (e.g. $a_n$ and $b_n$, AB), neutral loss pairs (e.g. $b_n$ and $b_n$-H$_2$O, N.L. ) and complimentary sequence pairs (e.g $b_1$ and $y_{n-1}$, BY), in top ranked SSM for all doubly charged precursor query spectra.

The plots presented in this section are a useful means to investigate the global structure of spectrum libraries and also to check their annotations. However these structures are not generally valid, but depend on mass spectrometer settings and sample preparation. A different enzyme to cut the proteins for example will produce different peptide termini and change the mass of typical short and long
sequence fragments. The main point here is that there is structure in the data, which has to be considered when creating decoy libraries. Since specific low mass m/z values and precursor mass correlated m/z values frequently occur in different spectra with high intensity, top-ranking incorrect SSM of distinct peptides are often explained by matching a limited list of these intense peaks, which are specific to a fragmentation type and protease. Therefore, the decoy library has to consider these high intensity peaks and their correlations in order to model the score distribution of incorrect matches correctly.

3.2) Comparing the DeLiberator algorithm to alternative approaches

When searching an experimental dataset against a DSL concatenated to its decoy version, top-ranked SSMs should originate from the two libraries at equal rates. This simple test indicates whether a decoy library contain sufficiently realistic spectra, and was used to evaluate the accuracy of various decoy generation strategies for both spectral library and sequence search tools [16, 1, 27, 28]. Figure 9 shows that that decoy libraries generated with SpectraST and the DeLiberator produce ratios that are slightly lower than one, but comparable to the results typically obtained when performing a similar evaluation of decoy databases for sequence searching [1]. The target-decoy similarity evaluation also presented in Figure 9 consistently shows that SpectraST decoy libraries are significantly more similar to their target version than DeLiberator spectral libraries.
When searching an experimental data against a DSL concatenated to its decoy version, top-ranked library candidates of the query spectra should originate from the two libraries at equal rates. This simple test indicates whether the decoy library contain sufficiently realistic spectra. Four decoy
creation algorithms were evaluated: The Deliberator algorithm, the decoy creation algorithm of SpectraST and the sampling and shifting algorithms, described in section 2.5. The ratio-axis give the ratio of target SSM to decoy SSM returned in the results output a) when dataset ExpH-ETD was searched against the spectrum library LibY-ETD concatenated to its decoy version, b) when dataset ExpH-CID-IT was searched against the spectrum library LibY-CID-IT concatenated to its decoy version and c) when dataset ExpH-CID-QTOF was searched against the spectrum library LibY-CID-QTOF concatenated to its target version. The similarity-axis indicated the mean dot-product score of top-ranked SSM when the spectral libraries (LibY-ETD, LibY-CID-IT, LibY-CID-QTOF) were searched against their decoy versions.

3.2.1) Controlling the similarity of target and decoy spectra

To provide some additional meaning to the calculated similarity values for target and decoy spectrum libraries we studied the score distributions of top and second rank SSMs. In a search where all SSMs are expected to be incorrect the rank 1 and rank 2 score distributions should be similar, where the latter will be slightly shifted to lower score values (see Figure 10A). This behavior is observed for rank 1 and rank 2 score distributions produced when screening the LibY-ETD against its decoy version compiled by DeLiberator (see Figure 10B). However, the score distribution of the two highest-ranking SSM produced when screening the LibY-ETD against its decoy version compiled by SpectraST is clearly bimodal, indicating that top-ranked SSM are assigned artificially high scores (see Figure 10C). Note that the search results presented in Figures 10B) and C) represent an untypical scenario where the decoy library contains a randomized version of the correct peptide identification of all query spectra. This is rarely the case as MS2 datasets typically contain a large number of poor quality spectra or spectra of peptides not present in the target library. Nevertheless these results explain why FDR estimates obtained from a SpectraST decoy search are expected to be conservative.

a)
Figure 10: The score distributions of top and second rank SSMs
The first and second rank dot-product score distributions of incorrect SSM are expected to have a similar shape, where the second rank distribution has a slightly lower mean score than the first rank score distribution as shown in a), where the experimental dataset ExpH-ETD was searched against the DSL LibY-ETD. To demonstrate the purpose of making efforts to reduce the similarity of target spectra and their decoy versions we studied the first and second rank score distributions when searching a spectrum library (LibY-ETD) against its decoy version created by DeLiberator (b) and SpectraST (c) respectively, and compared the results to a). The histogram shows the combined score distribution of rank 1 and rank 2 SSM dot-product scores and the light gray and dark gray lines show the separate score distributions of rank 1 and rank 2 SSMs.

Figure 11 compares the FDR estimates derived from concatenated target-decoy searches, when decoy libraries were created with SpectraST and DeLiberator, respectively. The FDR comparison show that decoy libraries created by SpectraST consistently provide more conservative FDR estimates than decoy libraries created using the DeLiberator, i.e. fewer valid SSM at typical FDR acceptance levels.

We compare the correlation of FDR estimates calculated from the results output of concatenated target-decoy searches, where decoy libraries were created using SpectraST and DeLiberator respectively. The plot display the FDR estimates derived from the two decoy libraries when screening dataset ExpY-ETD against LibY-ETD target-decoy libraries (dashed-dotted line), dataset ExpY-CID-IT against LibY-
CID-IT target decoy libraries (solid line) and dataset ExpY-CID-QTOF against LibY-CID-QTOF target-decoy libraries (dashed line).

Based on the results presented in this section we argue that a strict control of spectral similarity of target and decoy spectra leads to less conservative and more accurate FDR estimates. We control spectral similarity between target and decoy spectra in two steps of the shuffling and repositioning algorithm: when shuffling the peptide sequence and when positioning non-annotated peaks. In any decoy generation approach which relies on a randomization (sampling/shuffling etc) of peptide sequences, peptide sequences identical to those in the target database need to be avoided. Lam et al. [1] propose an iterative re-shuffling approach to ensure a maximum peptide sequence similarity of 80% between a target peptide and its decoy version. However as pointed out in Sherman et al. peptide sequence homology algorithms do not fully capture the expected spectral correlation of two distinct peptides [29]. What needs to be avoided is similarity in the m/z domain and not based on sequence. Consequently enforcing a threshold on the spectral level similarity between target and decoy spectral library entries is more effective than a sequence homology threshold.

17% 7% and 25% of the intensity sum of all library peaks is attributed to non-annotated peaks, in the LibY-CID-IT, LibY-CID-QTOF and LibY-ETD libraries respectively. Transferring all non-annotated peaks to the decoy spectrum with the same m/z values as in the target spectrum, as proposed by the SpectraST algorithm, may also produce decoy peptide spectra that are highly similar to the target version in cases where non-annotated peaks make up a large part the overall spectral intensity. Figure 12 displays a library spectrum (LibY-CID-IT) and its decoy version created by SpectraST. Even though the target and decoy peptides have a sequence similarity lower than 0.8, the target and decoy library entries have a very high spectral similarity (dot-product score 0.94), due to shared intense sequence ions and non-annotated peaks not repositioned by the decoy algorithm.
The upper spectrum represents the doubly charged peptide YENGEPPMEVYEVLR in the spectrum library LibY-CID-IT. The lower spectrum is a decoy version of this library entry created by the SpectraST algorithm (decoy peptide EYNGEPPYEMVELVR). Even though the target and decoy peptides have a sequence similarity lower than 0.8 (the similarity threshold implemented in the SpectraST algorithm) target and decoy library entries have a very high spectral similarity (dot-product score 0.94), resulting from shared high intensity sequence ions (solid black lines) and non-annotated ions (dotted lines). Solid gray lines show the non-matching peaks.

4) Concluding Remarks

In summary we have shown that the shuffling and repositioning algorithm implemented in DeLiberator produces sufficiently realistic decoy spectra. In addition we argue that decoy databases should be evaluated in terms of their similarity to the target database and conclude that strictly controlling the similarity between a target spectrum and its decoy version leads to a less conservative and more accurate estimation of the FDR.
Peptide fragmentation is a complex process therefore MS2 spectra are typically significantly different from their theoretical counterparts modeled by an MS2 sequence search tool. As a result sophisticated algorithms have been developed to calculate similarity scores of PSMs. Contrary to MS2 sequence search engines, relatively little attention has been paid to scoring issues for spectral library search tools. Studying the peak matching patterns of incorrect SSM provide guidance on how to develop more discriminant SSM scoring functions. We have seen that intense peaks of specific ion-types are frequently shared among spectra of the same charge state and similar precursor mass. Furthermore our analysis shows that high scoring SSM of distinct peptides often share multiple peaks originating from the same fragmentation site. Thus a data-type specific SSM scoring scheme which considers the ion-type and sequence position of the matched peaks may prove advantageous.

SRM is becoming the method of preference for targeted protein quantification [30]. Here proteins are quantified based on a limited list of MS2 fragment ions of proteotypic peptides; characterized by their uniqueness for a single protein and their detectability in a mass spectrometer. Peptide transitions need to be selected with care in order to avoid interfering signal from co-eluting peptides. Once again frequently occurring ions of distinct peptides should be avoided when selecting the most discriminative transitions for a given peptide and charge state, and the type of analysis presented in this study may be useful in guiding the determination of the best transitions.

5) Acknowledgment

The author would like to appreciate everyone who contributes this work. Without them, the project could not have stepped in advance. Erik Ahrné: for the great contribution of the work and his advice to solve a lot of problems; Markus Müller: for the fruitful discussion and his advice based on his research experience, Frederic Nikitin: for developing Java Proteome Library (JPL), Alexander Scherl: for the provides of ETD data and Frederique Lisacek: for the supervising the project as our group leader. The author also wishes to thank Patricia Palagi who is organizing the programs of Master's degree in Proteomics and Bioinformatics.

6) References

Ms2grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra., 


1285–1292.

Enhanced analysis of metastatic prostate cancer using stable isotopes and high mass accuracy 

[29] Sherman, J., McKay, M. J., Ashman, K., Molloy, M. P., How specific is my srm?: The issue of 