UNIVERSITY OF CALIFORNIA, SAN DIEGO

Maestro: Comprehensive, Multi-Stage Spectrum Identification in Protein Mass Spectrometry

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Computer Science

by

Julie Standig Wertz

Committee in charge:

Professor Nuno Bandeira, Chair
Professor Vineet Bafna
Professor Pavel Pevzner

2017

The Thesis of Julie Standig Wertz is approved, and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

Table of Contents

List of Figures

ABSTRACT OF THE THESIS

Maestro: Comprehensive, Multi-Stage Spectrum Identification in Protein Mass Spectrometry

by

Julie Standig Wertz

Master of Science in Computer Science

University of California, San Diego, 2017

Professor Nuno Bandeira, Chair

Tandem mass spectrometry has become a leading method of analyzing large-scale proteomics data, necessitating fast and accurate computational methods of interpreting mass spectrometer output to identify the contents of protein samples. A wide array of algorithms have been developed to this end – protein database searches are a common approach, but other types of searches, such as spectral library searches, and post-translational modification-based searches, may also be valuable tools. It is often advantageous to run a single dataset through multiple algorithms, given that one search algorithm may be able to identify spectra that other algorithms cannot identify, or identify certain types of spectra more quickly than other algorithms. However, combining the results from multiple searches manually is time-consuming and prone to error, and can make interpretation of the results difficult. The Maestro workflow is introduced here, which runs spectra automatically through multiple search algorithms, aggregates the results, and produces in-depth analyses and visualizations of the data. Maestro identifies a wide variety of peptides and modifications.

Chapter 1

Introduction

1.1 Mass Spectrometry

Mass spectrometry is a technique used to analyze the molecular mass and composition of a sample. In order to perform this analysis, the mass spectrometer first ionizes the sample. Ionization can be accomplished using various techniques, commonly electrospray ionization (ESI) [8] or matrix-assisted laser desorption/ionization (MALDI) [14]. Using ESI, a solution containing the sample is transported through a capillary tube, and a high voltage is applied. The charged molecules repel each other upon exiting the capillary, and after evaporation of the solvent, individual molecules of the sample remain. ESI is a soft ionization method, meaning that it generally does not cause fragmentation of molecular ions. MALDI is another soft ionization method, in which the sample is applied to a matrix, and ions are then released from the matrix via laser. Different ionization methods produce varying degrees of fragmentation and may differ in sensitivity, and in how effectively certain types of samples can be ionized.

After ionization, a mass analyzer separates the ions by their mass-to-charge ratio. Types of mass analyzers include quadrupoles, ion traps, time-of-flight, Fourier transform ion cyclotron resonance, and magnetic sector analyzers. The type of mass analyzer used affects how accurately the mass-to-charge ratio of each peak can be resolved, the range of mass-to-charge ratios covered, and how many spectra are generated per time period. A given mass analyzer may only be compatible with certain ionization methods. Once the ions have been separated by mass-to-charge ratio, a detector records the ion charge/ current, and a plot of

intensity with respect to mass-to-charge ratio is produced.

1.2 Tandem Mass Spectrometry

In tandem mass spectrometry (MS/MS), ions at a specific mass-to-charge ratio are isolated and fragmented. This entails multiple stages of mass analysis, with ion fragmentation taking place between stages. The spectra obtained from the first round of MS are called MS or MS1 spectra and the spectra obtained from the second round are called MS/MS or MS2 spectra. MS1 spectra are representative of intact peptide (precursor) ions, before fragmentation, and MS2 spectra characterize the product ions that are obtained after fragmentation. MS/MS can be performed either by combining multiple mass analyzers, or by using a single mass analyzer and performing each stage of analysis in sequence.

Various fragmentation methods may be used. Collision-induced dissociation (CID) [31] utilizes collisions between ions and neutral molecules to produce fragment ions. Electron-transfer dissociation (ETD) [26] induces cation fragmentation via electron transfer, and works well for highly-charged ions and long peptides. Higher-energy collisional dissociation (HCD) [23] is a type of CID used in conjunction with certain ion trap mass analyzers. It can be advantageous to use CID over ETD when analyzing small and singly-charged peptides, while ETD is preferable for large and multiply-charged peptides, and peptides with post-translational modifications (PTMs).

MS/MS has become a common method of identifying the proteins in a sample. Proteins can be identified by first purifying them, and digesting them into shorter peptides using a protease such as trypsin. The peptides are then fragmented via MS/MS, and the mass-to-charge ratios of the fragments are analyzed. The resulting spectra are characteristic of the peptides that produced them. Computational methods such as sequence database, spectral library, and blind searches can be used to identify the peptide corresponding to each spectrum.

1.3 Clustering

Individual peptides are frequently seen in multiple spectra in tandem mass spectrometry datasets (a single peptide may be repeatedly selected for fragmentation), creating redundancy when determining the set of peptides in a sample. Spectra that are similar to each other (that presumably originate from the same peptide) can be clustered together, so that once one spectrum in a cluster is identified, the other spectra in its cluster receive the same peptide identification. Clustering often speeds up analysis considerably, by reducing the number of spectra to be identified. It can also result in an increase in identifications, since the consensus spectrum for a cluster can be higher-quality, and therefore easier to identify, than some of its component spectra. Clustering can also reveal spectra that occur in multiple runs, even if they are not identified.

MS-Cluster is a clustering algorithm that is able to rapidly process large quantities of spectra [9]. The clustering algorithm starts with each spectrum as an individual node, and performs multiple rounds of merging nodes above a certain similarity threshold. The threshold is decreased each round. When combining spectra, a consensus spectrum is generated by combining the peaks in each spectrum in the cluster. To produce a consensus peak for a peak common to multiple spectra, the scaled sum of the peak intensities and a weighted average of the peak masses are used. Peaks that have low intensity relative to nearby peaks are filtered out. A normalized dot-product is used to determine similarity between consensus spectra. Two heuristics are used to improve clustering efficiency: pairs of spectra that have no overlap in their five highest-intensity peaks are ignored, and spectral similarity computations from earlier rounds of clustering are taken into account in later rounds.

1.4 Molecular Networking

Molecular networking [30] [29] can be used to identify groups of similar spectra using a spectral networks-based approach [1] [13]. Spectral networks analysis avoids having to search spectra against a database by creating pairs of peptides — peptides that differ by a single modification are paired, and peptides for which one is a substring of the other are paired. Edges between the spectra corresponding to paired peptides are created, forming a spectral network. The vertices of the network are clusters, and the edges are

cluster pairs.

Rather than using 3-to-4-amino acid tags to reconstruct the peptide sequence (as with database search), 7-to-9-amino acid peptides can be looked up in a hash of the database. This means that when identifying one spectrum in a spectral pair, the spectrum of the neighboring peptide can be used directly, without searching a database to find the peptide that corresponds to the query spectrum. The peptides can be reconstructed de novo.

Modifications are discovered by identifying a difference in parent mass between the spectra in a spectral pair. Once a spectrum is identified, spectral networking allows the masses of modifications on neighboring peptides to be found, as well as the approximate amino acids on which the modifications occurred. Unannotated neighboring spectra are annotated using this information. Unannotated spectra paired to the neighboring spectra are then annotated based on the annotations of the neighboring spectra. Annotations are propagated with each iteration in this way until the search is completed.

Spectral networking has several applications. It can be used for shotgun protein sequencing, by digesting a protein sample and forming a network from the resulting spectra. It also works especially well when finding uncommon PTMs; these might not be included in a spectral library or sequence database, but spectral networks analysis does not require knowing possible modifications in advance. Spectral pairing confers certain advantages. The pairing of unmodified and modified peptides helps to reduce noise. The spectral pairing approach takes advantage of the pairing of overlapping peptides, which is beneficial when processing nontryptic peptides, and tryptic peptides paired with semitryptic peptides. A disadvantage of spectral networking is that it has difficulty distinguishing modifications with small offsets from noise.

1.5 Database Search

A common method of identifying the peptides represented by mass spectra is to perform a sequence database search [2] [4] [7] [11] [27]. This involves comparing each spectrum to a database of known protein sequences (often in FASTA format) belonging to the organism of interest. Common contaminants, sequences containing mutations, and decoys are often included in the database.

In order to ascertain a peptide from a spectrum, the chain of differences in mass-to-charge ratio

between spectrum peaks is considered. Specific sequences of differences are known to correspond to certain peptides of length 3 to 5 amino acids, known as peptide sequence tags. The protein sequences in the database are virtually fragmented using same the fragmentation method that was used to process the original protein sample. Only database peptides that have a similar mass to the precursor mass are considered as matches; a parent/ precursor mass tolerance threshold defines how similar these two masses need to be. The optimal setting for this parameter depends on the accuracy of the MS1 mass analyzer.

The peptide sequence tags inferred from a spectrum are combined with the flanking masses and compared to the fragmented database protein sequences with appropriate parent mass in order to find the protein sequence that corresponds to that spectrum. Using this method allows for the discovery of post-translational modifications, since the mass of the inferred peptide will differ from the database mass if a PTM is present. PTM discovery requires that peptides with certain characteristic mass deviations relative to query spectra be included as candidates for spectrum identification. Possible sources of mass deviation include static mass changes to residues (such as cysteine alkylation), or changes that may occur on some residues and not on others (such as oxidation and lysine methylation).

In order to score peptide-spectrum matches, the query spectrum is compared to the spectrum predicted based on a given database peptide sequence. The number of matching peaks and peak intensities, possibly ignoring low-intensity peaks, are taken into account in computing the score. A shift in mass-to-charge ratio of one spectrum relative to the other may be applied. The scoring algorithm may also take into account how much better the best match is than other matches.

1.6 Library Search

In order to obtain high specificity in identifying the peptides corresponding to MS/MS spectra, it can be advantageous to perform a spectral library search [5] [20] [25] [34]. With this method, each query spectrum is compared to a library of previously-identified spectra collected under similar conditions, and the similarity between pairs of spectra is examined in order to match the query spectrum to a library spectrum. This type of search was originally used to identify small molecules, but has since proven useful in identifying proteins, as well.

In order to create a spectral library, a protein sample obtained from the organism of interest is analyzed using a method such as sequence database search. Multiple spectra are usually obtained for each library peptide. One method of combining the peptides corresponding to each spectrum is to create a consensus spectrum averaging the individual spectra, thus lowering redundancy and reducing the amount of noise in the library spectrum. High-quality spectra are often weighted more heavily than low-quality spectra in forming a consensus. Spectra with more replicates are more likely to produce a high-quality consensus spectrum, and are usually more likely to be included in a library than spectra with fewer replicates. (Some libraries exclude spectra with no replicates.) Alternatively, the highest-quality spectrum is sometimes retained in the library instead of forming a consensus. This method generally results in lower-quality spectra than the consensus method. Comprehensive spectral libraries are available from various sources, notably the National Institute of Standard and Technology (NIST). One disadvantage of performing a spectral library search is that it limits the number of modified peptides that can be identified, since spectral libraries typically include few spectra representing modified peptides.

In computing the similarity between a library and query spectrum, it is important to allow for the matching of spectra that are not completely identical — factors such as noise and contaminants can cause two spectra to be slightly different from each other, even though both spectra represent the same peptide. However, if this similarity threshold is set too low, there is an increased likelihood of an experimental spectrum being incorrectly matched to a library spectrum that does not correspond to that peptide. In order to maximize the number of correct matches while minimizing the number of incorrect matches, spectra can be preprocessed to remove low-quality spectra and low-intensity peaks.

A scoring function can be applied to quantify the extent to which two spectra are similar. One way to compute score is to bin peak intensity by mass-to-charge ratio, resulting in an intensity vector with one entry for each bin. The intensity vector of one spectrum is then compared with the intensity vector of the other spectrum in order to determine similarity. By comparing vectors, the number of peaks shared between the two spectra (which does not take intensity into account), and the dot product (cosine) of the two vectors (which strongly weights intensity) can be computed, for instance. How much better the best match of a library spectrum to a query spectrum is than other matches is also an indication of the quality of the match

(this difference will generally be large for a good match).

The score boundary between a spectrum being considered unidentified and considered identified is often defined by target-decoy searching. This involves including decoy spectra in the library that are not intended to match with any of the query spectra. A target ratio of false (decoy) matches to real (target) matches is set, and the score cut-off is determined accordingly. The decoy-to-target ratio is called the false discovery rate (FDR).

1.7 M-SPLIT

Mixture-Spectrum Partitioning using a Library of Identified Tandem mass spectra (M-SPLIT) is a spectral library search method that allows for the identification of spectra generated from individual peptides, as well as from mixtures containing two different peptides [28]. This type of mixture can arise when two peptides elute from the chromatography column at similar times, which can occur due to PTMs, for instance. M-SPLIT allows for more spectra to potentially be identified for a given amount of time spent on data collection than a search method which does not handle mixtures, and is effective at peptide ratios of up to 10:1 (although at higher ratios it becomes more difficult to distinguish peaks produced by the less-abundant peptide from noise).

In order to identify a mixture spectrum, the mixture is represented as a linear combination of individual library spectra. It is very time-consuming to search all pairs of spectra in the spectral library to determine which two comprise the mixture corresponding to a given query spectrum. Therefore, projected cosine-based filtration and branch-and-bound strategies are employed to reduce the number of candidate spectrum pairs. Projected cosine filtration constrains the number of individual spectra considered, while the branch-and-bound strategy constrains the number of pairs of spectra that are considered as matches to the query mixture spectrum.

When considering the similarity between a library spectrum and a query mixture spectrum, the query spectrum peaks that do not correspond to any peak in the library spectrum (which might arise from the other mixture component) are ignored. After subtraction of non-shared peaks, the cosine similarity of two spectra is determined. Using this projected cosine-based method, the library spectra which best correspond to the

query mixture spectrum are retained, while lower-scoring spectra are eliminated from consideration.

In order to perform branch-and-bound filtration, library spectra are sorted by cosine similarity to the query mixture spectrum. Every spectrum paired with a possible correct spectrum will have a cosine that is close to the highest cosine; specifically, the upper bound on the cosine will be greater than the highest cosine. Thus, the library spectrum with highest cosine similarity to the mixture spectrum is paired with other library spectra until a spectrum is found such that the upper bound of the cosine of the pair of library spectra with respect to the mixture spectrum is lower than the current highest cosine of any library spectrum pair with respect to the mixture spectrum. The original highest-similarity spectrum is then deleted from the library, and the process is repeated with the new highest-similarity spectrum.

In order to calculate cosine similarity between the mixture spectrum and library spectra, it is important to accurately estimate the ratio of the abundance of one library individual peptide spectrum to the abundance of the other library spectrum. One method of making this estimation is to remove peaks from the mixture spectrum that are in the library spectrum of the dominant peptide, and to calculate the ratio from the residual spectrum. It is also feasible to choose the ratio that maximizes cosine similarity between the mixture spectrum and the combination of library spectra. (This method performs better with uneven mixtures than the residual spectrum method.)

A given query spectrum must be established as either unidentified, as a single-peptide spectrum, or as a mixture spectrum. If the mixture cosine similarity is significantly higher than the cosine similarity for either individual library spectrum with respect to the query spectrum, the query spectrum is identified as a mixture. If the cosine similarities for both individual library spectra are low, the spectrum is considered unidentified.

1.8 MS-GF+

MS-GF+ [18] is a general-purpose sequence database search algorithm that extends MS-GFDB [17], which extends MS-GF (Mass Spectrometry - Generating Function) [16]. It works with several different fragmentation methods and enzymes. MS-GF+ can recognize post-translational modifications, which are specified as an input to the search.

In order to match peptides in the database with query spectra, a suffix array of the database is created, and each peptide in the suffix array is compared to spectra with a matching precursor mass. The quality and statistical significance of peptide-spectrum matches must then be determined. The spectrum is converted into a spectral vector (prefix residue mass spectrum), which has a score associated with each mass less than or equal to the parent mass. (This score is related to the probability that the peptide represented by the spectrum has a prefix with the given mass.) Masses are rounded to the nearest integer so that scores can be calculated quickly, and rescaled in order to reduce rounding errors. The score of a peptide-spectrum match is defined as the dot product of the vector representing the prefix residue mass spectrum and the vector representing the spectrum of the peptide. The statistical significance of each peptide-spectrum match is calculated using an E-value, based on the score distribution of all peptides.

In order to adhere to a given false discovery rate, a decoy database is generated by reversing each protein in the target database, and concatenated to the target database. The search is performed against the combined database, and an E-value threshold is chosen such that the given false discovery rate is maintained.

1.9 MODa

MODification via alignment (MODa) is a search algorithm that focuses on improving identification of spectra obtained from samples containing modified peptides [21]. Whereas most search methods require that possible PTMs be specified prior to the search, MODa does not have this requirement. Rather, MODa is an unrestrictive/ blind algorithm, meaning that PTMs are determined directly from experimental data and every PTM type is searched for simultaneously. Using an unrestrictive algorithm often means that a large proportion of spectra receive the wrong peptide identification (these identifications are referred to as false positives), or fail to be identified by the search (false negatives), at a given FDR. The possibility of multiple PTMs occurring on a single peptide also introduces a large amount of variability in the possible composition of the peptide including PTMs, compounding the issue of false negatives and false positives.

Despite the large search space, MODa performs a comparatively fast and accurate unrestrictive search, and constrains the number of false negatives and false positives. These results are achieved in part by inferring multiple 2-to-4-amino acid sequence tags from each spectrum and aligning those to a

protein database. Using sequence tags means that fewer database peptides need to be considered as matches for each spectrum, and also allows for PTMs to be identified by calculating the mass differences between experimental and theoretical peptides (this alignment is performed using dynamic programming).

The MODa algorithm addresses the possibility of multiple PTMs occurring on a single peptide by allowing for an unlimited number of modifications per peptide (peptides with more than one different tag indicate the presence of database peptides with multiple PTMs). MODa is particularly useful when attempting to identify new or uncommon PTMs, since many other unrestrictive searches run slowly, and searches that require specifying PTMs in advance are unlikely to identify rare PTMs.

Chapter 2

Maestro Workflow Implementation

2.1 Workflow Overview

Each method of peptide identification has advantages over other methods, in terms of speed, flexibility, specificity, and other metrics. Thus, when analyzing a set of spectra, it is beneficial to combine multiple types of searches in order to efficiently identify as many peptides as possible. It is also important to provide automatic analysis and visualization of the search results in order to allow for the user to easily focus in on interesting data, obtain a summary of the results, acquire in-depth information about each spectrum/ cluster/ network, or meet other objectives.

The Maestro workflow was created to this end, and combines the previously-described M-SPLIT, MS-GF+, and MODa algorithms into a single, comprehensive workflow. The workflow starts by pre-processing spectra and clustering them using MS-Cluster. M-SPLIT is then run on all spectra, MS-GF+ is run on the spectra that were not identified by M-SPLIT after FDR filtering, then MODa is run on the spectra that were not identified by M-SPLIT or MS-GF+ after FDR filtering. MODa is run with a reduced database that contains only proteins corresponding to clusters identified by M-SPLIT or MS-GF+. Peptide variant-level FDR is performed on the clusters passing cluster-level FDR, to determine which variants were identified. Molecular networking is not directly used for identification, but is performed alongside the three identification algorithms for verification purposes. Extensive data analysis and compilation of the results from the search algorithms, clustering, molecular networking, and other workflow nodes, is added to the

end of the workflow. The main workflow is shown in Figure 2.1(a). The result view generation part of the workflow is shown in Figure 2.1(b).

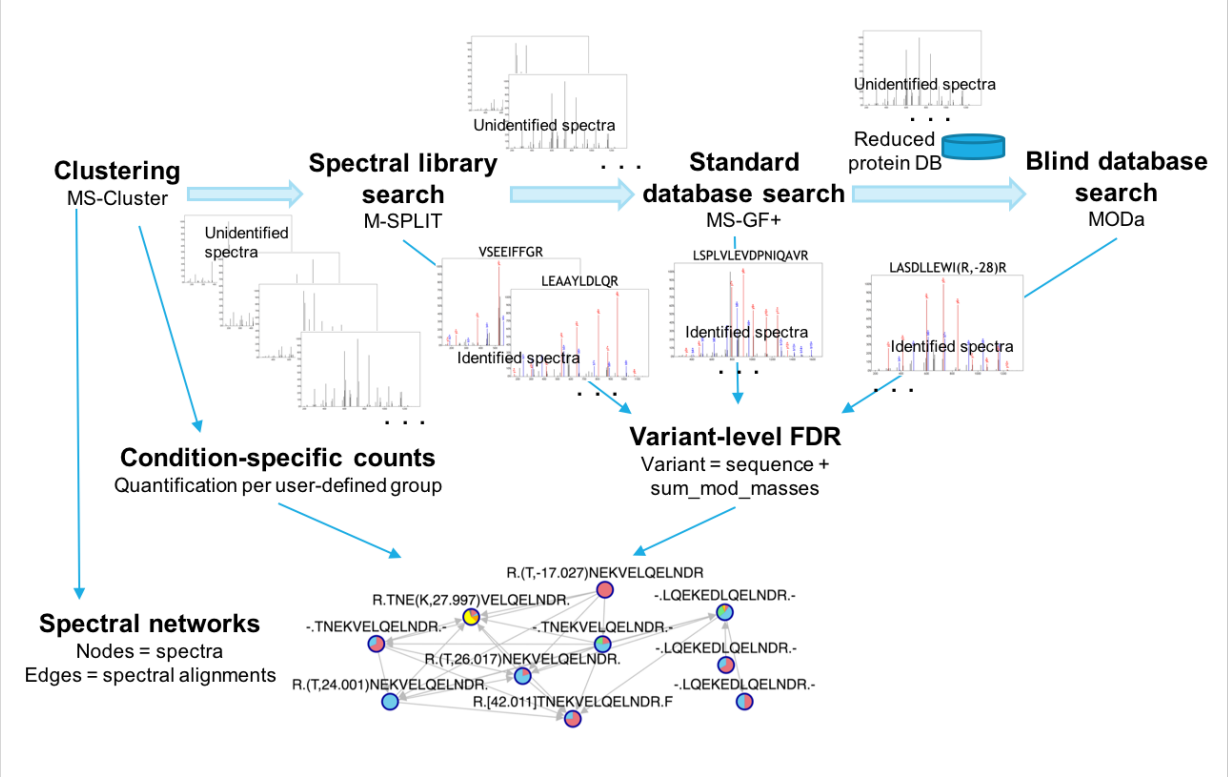


Figure 2.1 a) Main workflow. Spectra are clustered with MS-Cluster, and M-SPLIT library search is run on the full set of clusters. The M-SPLIT PSMs are FDR-filtered, and MS-GF+ database search is run on the clusters that were not identified by M-SPLIT. The MS-GF+ results are FDR-filtered, and MODa blind search is run on the clusters that were not identified by either M-SPLIT or MS-GF+. Variant-level FDR is performed on the set of combined PSMs from the three algorithms. Additionally, spectral networks are created from the set of clusters, and spectral counting is performed based on the group membership of the input spectra.

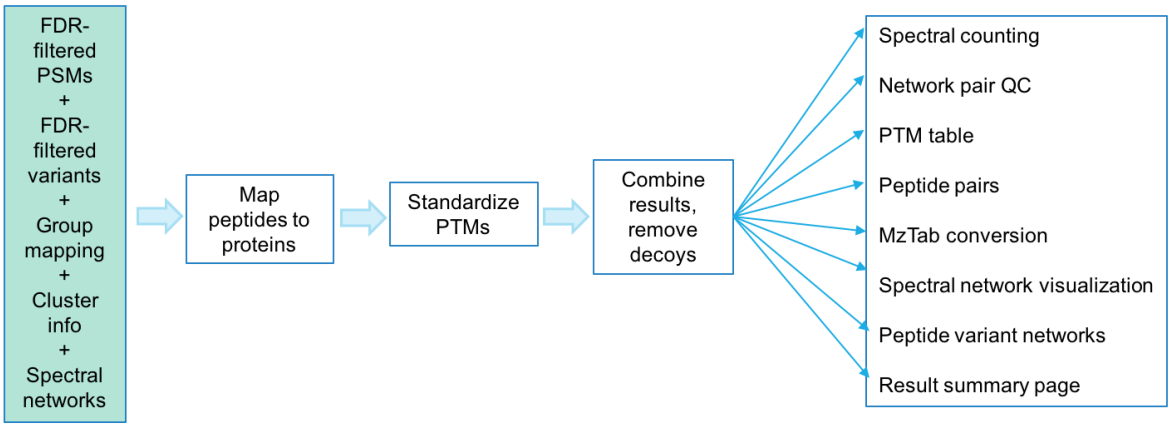


Figure 2.1 b) Creation of result views. At the conclusion of the search, the PSMs from each algorithm are combined with spectral networking results, as well as information about each spectrum. Per-spectrum results are created. Peptide modifications are converted into a standard format, and spectral counting is performed based on the group membership of each spectrum. These processed results are further processed in order to create several result views, including per-peptide and per-cluster views, a spectral networks view, a search summary page, a network pairs view, and a peptide pairs view.

2.2 ProteoSAFe Environment

The computational mass spectrometry platform ProteoSAFe (Proteomics Environment which is Scalable in utilizing distributed computing, Accessible via reconfigurable, easy-to-learn user interfaces, and Flexible in tool chaining) was used to create and run the workflow. ProteoSAFe allows for developer-created tools to be combined into workflows via XML workflow designations.

To run a workflow, first the user selects a workflow from the menu. Parameters for that particular workflow are displayed. The user selects values for these parameters, and submits the task. A graphic all of the nodes (tools) in the workflow, with colors indicating how far the task has progressed, is then displayed. If a node has a green border, that indicates that the task has completed. Nodes in progress are displayed with an orange border, and nodes with a black border have not yet been run. Each tool in the workflow is run on a computing cluster, and execution may be parallelized on a per-file basis. A running or completed task has a "Clone" button, which directs the user to the input form, filled out with the exact parameters used for the task. This is useful for seeing what parameters were used for a task, and for re-running a task with

minor changes. Tasks may also be restarted (which is useful in case a task encounters an error which was subsequently fixed), and deleted.

Once a task has completed, result views are displayed. One type of available result view is a server-side table with customizable columns, filtering and sorting functionality, and various other features. ProteoSAFe can invoke Lorikeet (see below) to display spectra from within these tables.

The implementation of each tool unique to this workflow will be described in detail. Pre-existing tools that are important to the workflow will be mentioned, as well.

2.3 Mass Correction, Charge Correction, and Kullback-Leibler Filtering

Precursor mass and charge correction are performed by precursor isotopic patterns in MS1 survey scans considering that a) the assigned precursor mass could be 1 isotope below or up to 3 isotopes above the theoretical mass of the peptide and b) the precursor charge is in the range of 1 to 5. Precursor mass and charge are then assigned by minimizing the Kullback-Leibler (KL) distance between the experimental and the theoretical isotopic patterns at each precursor mass/charge; theoretical distributions are calculated using averagine masses [22].

The workflow can also remove spectra whose precursor isotope patterns are not similar enough to the theoretical isotope pattern or not free from interference (e.g., coeluting precursors). The parameter "KL Filtering Threshold" (under "Advanced Filtering Options") indicates the KL threshold for the maximum divergence between the experimental and theoretical isotope patterns, where the theoretical distributions are calculated using averagine masses (as for precursor mass/charge reassignments) and also required to have zero intensity between theoretical isotope peaks. Spectra from precursors with isotope patterns having a KL divergence above the threshold are removed from further analysis. If the KL filtering threshold is set to 0, no spectra are filtered out, but mass correction and charge correction are still performed.

2.4 Clustering

If the "Run MS Cluster" parameter is selected, similar spectra are clustered together via the MS-Cluster algorithm.

2.5 Molecular Networking

Spectral networks are created from the spectral clusters. Each network node (cluster) contains associated information such as the consensus spectrum, cluster index, parent mass, number of spectra in the cluster, default groups to which the cluster belongs, charge, and peptide identification (if the cluster was identified). Each network pair (pair of nodes connected by an edge) also contains associated information, such as the cosine and difference in mass-to-charge ratio between the two clusters. These spectral networks can be used for several purposes, including manually validating clustering and identifications, inferring identifications for clusters that were not identified or that were incorrectly identified, and exploring peptide diversity.

2.6 Tag-based Pair Filtering

First, spectra are deconvoluted (peaks are converted to charge 1). If "Tag-based Pair Filter" in the input form is set to "yes", a tag file is generated. The number of tags per spectrum is also specified in the input form. Spectral pairs are then filtered, using tags if they were generated.

2.7 Parallelization on Spectrum Files

An MGF-format file with the combined contents of the mass-corrected input spectrum files, is split into 20 smaller MGF files, assuming that the original file is sufficiently large. These smaller MGF files are used as inputs during further workflow steps that use input spectra, since it is convenient for input spectrum files to be in MGF format, and the division of input files allows for per-file parallelization.

2.8 M-SPLIT

An M-SPLIT spectral library search is then performed on the spectra. The M-SPLIT results from each file are merged and filtered according to the spectrum-level false discovery rate specified by the user. A support vector machine (SVM) is used to determine significance of the results. This step is performed using the SVM-Light package. FDR filtering is performed separately on non-mixtures/ first components of mixtures (using the SVM1 score), and second components of mixtures (using the SVM2 score).

2.9 Subtraction of Identified Spectra

Spectra identified by M-SPLIT are eliminated from further processing by creating copies of the input spectrum files that exclude the spectra identified by M-SPLIT. This decreases the number of spectra run through subsequent search algorithms, which allows those searches to run faster.

2.10 MS-GF+

After removal of the spectra not identified by M-SPLIT, an MS-GF+ sequence database search is performed on the remaining spectra, followed by spectrum-level FDR filtering based on E-value.

2.11 MODa

Spectra that were identified by MS-GF+ are removed. A reduced database is created that contains only proteins corresponding to clusters identified by M-SPLIT or MS-GF+. This strategy can speed up MODa analysis, and is likely to produce more results at a given FDR than would be obtained with the full database. A MODa search is then performed on the spectra that were not identified by M-SPLIT or MS-GF+ against the reduced database, followed by spectrum-level FDR filtering. When performing FDR filtering, results are split by charge state and tripticity (whether the peptide has 0, 1 or 2 proper N-terminal and C-terminal tryptic cleavages, using a probability value between 0 and 1 computed by MODa).

2.12 Standardization of Peptide Modification Format

Peptides identified by all three algorithms are converted to a standard format. Non-N-terminal modifications are converted to the form (m, n), where m is the residue on which the modification occurs and n is the offset. N-terminal modifications are changed to the form [n]. Default flanking residues (hyphens) are added to peptides that lack flanking residues. All cysteine residues in each peptide string are modified with the protecting group mass if not otherwise modified.

2.13 Protein Reassignment

Since some algorithms only output a single protein containing a given peptide, the list of all proteins corresponding to each peptide must be found separately. Each peptide is mapped to the protein database, and a column containing a semicolon-separated list of all of the proteins containing the peptide string is added to the result files.

2.14 MzTab Conversion

An MzTab-formatted result file [12] is created based on the identified raw input spectra.

2.15 Peptide Variant Analysis

If a group of peptides have the same amino acid sequence, and the sum of their modification mass offsets are within the parent mass tolerance of each other, these peptides are considered to represent the same variant (and therefore belong to the same variant group). Variant analysis is used instead of peptide analysis because modifications are often localized incorrectly, and it is difficult to determine whether modifications are accurate without manually inspecting the spectra. Any two peptides in the same variant group may contain the same underlying modification.

A variant-level false discovery rate (as specified in the input form) is used to determine which variants to output. This is calculated using the q-score, which is comparable across different search algorithms,

and is defined as the ratio of target matches to decoy matches at or above the current position in the list of peptide-spectrum matches sorted by score or E-value. Q-scores are obtained separately from the spectrum-level false discovery rate results for non-mixtures/ first components of mixtures identified by M-SPLIT, second component of mixtures identified by M-SPLIT, MSGF+, and MODa. The peptide-spectrum match with the highest q-score for a given variant is chosen as the representative of that variant. The q-score threshold is then determined using the representative target and decoy peptide-spectrum matches, and the variants passing the threshold are retained.

Peptide variant regions (PVRs) are computed based on the variant-level FDR output. A PVR is a region of a protein where at least half of each identified peptide in the region overlaps with another identified peptide in the region.

2.16 Peptide Pair Categorization

MS-GF+ and MODa peptides are grouped into one (or more) of 7 categories by comparing them with peptides identified by the other searches. Peptides identified by MS-GF+ are compared with peptides identified by M-SPLIT, and peptides identified by MODa are compared with peptides identified by M-SPLIT and with peptides identified by MS-GF+. The possible categories for an MSGF+-MSPLIT peptide pair (meaning that each MS-GF+ peptide is being grouped with respect to the set of M-SPLIT peptides) are as follows:

Category 0: The MS-GF+ peptide has exact same sequence, the same number of modifications, and the same charge as an M-SPLIT peptide.

Category 1: The MS-GF+ peptide has the exact same sequence, the same number of modifications, and a different charge from an M-SPLIT peptide.

Category 2: The MS-GF+ peptide sequence is contained within an M-SPLIT peptide sequence.

Category 3: An M-SPLIT peptide sequence is contained within the MS-GF+ peptide sequence.

Category 4: An M-SPLIT peptide has the exact same sequence and more modifications than the MS-GF+ peptide.

Category 5: An M-SPLIT peptide has the exact same sequence and fewer modifications than the MS-GF+ peptide.

Category 6: Exactly one residue is changed to another in the MS-GF+ peptide with respect to an M-SPLIT peptide.

Category 7: Other.

The categories for the other two types of peptide pairs (MODA-MSPLIT pairs and MODA-MSGF+ pairs) are defined analogously. It is possible for one peptide to be in multiple categories (e.g., an MS-GF+ peptide is identical to one M-SPLIT peptide and is a substring of another M-SPLIT peptide).

2.17 Grouping/ Counts: Aggregation of Results from Each Search

Information from the M-SPLIT, MS-GF+, and MODa searches, as well as group mapping, cluster and network information files, is compiled into four files containing extensive information about each cluster. A file containing clusters identified by cluster-level FDR, a file containing peptide variants identified by variant-level FDR, and a file containing proteins for which component variants were identified, are produced. Two versions of the cluster-level FDR file are produced, one in which each mixture component is listed in a separate row, and one in which mixtures are listed as single-line entries with exclamation-point delimiters. These output files are directly shown in various result views, including the "Identified Clusters", "Identified Variants", "Identified Variants (Merged Protein Regions)", and "Identified Proteins" views (see below).

To perform spectral counting, the number of spectra for each variant (variant spectrum count) is calculated by adding up the spectra in the clusters identified as that variant. The variant spectrum count is stored separately for each default and user-defined group, so that the spectra per variant in each group can be determined in addition to the total spectra per variant. The overall and per-group protein spectrum counts, and overall and per-group unique protein spectrum counts (in which the peptide maps only to a single protein) are determined in the same way. Variant, protein, and unique protein spectrum counts are calculated separately aggregating over clusters and aggregating over variants. This produces spectrum counts for each cluster (to be used in result views that group by cluster) as well as for each variant (to be used in result

views that group by variant). For the latter, the protein and unique protein spectrum counts are calculated using the protein for each non-merged variant (rather than the list of all proteins) and sum over each cluster corresponding to the peptide, not just the representative cluster for the variant. Variant counts are only summed over the representative cluster for the variant, not all clusters identified as the variant. Outlier default and user groups are determined by taking the base-2 logarithm of the ratio of the highest number of spectra in a group to the second-highest number of spectra in a group, or the lowest number of spectra in a group to the second-lowest number of spectra in a group, whichever has greater absolute value (100 is used in place of infinity and -100 is used in place of negative infinity). The outlier group is considered to be the group with the largest absolute outlier group ratio.

Each output file contains for each cluster/ variant/ protein the algorithm, filename, cluster index, peptide, unmodified peptide, identification charge, groups to which the cluster belongs, number of network neighbors, user groups, default groups, proteins, precursor charge, number of network neighbors for the peptide, number of network neighbors for the variant, number of network neighbors for the protein, number of modifications, sum of modifications, list of modifications, start amino acid on the protein, end amino acid on the protein, variant group, spectral network component index, spectral network component indices for the variant, protein region, FDR, spectral probability, MQ score, exact mass, variant-level FDR, variants per unmodified peptide, cluster indices corresponding to the variant, spectral counts, and other information. Total and per-default and per-user group counts and outlier groups and outlier group ratios are output for clusters, peptides, peptide variants, proteins, and unique proteins (proteins for which the peptide maps only to that protein).

2.18 Spectral Networks

An interactive spectral network visualization is created from the nodes and edges in the network, and the spectrum identifications, using JavaScript code that invokes Cytoscape. This code makes use of the Cytoscape.js library [10].

2.19 Variant Networks

Peptide variant networks are created from spectral network nodes and edges. This is done by merging the spectral network nodes corresponding to each variant group into a single node. The edge with the highest cosine is chosen as the representative edge between variant nodes. Edges between variants that do not have any protein variant regions in common, are removed. Network components containing the same unmodified peptide are merged. Variant networks consist only of identified nodes.

2.20 Network Pair QC

Actual and theoretical network edges are output by this tool. For each network pair, the output shows basic information about each PSM in the pair, and indicates whether the pair is present in the network, and whether the pair is correct.

2.21 Summarization of Results

This tool produces an HTML file containing the overall statistics displayed in the "Summary Report" result view. The precursor charges and default groups corresponding to each cluster index are obtained from a cluster information file. The modifications searched for during MS-GF+ are obtained from a parameters file. Each type of modification has a name, an offset, residues, and options associated with it, which are included in the tool output. The Spectrum Counts output file is parsed to obtain the number of raw and identified spectra for each precursor charge, number of clusters for each precursor charge, number of clusters identified by each workflow (M-SPLIT, MS-GF+, and MODa) for each precursor charge, and unique unmodified versions of peptide sequences in each group. The fifteen most common mass offsets, and the number of occurrences of each, are also included in the output file.

2.22 Spectrum Viewer

Lorikeet is a JQuery plugin for visualizing, annotating, and analyzing MS/MS spectra, and works with both unmodified and modified peptides [24]. The spectrum filename, scan number, and peptide annotation may be passed as parameters to Lorikeet. When a Lorikeet icon in a ProteoSAFe result table is clicked, an image of the annotated spectrum is shown. The user can change the annotation, choose which type of ions to display, change the scaling of the spectrum, and manipulate various other spectrum parameters. Peptide peaks that are present in the spectrum are highlighted with colors corresponding to the ion type. Peptide modifications are supported.

Chapter 3

Running the Maestro Workflow

In order to run the workflow, the first step is to visit the ProteoSAFe website and to select "Maestro" from the Workflow category. The following parameters are then specified:

3.1 Basic Options

Spectral library: The spectral library is matched to experimental spectra during the M-SPLIT portion of the search. Built-in NIST, SWATH Atlas, and MassIVE-KB libraries are present in the CCMS_SpectralLibraries directory in the Select Input Files window.

Spectrum files: At least one input MS/MS spectrum file is required. Possible formats include mzXML, MGF and mzML. If spectrum files are grouped using default groups G1 through G6, or using groups specified in a group mapping file uploaded by the user, group-specific results and statistics will be provided after completion of the search. Spectra derived from the same type of tissue might be grouped together, for instance.

Sequence database: Experimental spectra are compared to the sequences in the selected protein database. Built-in human, mouse, and yeast Uniprot databases are present in the CCMS_ProteomeDatabases directory in the Select Input Files window.

Group mapping: A group mapping file can be used instead of or alongside assigning default groups in the input file selection window. A maximum of 16 user-defined groups will be displayed in result views,

although downloaded result view files will contain all groups.

Attribute mapping: An attribute file can be used to further organize groups, in which case the attributes will appear as columns in the output.

Instrument: The type of tandem mass spectrometer used to generate the spectra. Options are ESI-ION-TRAP and QTOF. The instrument type is used to determine scoring.

Cysteine protecting group: The modification that was made to the cysteine residues in the protein sample in order to decrease cysteine side-chain reactivity. Options are carbamidomethylation (+57), carboxymethylation (+58), NIPIA/NIPCAM (+99), and none.

Number of allowed 13C: The number of mass units by which the molecular ion peak can be shifted. Carbon-13 (a naturally-occurring isotope) shifts the molecular ion peak one mass unit higher. Options are 0 through 2.

Parent mass tolerance: Parent mass tolerance (PMT) is the maximum precursor mass difference between a known spectrum and a query spectrum, in order for them to be considered as a possible match. PMT can be specified in daltons (absolute units) or parts per million (fractional units), and should be determined based on the accuracy of the mass analyzer. For certain types of mass analyzers, mass accuracy might differ between higher and lower masses, in which case it is preferable to specify mass tolerance in ppm. PMT is specified separately for M-SPLIT, MS-GF+, and MODa, and peptide variants.

Fragmentation method: Method by which molecular ions were fragmented. Options are "Specified in spectrum file", "CID", "ETD", "HCD", and "Merge spectra from the same precursor". The last option indicates that multiple spectra were acquired from the same precursor ion (CID/ETD pairs may have been obtained, for instance), and that the aggregate spectrum should be searched instead of the individual spectra.

Protease: The enzyme used for protein digestion. Options are trypsin, chymotrypsin, Lys-C, Lys-N, Arg-C, Glu-C, Asp-N, and none. During database search, the protein database is virtually digested using this enzyme to obtain the known PSMs.

Number of allowed non-enzymatic termini: Maximum number of termini that do not match the cleavage specificity of the selected enzyme. PSMs with non-enzymatic termini receive a lower score. Options are 0 through 2.

Ion tolerance: Maximum shift of b and y peaks from expected masses, in daltons. Ion tolerance is important (along with parent mass tolerance) in reducing false positive identifications.

Include common contaminants: If this parameter is selected, common contaminants (trypsin and keratin) are included in the database used for database search.

Clustering on/ off: If selected, similar spectra will be clustered together using the MS-Cluster algorithm at the beginning of the workflow.

Minimum cluster size: The minimum number of spectra in a cluster for the cluster to be retained.

3.2 Advanced Spectral Network Options

Minimum pair cosine: The score threshold in order for a pair to be accepted in the network.

Maximum connected component size: The maximum size of a connected component in order for it to be included in the network. For networks that are larger than this size, the network will be divided into smaller networks by increasing the cosine threshold for the particular network.

Minimum matched peaks: The minimum number of common peaks between two spectra for inclusion in the network.

Apply tag-based pair filter: Whether to use tags to filter network edges.

Number of tags per spectrum: Minimum tag size to use for tag-based network pair filter.

Deconvolute MS/MS: Convert multiply-charged peaks to a charge state of 1.

3.3 Advanced Filtering Options

PSM-Level FDR: The false discovery rate (FDR) is used for statistical validation of the results and is defined as the target ratio of decoy (false) matches to target matches. The FDR determines the score threshold of spectrum or peptide matches, in order for a spectrum to be considered identified. FDR is specified separately for M-SPLIT, MS-GF+, and MODa.

Variant-Level FDR: Variant-level FDR used to filter results from each algorithm. This is specified separately for M-SPLIT, MS-GF+, and MODa.

Overall Variant-Level FDR: Variant-level FDR used to filter combined results from all three algorithms.

Standard deviation for peak filtering: The least intense spectrum peaks (relative to the 25 percent of peaks with lowest peak intensity) are filtered out according to this parameter.

Filter precursor window: If selected, peaks near the precursor mass are removed.

Filter peaks in 50Da window: If selected, every peak that is not within the top 6 most intense peaks in the range from 50 daltons below its mass-to-charge ratio to 50 daltons above, will be removed.

Minimum peak intensity: The lowest possible intensity in order for a peak to be retained in a spectrum.

M-SPLIT SVM1 Threshold: M-SPLIT PSMs with SVM1 scores below this threshold will be filtered out prior to FDR calculation.

KL Filtering Threshold: Spectra with a Kullback-Leibler divergence above this value will be filtered out at the beginning of the workflow.

3.4 Allowed Post-Translational Modifications

MODa Blind mode: Options are blind search and multi-blind search. For blind search, only one modification per peptide is allowed, whereas for multi-blind search, multiple modifications can occur on a single peptide.

Modification mass range: The mass range of modifications to be considered. Specified in daltons.

Maximum number of PTMs per peptide: Indicates specifically how many PTMs can occur per peptide (as opposed to blind mode, which indicates only whether one or more than one modification per peptide is permitted).

PTMs: Possible post-translational modifications to be considered. Options are oxidation, lysine methylation, pyroglutamate formation, phosphorylation, N-terminal carbamylation, N-terminal acetylation, and deamidation. The user can also specify custom modifications.

Chapter 4

Maestro Result Views

4.1 Search Result Summary

This view is a good place to start when looking at results, and is an HTML file consisting of several tables. The first table displays overall information about the search, such as how many spectra were searched and identified, how many spectra are in networks, how many clusters were created and identified, and how many variants and proteins were discovered. These numbers are listed in total, and per charge (Figure 4.1(a)). The prevalence of searched modifications is shown in the next table. It lists each searched modification, and how many PSMs containing that modification were found (Figure 4.1(b)). The prevalence of discovered modifications is shown in the next table. The number of PSMs and variants at various mass offsets, and the most common amino acids on which each mass offset occurs, are listed. This may be used to discover modifications that were not searched, but should have been (Figure 4.1(c)). Three tables relating to spectral networks are shown. One shows network size with respect to percent identified. Networks with many spectra, but a low percent identified, suggest the presence of non-peptides or peptides missing from the library and database. The other network tables help assess the accuracy of networking. The number of true positive, false positive, and false negative network pairs are displayed, along with the precision and recall for different cosine thresholds. This information can be used, for instance, to decide on an appropriate cosine threshold for finding rare modifications (Figure 4.1(d)). The groups section gives insight into how many differentially-present peptides and proteins were found in each group, was well as which groups tend

to co-occur in clusters. One table in this section shows the number of peptide sequences, variants, proteins, and unidentified clusters in each group, and can be filtered by spectrum count and group outlier ratio. Another lists the number of clusters that contain spectra from each top combination of groups (Figure 4.1(e)). A peptide section shows the similarity between peptides identified by different search algorithms. A table shows each combination of algorithms, and the number of peptide pairs for which the two peptides are the same, the peptides are the same but have different charges, one peptide is a substring of the other, one peptide has more modifications than the other, one amino acid is changed in one peptide with respect to the other, and none of the above (Figure 4.1(f)).

Identification Results

	Total	Z = 1	Z = 2	Z = 3	Z >= 4	Z = Undetermined
Original MS/MS Spectra	12254429	0	6859017	4274073	1121339	0
Filtered, Charge-Corrected MS/MS Spectra	10338029	0	5862824	3584268	890937	0
Identified MS/MS Spectra	7046088 (68.16%)	0 (--)	3925356 (66.95%)	2599169 (72.52%)	521563 (58.54%)	0 (--)
Clusters	446730	0	245110	155276	46344	0
Identified Clusters	234607 (52.52%)	0 (--)	123888 (50.54%)	88920 (57.27%)	21799 (47.04%)	0 (--)
Identified M-SPLIT Clusters	113144 (48.23%)	16 (--)	74689 (60.29%)	43330 (48.73%)	8225 (37.73%)	0 (--)
Identified MS-GF+ Clusters	73989 (31.54%)	0 (--)	34953 (28.21%)	31439 (35.36%)	7597 (34.85%)	0 (--)
Identified MODa Clusters	47563 (20.27%)	0 (--)	19976 (16.12%)	19514 (21.95%)	8073 (37.03%)	0 (--)
Clusters in Networks	325987 (72.97%)	0 (--)	184209 (75.15%)	109908 (70.78%)	31870 (68.77%)	0 (--)
Identified Clusters in Networks	169992 (52.15%)	0 (--)	93474 (50.74%)	62000 (56.41%)	14518 (45.55%)	0 (--)
Unidentified Clusters With Edges to Identified Clusters	84714 (18.96%)	0 (--)	49587 (20.23%)	27229 (17.54%)	7898 (17.04%)	0 (--)
Unidentified Clusters in Networks with Identified Clusters, With No Edges to Them	45041 (10.08%)	0 (--)	27768 (11.33%)	13054 (8.41%)	4219 (9.10%)	0 (--)
Unidentified Clusters in Networks With No Identified Clusters	26240 (5.87%)	0 (--)	13380 (5.46%)	7625 (4.91%)	5235 (11.30%)	0 (--)
Clusters Identified or in Networks	390602 (87.44%)	0 (--)	214623 (87.56%)	136828 (88.12%)	39151 (84.48%)	0 (--)
Peptide Variants	159953	2	89139	57206	13606	0
Proteins	46144	24	42396	31157	13678	0

Figure 4.1 a) Identification Results section of the Summary Report view

All Searched Mods (MSGF+)

Searched Mod	Mass Offset	Amino Acids	Options	Total Count
Deamidation	0.984016	NQ	OPTIONAL	4945
Oxidation	15.994915	M	OPTIONAL	5201
Pyroglutamate Formation	-17.026549	Q	OPTIONAL, N-TERMINAL	809
N-terminal Acetylation	42.010565	*	OPTIONAL, N-TERMINAL	4493

Figure 4.1 b) Searched Modifications section of the Summary Report view.

Top 15 Discovered Mods (MODa)

Mass Offset	# PSMs	# Variants	Top Five Sites (# PSMs with mod)
57	10394	519	P(68), H(61), V(56), M(49), G(43)
14	5156	4246	K(2126), G(398), E(366), S(307), P(295)
30	5066	4539	L(580), A(507), E(506), V(492), S(346)
28	3436	2812	K(2200), G(142), E(124), S(117), V(103)
12	3202	2668	P(556), A(210), S(207), D(198), Y(193)
53	2026	1528	P(260), E(202), A(175), G(175), L(161)
-1	1607	1504	L(237), E(188), V(184), A(155), I(133)
54	1135	921	P(154), L(134), D(97), A(89), V(89)
56	967	849	E(117), A(95), L(91), V(83), P(77)
16	834	658	P(164), Y(132), W(131), M(72), H(42)
40	764	677	C(701), K(10), V(8), P(6), E(5)
70	732	644	C(303), H(201), S(34), A(25), G(25)
-18	683	587	D(146), E(120), T(68), S(55), Q(54)
32	601	531	W(186), M(79), P(44), Y(31), C(30)
-19	583	502	K(484), R(29), D(10), L(10), T(9)

Figure 4.1 c) Discovered Modifications section of the Summary Report view.

Spectral Networks

	0% ID	1 - 24% ID	25 - 49% ID	50 - 74% ID	75 - 89% ID	90 - 99% ID	100% ID
Network size = 2	4571	188	566	1136	1036	688	8703
Network size = 3	1149	132	217	412	527	607	2596
Network size = 4	442	104	135	239	318	432	1029
Network size = 5	203	61	54	128	166	344	451
Network size = 6 - 10	255	139	169	262	375	728	547
Network size = 11 - 15	35	35	37	65	143	239	68
Network size = 16 - 30	27	31	31	90	147	188	34
Network size = 31 - 50	9	14	12	47	74	50	0
Network size = 51 - 100	111	61	147	441	600	256	0
Network size >= 101	14	43	92	228	130	19	0

Total Network Pairs	1317633
# True Positives	209953
% True Positives	70.74%
# False Positives	86837
% False Positives	29.26%
# False Negatives	174481

Cosine	Precision	Recall	# Pairs	% Pairs
0.5	70.7413	54.6135	1317633	100.00%
0.6	70.7413	54.6135	1317633	100.00%
0.7	83.7333	28.8083	645580	49.00%
0.8	93.9224	10.3875	276794	21.01%
0.9	97.6642	1.7837	112226	8.52%

Figure 4.1 d) Spectral Networks section of the Summary Report view.

Outlier Group Counts

Only the highest-count or lowest-count group for each cluster is counted. (Whichever has a greater absolute outlier ratio.)
If there is a tie in the group outlier ratio, all outlier groups are counted.

	G1	G2	G3	G4	G5	G6	Total
Unique Peptide Sequences	8802	8971	12334	13348	66747	21037	120212
Peptide Variants	10711	11389	16277	19046	88995	28828	159953
Proteins	2317	2395	2446	686	14322	1178	22976
Unidentified Clusters	5128	10108	5438	54168	105230	51647	212123

	Spectrum Count >= 0	Spectrum Count >= 4	Spectrum Count >= 7
|Group Outlier Ratio| >= 0	Apply Filter	Apply Filter	Apply Filter
|Group Outlier Ratio| >= 1	Apply Filter	Apply Filter	Apply Filter
|Group Outlier Ratio| >= 2	Apply Filter	Apply Filter	Apply Filter
|Group Outlier Ratio| >= 4	Apply Filter	Apply Filter	Apply Filter
|Group Outlier Ratio| = Inf	Apply Filter	Apply Filter	Apply Filter

Top 10 Group Combinations

Default Groups	Number of Clusters
G4 and G5 and G6	177579
G5 and G6	46446
G5	37291
G4 and G5	32262
G4 and G6	28053
G1 and G2 and G3 and G4 and G5 and G6	13199
G2 and G4 and G5 and G6	9310
G1 and G4 and G5 and G6	9202
G1 and G2 and G4 and G5 and G6	8184
G4	7817

Figure 4.1 e) Groups section of the Summary Report view.

Peptides

Number of Peptide Pairs per Category

Category descriptions:
Type 0: Peptides are the same
Type 1: Peptides are the same except charges differ
Type 2: Peptide2 (from later search) is a substring of Peptide1 (from earlier search)
Type 3: Peptide1 is a substring of Peptide2
Type 4: Peptide1 has more modifications than Peptide2
Type 5: Peptide2 has more modifications than Peptide1
Type 6: Peptides differ by one residue
Type 7: None of the above

	Type 0	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Total
MS-GF+ after M-SPLIT	15012	22995	57797	68790	7393	22421	8836	34285	237529
MODa after M-SPLIT	15521	15666	62677	179734	3906	106351	18173	7171	409199
MODa after MS-GF+	10279	10729	39966	101046	3814	40014	7226	10473	223547
Total	40812	49390	160440	349570	15113	168786	34235	51929	870275

Figure 4.1 f) Peptides section of the Summary Report view.

4.2 Identified Clusters

This view contains the combined PSM-level results from all three search algorithms (Figure 4.2). Basic information about each cluster is shown, as well as information useful for delving into the results. Various spectral counting calculations are displayed for each cluster, such as the total spectra in the cluster, the number of spectra per default group and user-defined group, and the number of spectra identified as the variant. These numbers allow for an estimation of how frequently a given peptide occurred in the sample, and which peptides might be group-specific. (The Identified Peptide Variants result view has more information on this.) A link to a visualization of the spectral network containing the cluster is shown, allowing for easy viewing of different versions of the peptide that are present in the dataset (e.g., different modifications). The PSM and Variant FDR are also listed here, providing an indication of the confidence of the match. An interactive visualization of each spectrum is shown, showing which peaks are matched by the annotation, and allowing for the annotation, ion types, mass tolerance, and other parameters to be adjusted.

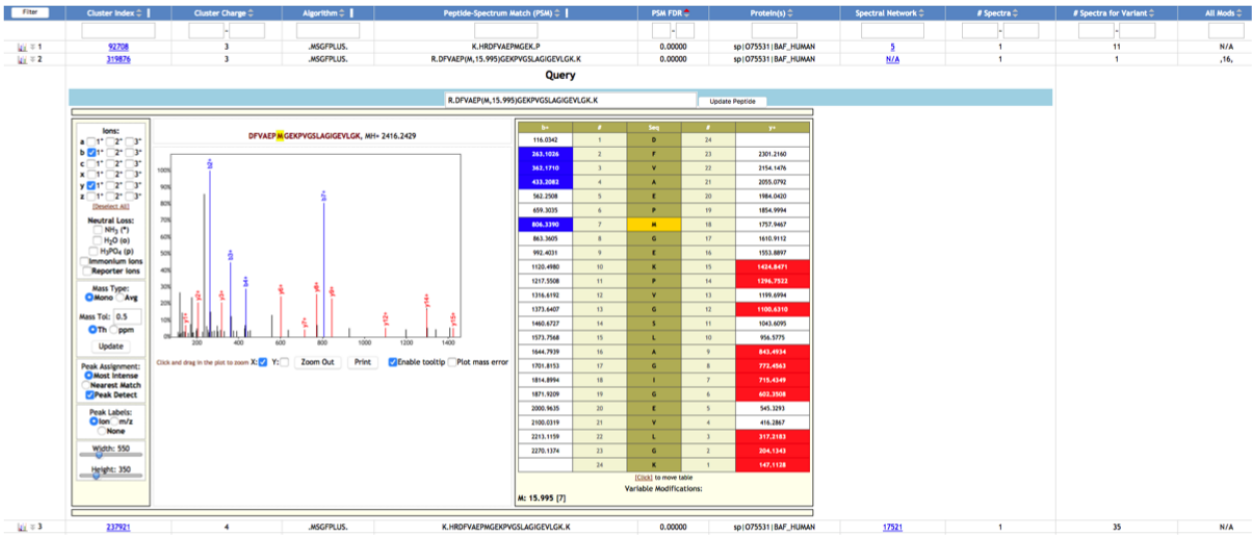
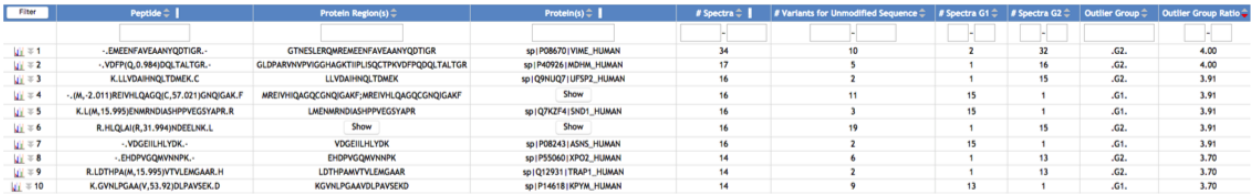


Figure 4.2 Selection of columns from the Identified Clusters view.

4.3 Identified Peptide Variants

Peptides with the same sequence and a similar total modification mass are grouped together as a peptide variant. This view contains each peptide variant in the dataset, and is useful for looking at the set of variants for a given peptide sequence, or looking at the set of peptides that contain a given modification mass. For each variant, the number of spectra assigned to the variant is shown, and this number is also broken down by user-defined group and default group (Figure 4.3). The outlier group and outlier group ratio are shown, providing an estimation of how group-specific the variant is. Sorting or filtering by the outlier group ratio is useful for viewing the variants that are disproportionately present or absent in one group with respect to the other groups. The number of variants corresponding to the peptide sequence is also displayed, allowing an estimation of how diverse the peptide is. The view contains information about the protein region (i.e., the region on the protein where the peptide occurs).

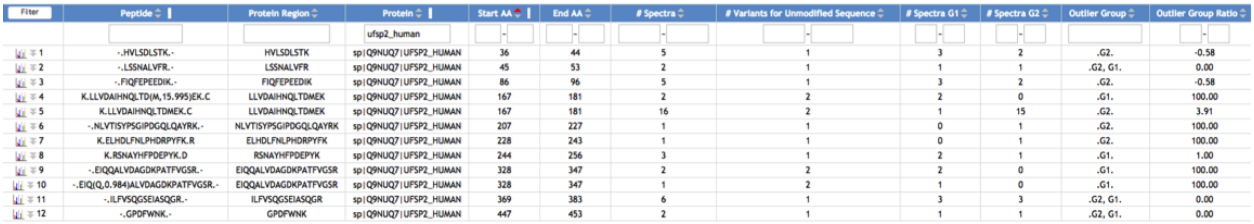


	Peptide	Protein Region(s)	Protein(s)	# Spectra	# Variants for Unmodified Sequence	# Spectra G1	# Spectra G2	Outlier Group	Outlier Group Ratio
1	-.EMEENFAVEAANYQDTIGR.-	GTNESLERQMREMEENFAVEAANYQDTIGR	sp|P08670|VIME_HUMAN	34	10	2	32	.G2.	4.00
2	-.VDFP(Q,0.984)DQLTALTGR.-	GLDPARVNVPVIGGHAGKTIPLISQCTPKVDFPQDQLTALTGR	sp|P40926|MDHM_HUMAN	17	5	1	16	.G2.	4.00
3	K.LLVDAIHNQLTDMEK.C	LLVDAIHNQLTDMEK	sp|Q9NUQ7|UFSP2_HUMAN	16	2	1	15	.G2.	3.91
4	-.(M,-2.011)REIVHLQAGQ(C,57.021)GNQIGAK.F	MREIVHIQAGQCGNQIGAKF;MREIVHLQAGQCGNQIGAKF	Show	16	11	15	1	.G1.	3.91
5	K.L(M,15.995)ENMRNDIASHPPVEGSYAPR.R	LMENMRNDIASHPPVEGSYAPR	sp|Q7KZF4|SND1_HUMAN	16	3	15	1	.G1.	3.91
6	R.HLQLAI(R,31.994)NDEELNK.L	Show	Show	16	19	1	15	.G2.	3.91
7	-.VDGEIILHLYDK.-	VDGEIILHLYDK	sp|P08243|ASNS_HUMAN	16	2	15	1	.G1.	3.91
8	-.EHDPVGQMVNNPK.-	EHDPVGQMVNNPK	sp|P55060|XPO2_HUMAN	14	6	1	13	.G2.	3.70
9	R.LDTHPA(M,15.995)VTVLEMGAAR.H	LDTHPAMVTVLEMGAAR	sp|Q12931|TRAP1_HUMAN	14	2	1	13	.G2.	3.70
10	K.GVNLPGAA(V,53.92)DLPAVSEK.D	KGVNLPGAAVDLPAVSEKD	sp|P14618|KPYM_HUMAN	14	9	13	1	.G1.	3.70

Figure 4.3 Selection of columns from the Identified Variants view, showing highly-differential variants.

4.4 Identified Peptide Variants by Protein Region

This view can be used to look at peptides covering a region of interest in a protein (this region may be a functional site, etc.), or peptides covering a certain amino acid sequence. This is the same as the Identified Peptide Variants view, except that instead of each protein region for a given variant being merged into one row, each protein region is listed in a separate row. This view also displays information that the Identified Peptide Variants view does not – specifically, the start and end amino acid on the protein (Figure 4.4).



	Peptide	Protein Region	Protein	Start AA	End AA	# Spectra	# Variants for Unmodified Sequence	# Spectra G1	# Spectra G2	Outlier Group	Outlier Group Ratio
			ufsp2_human								
1	-.HVLSDLSTK.-	HVLSDLSTK	sp|Q9NUQ7|UFSP2_HUMAN	36	44	5	1	3	2	.G2.	-0.58
2	-.LSSNALVFR.-	LSSNALVFR	sp|Q9NUQ7|UFSP2_HUMAN	45	53	2	1	1	1	.G2, G1.	0.00
3	-.FIQFEPEEDIK.-	FIQFEPEEDIK	sp|Q9NUQ7|UFSP2_HUMAN	86	96	5	1	3	2	.G2.	-0.58
4	K.LLVDAIHNQLTD(M,15.995)EK.C	LLVDAIHNQLTDMEK	sp|Q9NUQ7|UFSP2_HUMAN	167	181	2	2	2	0	.G1.	100.00
5	K.LLVDAIHNQLTDMEK.C	LLVDAIHNQLTDMEK	sp|Q9NUQ7|UFSP2_HUMAN	167	181	16	2	1	15	.G2.	3.91
6	-.NLVTISYPSGIPDGQLQAYRK.-	NLVTISYPSGIPDGQLQAYRK	sp|Q9NUQ7|UFSP2_HUMAN	207	227	1	1	0	1	.G2.	100.00
7	K.ELHDLFNLPHDRPYFK.R	ELHDLFNLPHDRPYFK	sp|Q9NUQ7|UFSP2_HUMAN	228	243	1	1	0	1	.G2.	100.00
8	K.RSNAYHFPDEPYK.D	RSNAYHFPDEPYK	sp|Q9NUQ7|UFSP2_HUMAN	244	256	3	1	2	1	.G1.	1.00
9	-.EIQQALVDAGDKPATFVGSR.-	EIQQALVDAGDKPATFVGSR	sp|Q9NUQ7|UFSP2_HUMAN	328	347	2	2	2	0	.G1.	100.00
10	-.EIQ(Q,0.984)ALVDAGDKPATFVGSR.-	EIQQALVDAGDKPATFVGSR	sp|Q9NUQ7|UFSP2_HUMAN	328	347	1	2	1	0	.G1.	100.00
11	-.ILFVSQGSEIASQGR.-	ILFVSQGSEIASQGR	sp|Q9NUQ7|UFSP2_HUMAN	369	383	6	1	3	3	.G2, G1.	0.00
12	-.GPDFWNK.-	GPDFWNK	sp|Q9NUQ7|UFSP2_HUMAN	447	453	2	1	1	1	.G2, G1.	0.00

Figure 4.4 Selection of columns from the Identified Variants per Protein Region view, showing variants and corresponding regions for the protein sp|Q9NU97|UFSP2_HUMAN.

4.5 Identified Proteins

This view lists each protein that contains identified variants, and is useful for directly looking at biologically-relevant information. For each protein, the number of spectra that map to it is listed, and broken down by group (Figure 4.5). This can be used to look at which proteins are commonly seen in a certain group (by filtering by the number of spectra present in that group). Since some peptide sequences may map to multiple proteins, and therefore may not be indicative of the given protein, the number of spectra

that uniquely map to the protein (i.e., do not map to any other proteins) is also shown.

Filter	Protein	# Spectra G1	# Unique Spectra G1	# Spectra G2	# Unique Spectra G2	Outlier Group	Outlier Group Ratio	Unique Outlier Group	Unique Outlier Group Ratio
1	sp|P34059|GALNS_HUMAN	9	9	1	1	.G1.	3.17	.G1.	3.17
2	sp|Q6N5J5|LRC8E_HUMAN	11	9	1	1	.G1.	3.46	.G1.	3.17
3	sp|Q9BV19|CA050_HUMAN	9	9	1	1	.G1.	3.17	.G1.	3.17
4	sp|Q9BXP2|S12A9_HUMAN	9	9	1	1	.G1.	3.17	.G1.	3.17
5	sp|Q9HA47|UCK1_HUMAN	12	9	2	1	.G1.	2.58	.G1.	3.17
6	sp|Q98R77|CCD77_HUMAN	8	8	1	1	.G1.	3.00	.G1.	3.00
7	sp|Q9H147|TDIF1_HUMAN	8	8	1	1	.G1.	3.00	.G1.	3.00
8	sp|Q9H497|TOR3A_HUMAN	23	23	3	3	.G1.	2.94	.G1.	2.94
9	sp|O15484|CAN5_HUMAN	6	6	1	1	.G1.	2.58	.G1.	2.58
10	sp|O94811|TPPP_HUMAN	6	6	1	1	.G1.	2.58	.G1.	2.58

Figure 4.5 Selected columns from the Identified Proteins view, showing proteins that have a high ratio of spectra in group G1 to group G2, considering spectra that uniquely map to each protein.

4.6 Identified Protein Regions

This view can be used to get an overview of the protein regions present in the dataset. It lists each protein region found, and the number of variants that map to it (Figure 4.6). Sorting by the number of variants reveals which regions have the most coverage/ diversity.

Filter	Protein Region Index	Protein Region	Protein	Variant Count	Start AA	End AA
1	14387	DGVTVAKSIDLKDKYKNIGAKLVQDVANNTNEEAGDGTTTATVLAR	sp|P10809|CH60_HUMAN	76	75	120
2	34667	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|P58876|H2B1D_HUMAN	71	30	72
3	34221	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|Q99879|H2B1M_HUMAN	71	30	72
4	34217	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|O60814|H2B1K_HUMAN	71	30	72
5	34211	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|P62807|H2B1C_HUMAN	71	30	72
6	34050	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|P57053|H2BFS_HUMAN	71	30	72
7	34040	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|Q99877|H2B1N_HUMAN	71	30	72
8	33789	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|Q5QNW6|H2B2F_HUMAN	71	30	72
9	33776	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|Q99880|H2B1L_HUMAN	71	30	72
10	33770	KRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFER	sp|Q93079|H2B1H_HUMAN	71	30	72

Figure 4.6 Protein regions that have many corresponding peptide variants in the Identified Protein Regions view.

4.7 Identified Spectra

All of the unclustered spectra that were are identified, are displayed here. This view may be useful for looking at a cluster in depth (via its component spectra), or for looking at spectrum-specific information such as original filename (which can be used to find experimental details), precursor intensity, and retention time (Figure 4.7). Both the input spectrum and the cluster consensus spectrum are shown in the spectrum

viewer. This view also contains identifications, and loads more quickly than the All Spectra view for large datasets.

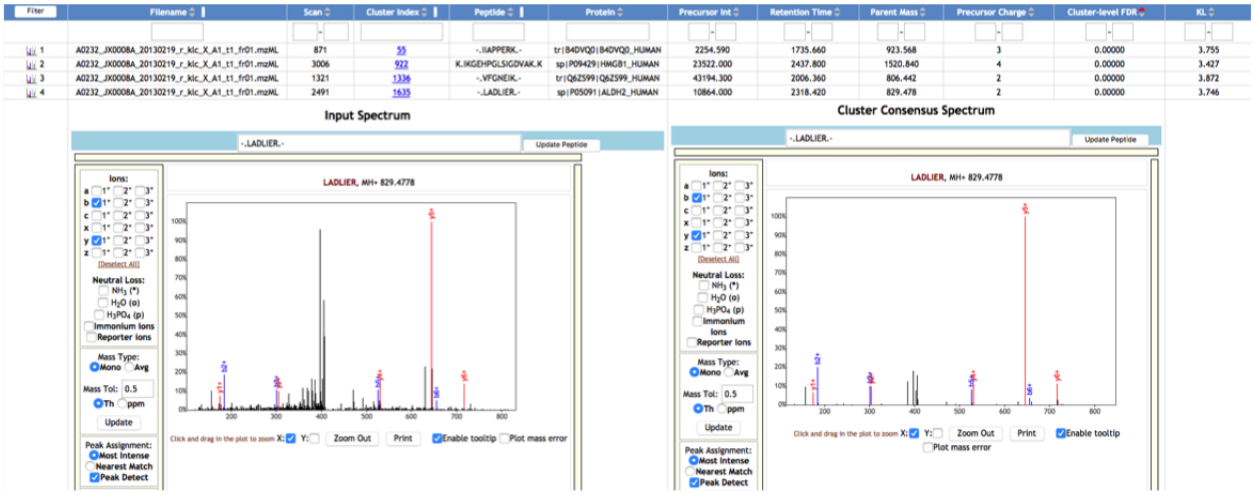


Figure 4.7 Identified Spectra view.

4.8 All Clusters

All clusters, whether identified or not, are listed in this view (Figure 4.8). This can be used to look at clusters that did not get identified. This view also contains per-group spectral counting information. The view can be filtered by mass or charge to narrow down results, or sorted by the number of spectra to see the smallest/ largest clusters.

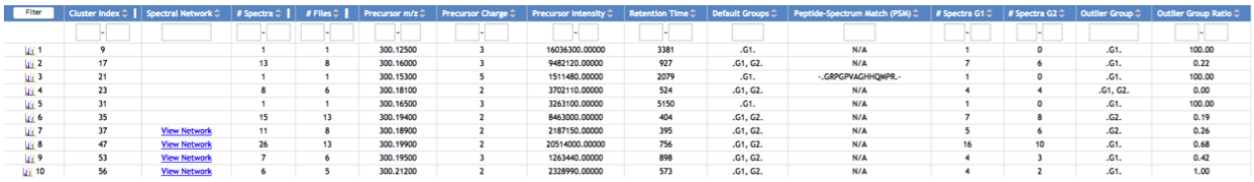


Figure 4.8 All Clusters view.

4.9 All Spectra

All spectra, whether identified or not, are listed in this view (Figure 4.9). This can be used to look at spectra that did not get identified, and view these spectra in the spectrum viewer along with possible an-

notations. Values such as precursor intensity, retention time, parent mass, and Kullback-Leibler divergence are shown. The unannotated input spectrum and cluster consensus spectrum are displayed in the spectrum viewer.

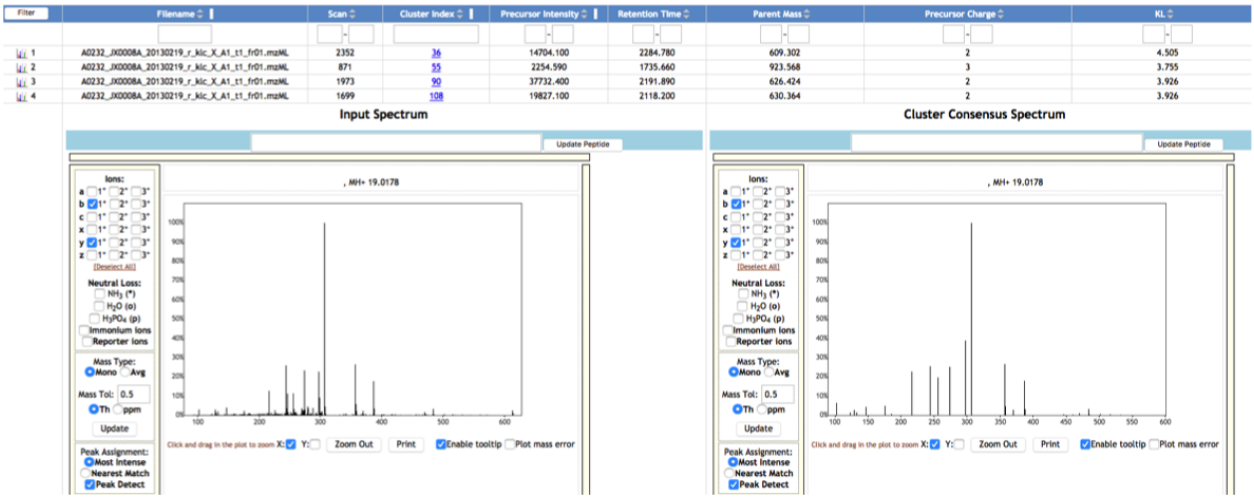


Figure 4.9 All Spectra view.

4.10 MzTab Result Files

It may be useful to have results in a standardized format, if further processing is required. This view contains the results converted into the community standard mzTab format (Figure 4.10). These results can be directly used for MassIVE dataset submission, running the Results Comparison workflow, etc.

Filter	PSM ID	Peak List File	Scan	Peptide	Protein	Charge	Valid	cluster_index	AllFiles	sum(precursor_intensity)	RTMean	RTStdErr	parent_mass
1	1	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	3428	FARPDFR	tr|B4DWQ6|B4DWQ6_HUMAN	3	VALID	24	inputspectra/spec-00000.mgf	40141	1960	0	908
2	2	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	3494	FARPDFR	tr|B4DWQ6|B4DWQ6_HUMAN	3	VALID	24	inputspectra/spec-00000.mgf	221852	1975	0	908
3	3	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	2478	FGIAAK	tr|B4DKM5|B4DKM5_HUMAN	2	VALID	26	inputspectra/spec-00000.mgf	57489	1739	0	606
4	4	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	2543	FGIAAK	tr|B4DKM5|B4DKM5_HUMAN	2	VALID	26	inputspectra/spec-00000.mgf	993153	1754	0	606
5	5	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	2615	FGIAAK	tr|B4DKM5|B4DKM5_HUMAN	2	VALID	26	inputspectra/spec-00000.mgf	304328	1771	0	606
6	6	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	1956	FVHGEGLR	sp|P30536|TSPOA_HUMAN	3	VALID	38	inputspectra/spec-00000.mgf	14033	1612	0	915
7	7	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	2024	FVHGEGLR	sp|P30536|TSPOA_HUMAN	3	VALID	38	inputspectra/spec-00000.mgf	93058	1628	0	914
8	8	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	2096	FVHGEGLR	sp|P30536|TSPOA_HUMAN	3	VALID	38	inputspectra/spec-00000.mgf	290774	1645	0	914
9	9	nuno/ccms-ftp01/Cancer/CPTAC2/TCGA_Colorectal_Cancer/TCGA-A6-3807-01A-22_W_VU_20121019_A0218_4I_R_FR01.mzXML	2170	FVHGEGLR	sp|P30536|TSPOA_HUMAN	3	VALID	38	inputspectra/spec-00000.mgf	11011	1663	0	914
10	10	jswertz/cfdb72dbf2dc4d08ad394890243c62e6/spec/specthree-00000.mgf	7053	LVLLR	sp|P35557|HXK4_HUMAN	2	VALID	43	inputspectra/specthree-00000.mgf	773291	2396	0	613

Figure 4.10 Selected columns from the PSM results within the MzTab Result Files view.

4.11 M-SPLIT Clusters

Since the overall views cannot include algorithm-specific columns, per-algorithm result views are displayed. Each algorithm contains scores that are specific to it. In the M-SPLIT library search result views, the query spectrum is displayed next to its matching library spectrum (Figure 4.11). The M-SPLIT result views also show mixtures, which are separated into components in the overall views.

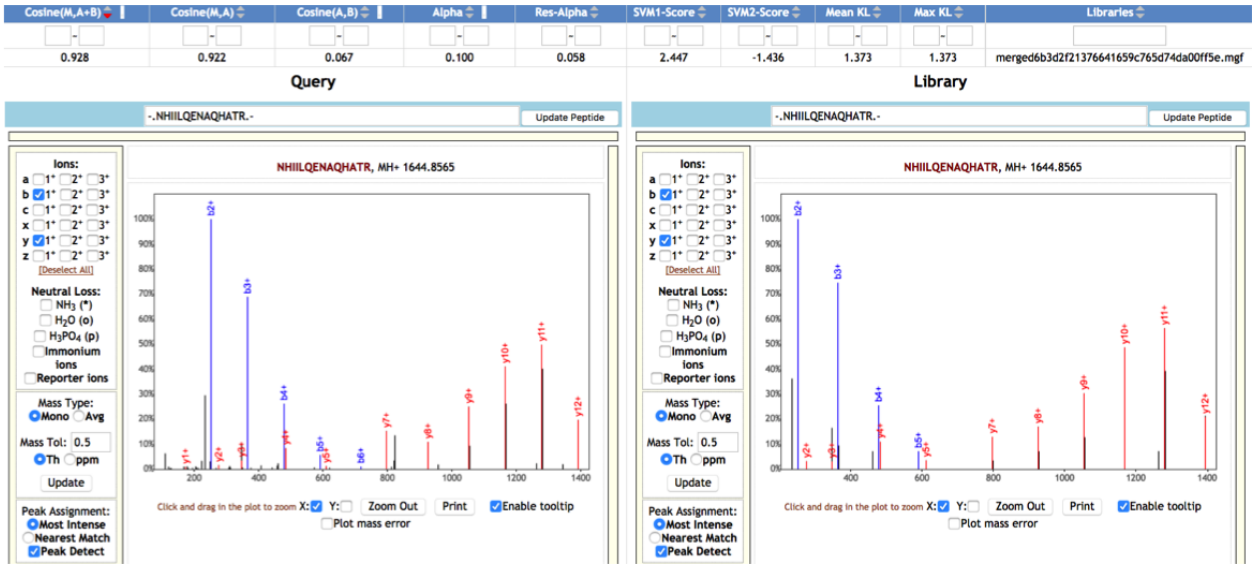


Figure 4.11 Selected columns from the M-SPLIT Clusters view, sorted by decreasing cosine score.

4.12 M-SPLIT Peptides

This view has the same columns as the M-SPLIT Clusters view, but groups M-SPLIT identifications by unmodified peptide instead of showing each cluster individually.

4.13 M-SPLIT Proteins

This view has the same columns as the M-SPLIT Clusters view, but groups M-SPLIT identifications by protein instead of showing each cluster individually.

4.14 MS-GF+ Clusters

This view shows clusters identified by MS-GF+ database search, and contains spectral probabilities and other columns specific to MS-GF+ scoring (Figure 4.12).

Figure 4.12 Selected columns from the MS-GF+ Clusters view, sorted by decreasing spectral probability.

4.15 MS-GF+ Peptides

This view has the same columns as the MS-GF+ Clusters view, but groups MS-GF+ identifications by unmodified peptide instead of showing each cluster individually.

4.16 MS-GF+ Proteins

This view has the same columns as the MS-GF+ Clusters view, but groups MS-GF+ identifications by protein instead of showing each cluster individually.

4.17 MODa Clusters

MODa views can be used to look at peptides with unexpected modifications. This view shows clusters identified by MODa blind search, and contains columns specific to MODa scoring (Figure 4.13).

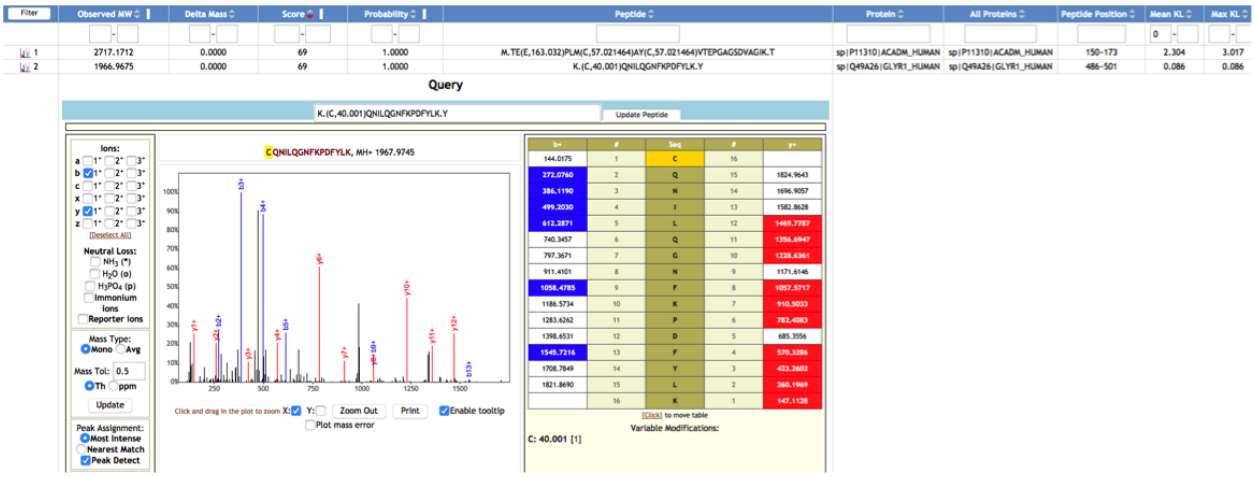


Figure 4.13 Selected columns from the MODa Clusters view.

4.18 MODa Peptides

This view has the same columns as the MODa Clusters view, but groups MODa identifications by unmodified peptide instead of showing each cluster individually.

4.19 MODa Proteins

This view has the same columns as the MODa Clusters view, but groups MODa identifications by protein instead of showing each cluster individually.

4.20 Identified Network Pairs

This view compares identified clusters based on spectral alignment. Each pair of identified clusters in the network is shown, as well as theoretical pairs that were not discovered by spectral networking (Figure 4.14). This may be useful for finding clusters that have similar identifications, that were not paired in the network. The cosine and mass difference between each pair is listed, as well as whether the pair is in the network, and whether the pairing is correct (based on prefix residue masses).

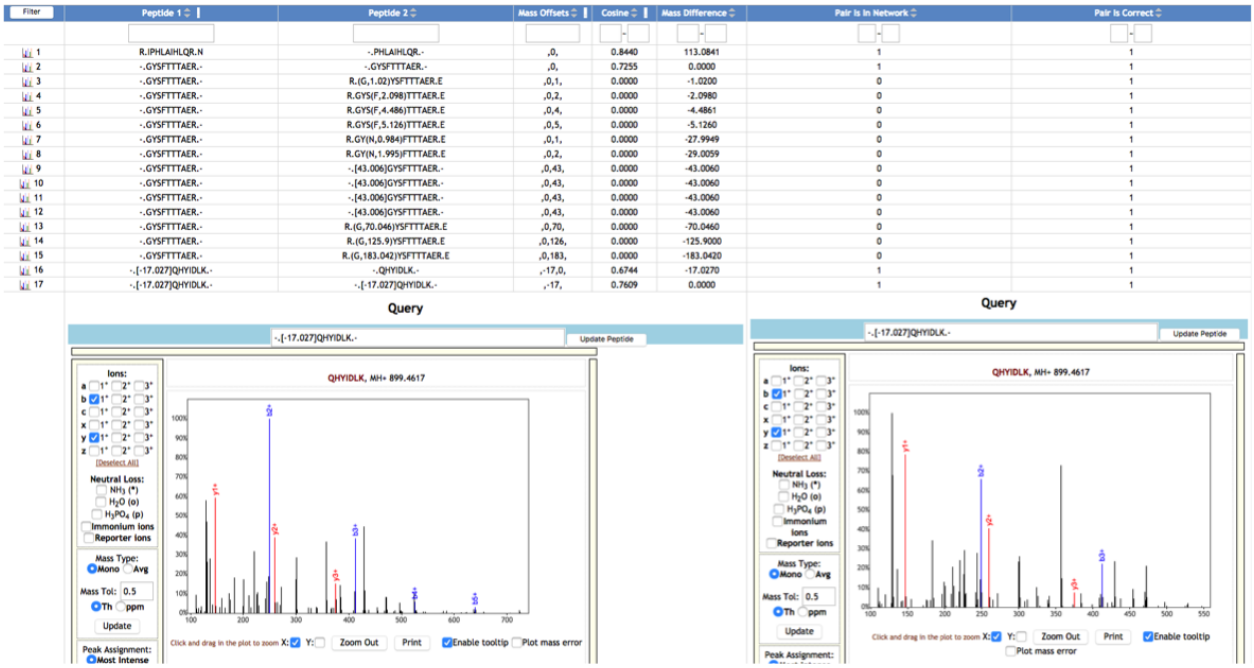


Figure 4.14 Selected columns from the Identified Network Pairs view.

4.21 Network Pair Modifications

This view offers a look at the frequency of parent mass differences based on spectral alignment of network pairs. For each mass offset in the network, this view shows how many network pairs contain a PSM with that offset. This number is broken down by cosine bin (Figure 4.15).

Filter	Mass Offset	Cos < .45	.45 <= Cos < .55	.55 <= Cos < .65	.65 <= Cos < .75	.75 <= Cos < .85	.85 <= Cos < .95	Cos >= .95
1	0	0	0	22700	27696	10284	2145	55
2	57	0	0	5805	6932	2508	504	5
3	1	0	0	2622	3371	1326	250	6
4	16	0	0	3383	4119	1394	174	0
5	43	0	0	2707	3368	981	154	1
6	-17	0	0	1154	1442	501	130	2
7	42	0	0	762	1061	467	123	1
8	-1	0	0	613	800	416	101	8
9	28	0	0	797	1011	342	45	0
10	-18	0	0	285	400	144	40	0

Figure 4.15 Top 10 most common mass offsets with a cosine between 0.85 and 0.95 via the Network Pair Modifications view.

4.22 All Network Pairs

This view shows all of the cluster pairs detected by spectral alignment (spectral network edges), whether identified or not. The cosine and mass difference between each pair is listed in the view (Figure 4.16). This can be used to look at unidentified pairs in the spectrum viewer, and check possible annotations. It may also be useful to to obtain all of the pairs with a certain mass difference, by applying mass difference filters.

Figure 4.16 All Network Pairs view.

4.23 Peptide Pairs

This view can be used to look at peptide agreement between search algorithms. Each pair of peptides is categorized based on whether the peptides are the same, the peptides have a charge difference between them, one peptide is a substring of the other, one peptide has more or fewer modifications than the other, or one residue is changed in one peptide with respect to the other (Figure 4.17).

Figure 4.17 Selected columns from the Peptide Pairs view, showing peptide pairs for which a residue is changed in a MODa peptide with respect to an MS-GF+ peptide.

4.24 Peptide Modifications

This view can be used to look at the frequency of modifications across all variants. A table is shown listing the number of occurrences of each mass offset on each amino acid (Figure 4.18).

Mass Offset	AC	CC	DC	EC	FC	GC	HC	IC	KC	LC	MC	NC	PC	QC	RC	SC	TC	VC	WC	YC	Sum
57	38	29885	36	17	15	45	18	45	16	51	59	26	61	16	11	36	24	61	0	25	30364
16	25	0	11	24	20	31	4	27	9	23	9635	24	56	15	5	28	13	22	297	33	10302
43	30	4	21	35	5	26	7	21	115	33	10	5	23	18	20	14	18	22	2	12	8688
1	132	0	41	116	96	53	23	128	34	224	16	2861	113	2391	7	99	88	180	7	61	6670
183	3	1	6	4	0	6	0	4	1540	3	0	1	11	0	0	9	5	6	1	1969	3569
-17	0	2	12	32	5	0	2	24	3	26	1	126	14	182	21	10	16	19	2	4	3528
-1	287	0	99	263	130	0	59	292	96	450	65	112	242	234	18	234	194	397	9	123	3304
42	17	19	4	13	2	11	6	10	26	17	9	6	4	10	3	15	8	13	0	2	3102
28	56	1	66	62	21	110	61	41	58	53	5	64	84	40	35	597	200	53	5	19	1631
80	5	3	0	3	1	6	1	0	0	2	1	1	15	1	2	1074	187	3	0	38	1343
54	103	13	104	125	47	60	5	81	2	112	15	74	114	89	0	94	72	100	8	32	1250
302	83	0	11	61	60	19	14	170	14	240	22	15	88	77	9	48	41	193	14	40	1219
32	13	1	9	10	6	22	3	8	1	18	98	12	29	12	5	21	11	16	671	9	975
2	56	0	31	57	30	33	6	61	20	110	16	36	51	49	2	55	40	87	1	31	772
-2	52	5	16	43	26	0	9	48	11	70	202	29	47	35	7	36	21	70	15	30	772
40	4	600	2	5	2	4	1	1	0	3	4	5	0	8	0	3	1	3	0	1	647
-18	0	0	87	224	13	0	2	12	10	22	5	11	13	75	1	27	71	13	0	5	591
3	13	0	14	30	20	14	2	40	6	51	6	14	17	25	4	12	18	58	2	18	364
126	2	0	7	2	2	11	0	1	0	5	0	4	4	2	0	7	5	3	23	277	355
-3	24	0	9	21	17	0	3	29	6	47	15	18	27	19	4	28	18	39	5	17	346
17	14	2	8	17	3	30	3	9	0	19	102	9	40	14	2	17	11	17	19	7	343
15	10	0	5	12	4	10	1	15	0	14	128	6	18	5	2	8	7	11	72	9	337
4	11	1	8	18	7	16	0	18	2	31	6	10	14	12	6	10	13	32	88	9	312
44	13	4	8	9	5	19	2	7	9	14	17	2	1	6	1	9	5	9	89	1	230
14	4	6	9	11	10	26	10	3	8	6	29	9	19	4	9	23	3	21	8	2	220
22	10	1	10	11	2	22	0	5	0	10	1	16	31	19	0	20	23	20	1	2	204
56	4	116	2	2	6	9	0	4	3	1	9	2	9	2	1	5	3	10	0	4	192
58	10	105	0	1	3	8	1	3	0	2	23	6	2	0	2	1	6	7	0	2	182
250	8	0	1	11	11	6	4	23	5	23	2	4	13	6	1	4	7	33	0	5	167
48	1	108	0	0	0	1	0	1	1	1	10	2	3	2	0	1	2	2	16	4	155

Figure 4.18 Selected columns from the Peptide Modifications view, sorted by the most common modifications.

4.25 Spectral Networks

Spectral networks can be used to extrapolate identifications of correctly-identified spectra in a network component to incorrectly-identified or unidentified spectra in the component, and assess why those spectra are missed by identification algorithms. They are also useful for determining the different versions of a peptide that are present in the dataset. This view shows information about each spectral network component, as well as an interactive visualization of each component (Figure 4.19). Information shown about each component includes a listing of the clusters in the network, a listing of the peptides in the network, the number of spectra in the network (overall and per group), and the percentage of clusters in the network that were identified.

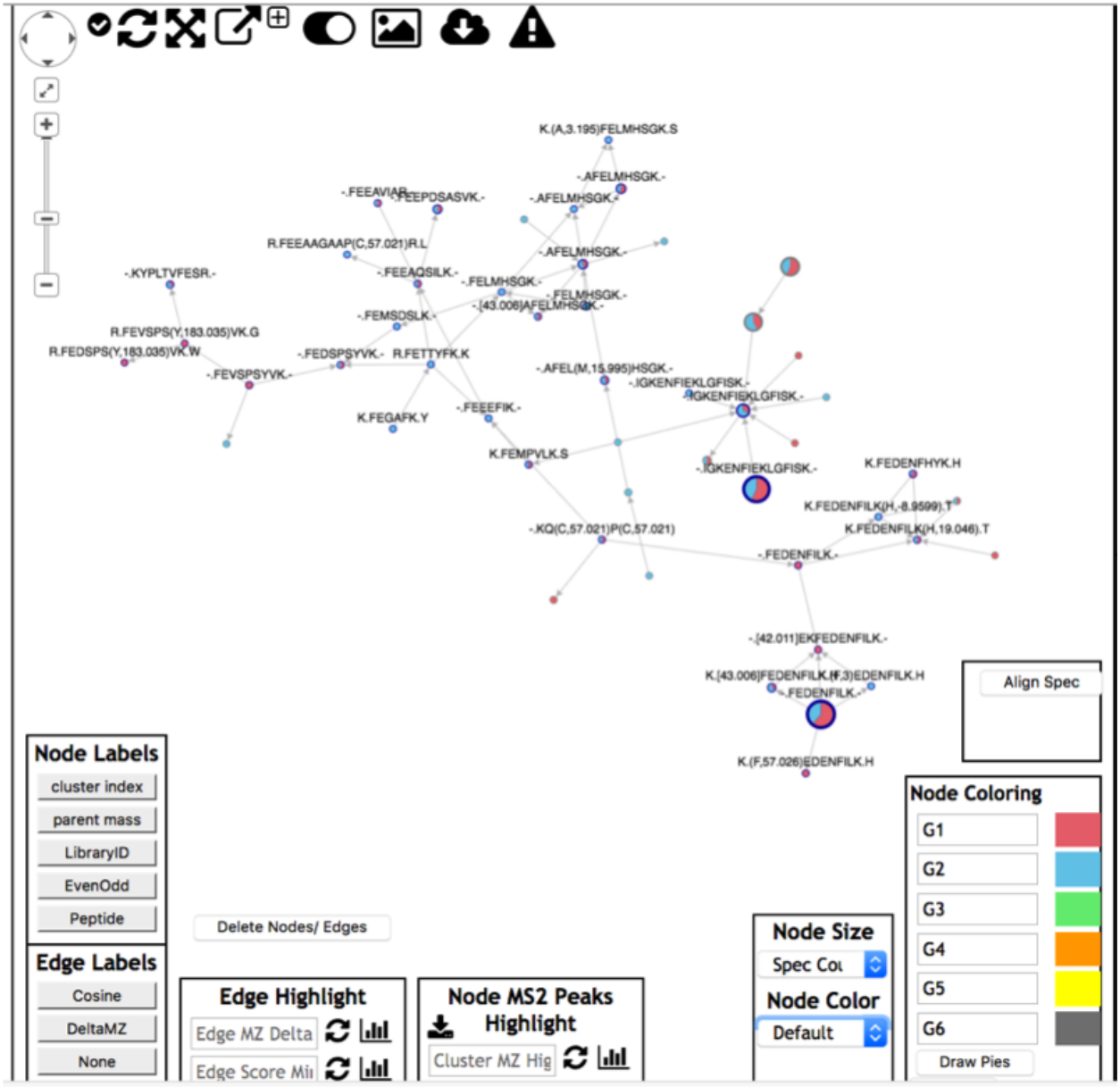


Figure 4.19 Visualization of a spectral network in the Spectral Networks view. Nodes are sized by spectral count and colored by group ratio.

4.26 Peptide Variant Networks

In order to look at the versions of a peptide present, variant networks may be more useful than spectral networks. Variant networks show all of the variants in each protein region, even if some of the

variants are in separate spectral network components. Each variant is also shown in a single node in this view, whereas spectral networks often contain the same peptide in multiple clusters, cluttering the network visualization. This view shows information about each variant network component, as well as an interactive visualization of each component (Figure 4.20). Information shown about each component includes a listing of the variants in the network, a listing of the clusters in the network, and the number of variants in the network.

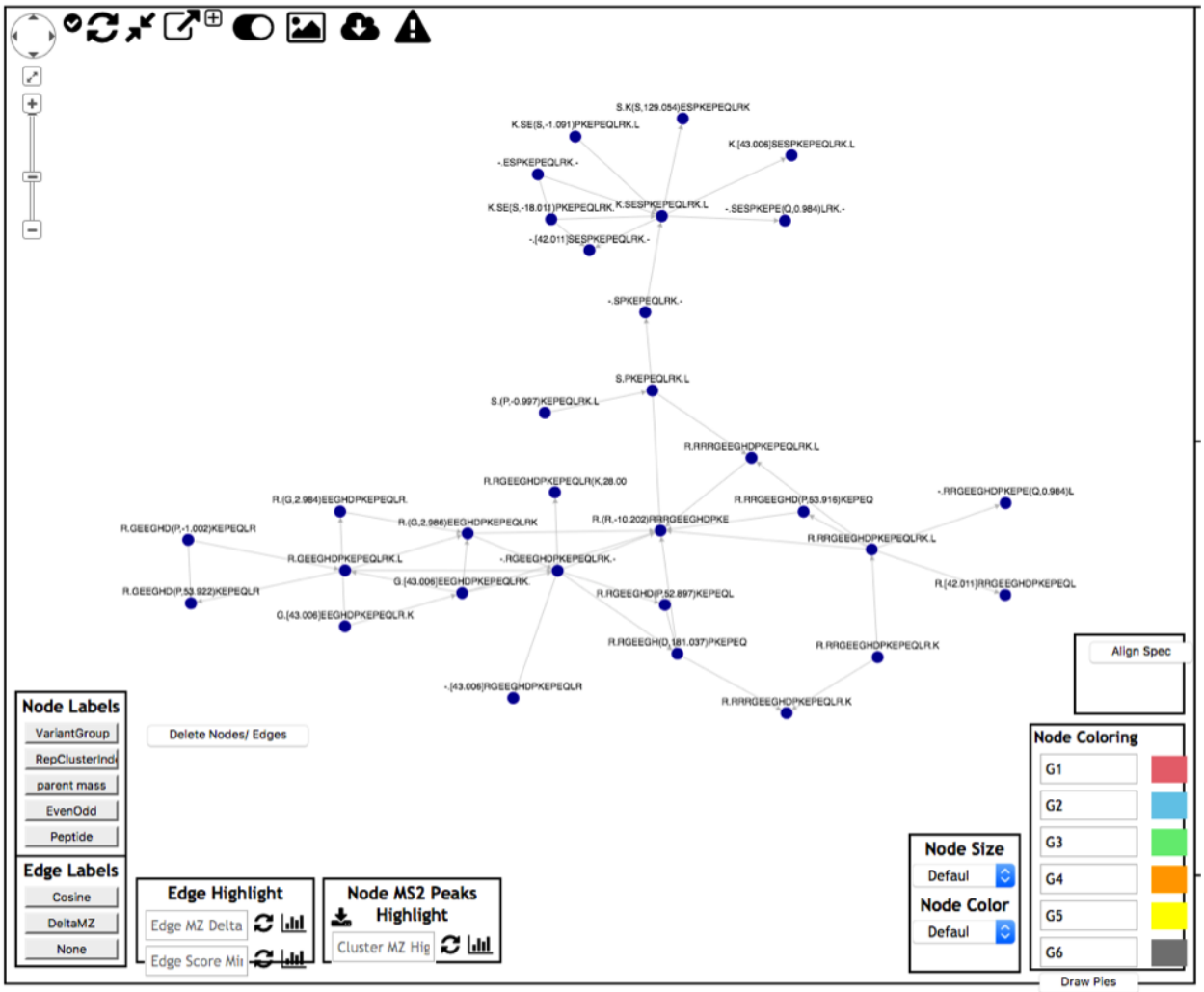


Figure 4.20 Visualization of a variant network in the Variant Networks view.

Chapter 5

Results

5.1 Maestro identifies more PSMs than individual search algorithms

A modified version of Maestro was created that runs M-SPLIT, MS-GF+, and MODa individually on all input spectra, and does not remove identified spectra or run MODa on a reduced protein database. This workflow was run on a proteome-wide HEK 293 dataset [3] in a semi-tryptic search with an ion tolerance of 0.01Da; searched modifications of oxidation, pyroglutamate formation, phosphorylation, N-terminal carbamylation, and N-terminal acetylation; and parent mass tolerances of 2.5Da, 5ppm, and 0.05Da for M-SPLIT, MS-GF+, and MODa, respectively. An FDR of 1 percent was applied. M-SPLIT identified 160603 clusters, MS-GF+ identified 137322 clusters, and MODa identified 194385 clusters. Maestro run with the same parameters and filtered at 1 percent FDR identified 230527 clusters – significantly more than any individual search algorithm.

5.2 Maestro discovers a wide variety of peptides and modifications

To evaluate discovery of peptide variants, Maestro was run on the above HEK 293 dataset and filtered at 1 percent FDR. Over 1100 different modifications (mass offsets on a particular amino acid) were initially found. These modifications were searched in the Unimod protein modification database [6], and evaluated for plausibility based on presence in spectral networks, cosine to network neighbors identified as

the unmodified version of the peptide, the presence of network neighbors with a different peptide sequence, the number of spectrum peaks surrounding the putative modification, and other factors. 371 modifications were found to have strong evidence supporting their presence (Figure 5.1).

Modification Name	Unimod Classification	Cosine	Modification Name	Unimod Classification	Cosine	Modification Name	Unimod Classification	Cosine
Acetyl(Nterm)	Multiple	0.946	W+15	Novel	0.854	Labile+477	Novel	0.824
Glu->pyro-Glu(E)	Artefact	0.933	Oxidation(N)	Post-translational	0.853	Labile+316	Novel	0.824
Dehydrated(D)	Chemical derivative	0.93	Oxidation(Y)	Post-translational	0.852	Labile+419	Novel	0.824
Nterm+183	Novel	0.925	Labile+92	Novel	0.852	Hex(1)HexNAc(1)(T)	O-linked glycosylation	0.823
Deamidated(Q)	Artefact	0.92	Labile+301	Novel	0.851	Dehydrated(T)	Post-translational	0.822
Deamidated(N)	Artefact	0.918	M->E/2(M)	AA substitution	0.847	Labile+330	Novel	0.822
Labile+250	Novel	0.917	Carbamyl(Nterm)	Multiple	0.847	Nterm-20	Novel	0.821
Labile+181	Novel	0.912	L->D/2(L)	AA substitution	0.847	Labile+300	Novel	0.821
L->N/2(L)	AA substitution	0.907	Labile+93	Novel	0.847	Labile+318	Novel	0.821
Formyl(S)	Artefact	0.906	Propionyl(Nterm)	Multiple	0.845	Labile+285	Novel	0.819
Carboxymethyl(C)	Chemical derivative	0.906	Dioxidation(W)	Chemical derivative	0.844	Formyl(Nterm)	Artefact	0.818
Formyl(T)	Artefact	0.902	Carbamidomethyl(Nterm)	Artefact	0.844	Labile+222	Novel	0.817
Methyl(R)	Post-translational	0.896	Labile+324	Novel	0.844	AEBS(K)	Artefact	0.816
Oxidation(W)	Artefact	0.895	Labile+426	Novel	0.841	Labile+377	Novel	0.815
Labile+408	Novel	0.894	C->S/1(C)	AA substitution	0.84	Deoxy(S)	Chemical derivative	0.814
Ammonium(E)	Artefact	0.882	Ammonia-loss(N)	Chemical derivative	0.84	Delta:H(4)C(3)(Nterm)	Other	0.813
AEBS(Y)	Artefact	0.882	Y->E/2(Y)	AA substitution	0.838	Labile+185	Novel	0.813
F->E/3(F)	AA substitution	0.882	T->P/1(T)	AA substitution	0.838	E-20	Novel	0.813
Q-18	Novel	0.879	Carboxy(W)	Chemical derivative	0.837	Labile+50	Novel	0.813
I->V/1(I)	AA substitution	0.878	Cation:Fe[II](E)	Artefact	0.837	Labile+254	Novel	0.811
Labile+474	Novel	0.877	Labile+271	Novel	0.837	Dimethyl(R)	Post-translational	0.81
HexNAc(2)(T)	O-linked glycosylation	0.876	Labile+230	Novel	0.836	Labile+308	Novel	0.81
Labile+302	Novel	0.871	W+126	Novel	0.835	Labile+454	Novel	0.81
Phospho(S)	Post-translational	0.869	Thiazolidine(W)	Chemical derivative	0.834	V->T/2(V)	AA substitution	0.807
Iodo(Y)	Chemical derivative	0.867	Labile+171	Novel	0.834	Pyro-carbamidomethyl(C)	Artefact	0.807
Dicarbamidomethyl(Nterm)	Artefact	0.867	Labile+282	Novel	0.834	Labile+179	Novel	0.806
Labile+256	Novel	0.866	Labile+14	Novel	0.833	ct/->K(Z)	addedAA	0.805
Labile+91	Novel	0.865	K-52	Novel	0.833	Methyl(C)	Post-translational	0.804
Nterm+229	Novel	0.864	Carbamidomethyl(S)	Artefact	0.832	L+302	Novel	0.804
Labile+276	Novel	0.864	A->S/1(A)	AA substitution	0.83	P+129	Novel	0.804
Trp->Kynurenin(W)	Chemical derivative	0.862	P->N/2(P)	AA substitution	0.83	Carbamidomethyl(D)	Artefact	0.803
Labile+319	Novel	0.862	Cation:K(E)	Artefact	0.83	V->IL/1(V)	AA substitution	0.802
Labile+210	Novel	0.86	M->N/2(M)	AA substitution	0.83	Labile+348	Novel	0.802
Oxidation(M)	Artefact	0.859	Labile+283	Novel	0.828	Dehydrated(S)	Post-translational	0.801
Labile+41	Novel	0.859	Labile+306	Novel	0.827	Q-17	Novel	0.8
E->M/2(E)	AA substitution	0.858	Thiazolidine(Nterm)	Chemical derivative	0.826	Labile+249	Novel	0.8
Labile+412	Novel	0.857	V->H/2(V)	AA substitution	0.826	Labile+305	Novel	0.8
S-19	Novel	0.855	Labile+286	Novel	0.825			

Figure 5.1 Several hundred different modifications were found, with supporting evidence, when running Maestro on a deep HEK 293 dataset and comparing the resulting PSMs to the Unimod database. A portion of these is shown here.

In order to compare Maestro to existing PTM discovery tools, Maestro and MSFragger were both run on the same dataset. MSFragger is a database search tool that makes use of fragment ion indexing to perform fast open searches [19]. An open MSFragger search and an unclustered Maestro search were run on the above dataset. Both were filtered at 1 percent PSM-level and variant-level FDR. Maestro identified 148690 unmodified peptides total, whereas MSFragger identified 120899. There were 106624 peptides identified by both searches, while Maestro alone identified 42066 peptides, and MSFragger alone identified 14275 peptides (Figure 5.2).

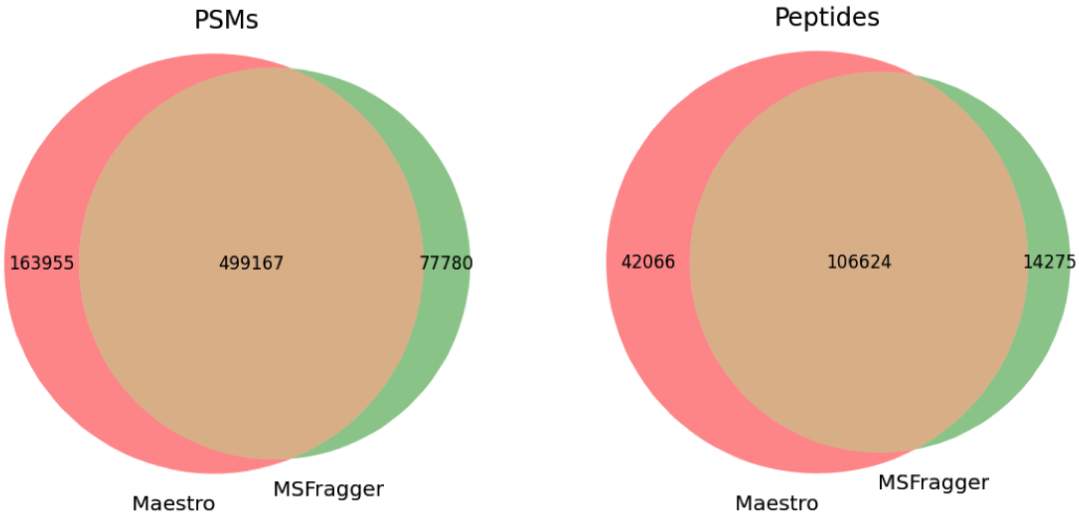


Figure 5.2 Overlap between Maestro and MSFragger identifications. PSM-level results are shown on the left and peptide-level results are shown on the right. Maestro and MSFragger identified 499167 PSMs in common on a HEK 293 dataset, while Maestro identified 163955 PSMs that MSFragger did not identify, and MSFragger identified 77780 PSMs that Maestro did not identify. Maestro and MSFragger identified 106624 peptides in common, while Maestro identified 42066 peptides that MSFragger did not identify, and MSFragger identified 14275 peptides that Maestro did not identify.

Maestro was run on multiple cancer and non-cancer colon datasets in a single search [15] [32] [33], and filtered at 1 percent FDR. This revealed a wide variety of peptide modifications. 26 protein regions had over 100 different associated variants, and 83 protein regions had over 50 different associated variants (Figure 5.3).

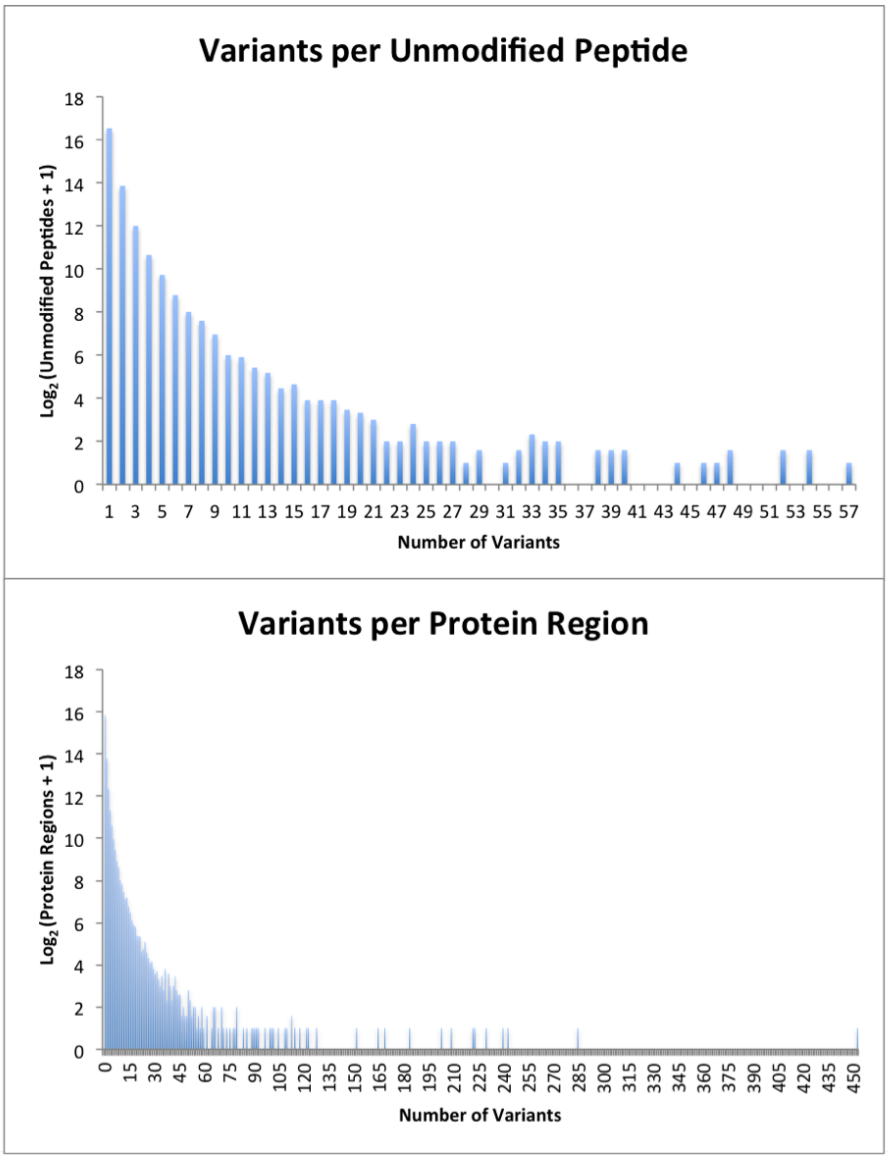


Figure 5.3 a) Distribution of variants per peptide and protein region, from a Maestro run on a mixture of cancer and non-cancer colon datasets. The number of unmodified peptides that has each number of variants is shown on the top, and the number of protein regions that has each number of variants is shown on the bottom.

Figure 5.3 b) Highly-variable protein region with 101 different variants, from a Maestro run on a mixture of cancer and non-cancer colon datasets. Modified amino acids are shown in red, with associated mass offsets above.

Chapter 6

Discussion

Maestro makes use of the unique advantages of library search, database search, and blind PTM search to maximize the number of PSMs found at a given FDR, and find a large number of peptide variants. Library search has high sensitivity, and helps identify mixtures and short peptides that might be missed by database search. Database search increases the search space to identify peptides that are not in the library. Blind search allows for the discovery of unanticipated modifications and highly-modified peptides. It has been demonstrated that by using this strategy, Maestro is able to identify substantially more PSMs at a given FDR than individual search algorithms, and more unique peptides than current tools designed for PTM discovery.

Maestro also produces many result views that aid in interpreting results, and particularly in assessing variant diversity. The Identified Peptide Variants, Identified Peptide Variants by Protein Region, Identified Protein Regions, Identified Network Pairs, Network Pair Modifications, All Network Pairs, Peptide Pairs, Peptide Modifications, Spectral Networks, and Peptide Variant Networks views allow for the variants in a dataset to be viewed and interpreted in a number of ways.

There are many ways in which Maestro could be extended – for example, by adding support for isotopically-labeled data, employing a more rigorous method of computing spectral network alignments, or using Unimod to automatically label and classify peptide modifications. These paths are currently being explored.

Chapter 7

Appendix I: Result View Details

The combined results from M-SPLIT, MS-GF+, and MODa searches, along with other information about spectra, clusters, spectral networks, etc., are analyzed and displayed in several different ways. The following result views are shown to the user at the conclusion of their task.

7.1 Search Result Summary

This view contains overall information about the search results. Several HTML tables are shown, containing aggregate statistics relating to input spectra, clustering, cluster identification, post-translational modifications, and spectral networks. Some statistics are displayed as text and some as hyperlinks to relevant result views.

The first table, titled "Identification Results", contains the following statistics, in total and for each charge.

- Original MS/MS Spectra: Number of raw input spectra, grouped by charge.
- Filtered, Charge-Corrected MS/MS Spectra: Number of input spectra after Kullback-Leibler divergence filtering and charge correction and other pre-filtering, grouped by corrected charge (links to All Spectra view pre-filtered by charge).
- Identified MS/MS Spectra: Number and percentage of spectra identified by M-SPLIT, MS-GF+, or

MODa, grouped by original charge. Percentage is calculated as identified spectra divided by total spectra.

- Clusters: Number of clusters of spectra, grouped by original charge (links to All Clusters view).
- Identified Clusters: Number and percentage of identified clusters, grouped by original charge (links to Identified Clusters view pre-filtered by charge). Percentage is calculated as identified clusters divided by total clusters.
- Identified M-SPLIT Clusters: Number and percentage of clusters identified by M-SPLIT, grouped by identification charge (links to M-SPLIT Clusters view pre-filtered by charge). Percentage is calculated as clusters identified by M-SPLIT divided by clusters identified.
- Identified MS-GF+ Clusters: Number and percentage of clusters identified by MS-GF+, grouped by identification charge (links to MS-GF+ Clusters view pre-filtered by charge). Percentage is calculated as clusters identified by MS-GF+ divided by clusters identified.
- Identified MODa Clusters: Number and percentage of clusters identified by MODa, grouped by identification charge (links to MODa Clusters view pre-filtered by charge). Percentage is calculated as clusters identified by MODa divided by clusters identified.
- Clusters in Networks: Number and percentage of clusters that belong to spectral networks, grouped by original charge. Percentage is calculated as clusters in networks divided by clusters.
- Identified Clusters in Networks: Number and percentage of clusters that belong to networks and were identified by M-SPLIT, MS-GF+, or MODa, grouped by identification charge. Percentage is calculated as identified clusters in networks divided by clusters in networks.
- Unidentified Clusters With Edges to Identified Clusters: Number and percentage of clusters that are immediate neighbors of (adjacent to in the network) identified clusters, but are not identified themselves, grouped by original charge. Percentage is calculated as immediate neighbor clusters divided by clusters.

- Unidentified Clusters in Networks with Identified Clusters, With No Edges to Them: Number and percentage of clusters that are in networks with identified clusters, but are not immediate neighbors and are not identified themselves, grouped by original charge. Percentage is calculated as other neighbor clusters divided by clusters.

- Unidentified Clusters in Networks With No Identified Clusters: Number and percentage of clusters in networks containing no identified clusters, grouped by original charge. Percentage is calculated as clusters with no identifications in the network divided by clusters.

- Clusters Identified or in Networks: Number and percentage of clusters that were identified and/or are in a network, grouped by original charge. Percentage is calculated as clusters identified or in networks divided by clusters.

- Peptide Variants: Number of identified peptide variants, grouped by representative spectrum precursor charge (links to Identified Peptide Variants view pre-filtered by charge).

- Proteins: Number of proteins that have mapping peptide identifications (links to Identified Proteins view). For each charge, the number of proteins containing any variant whose representative spectrum has that precursor charge, is displayed.

Charge categories are 1, 2, 3, 4 or greater, and undetermined.

Two tables relating to post-translational modifications are displayed. The first, titled "All Searched Mods (MS-GF+)" shows default modifications searched for during MS-GF+. The following information about each modification is shown:

- Name of the modification (Oxidation, Lysine Methylation, Pyroglutamate Formation, Phosphorylation, N-terminal Carbamylation, N-terminal Acetylation, or Deamidation)

- Mass offset

- Amino acids on which the modification can occur (an asterisk means that the modification can occur on any amino acid)

- Options (whether the modification is optional or fixed, whether it is N-terminal)
- Number of clusters with that modification

The second PTM table, titled "Top 15 Discovered Mods (MODa)", shows the top 15 modifications discovered during MODa blind search. For each modification, the following information is shown:

- Mass offset
- Total clusters with a modification at given mass offset
- Total variants (distinct peptides) with a modification at given mass offset
- The top five amino acids on which the modification occurs, and the total modifications found on each of these, for given mass offset

A section titled "Spectral Networks" contains three tables.

The first displays the number of components at various network size and percent identified bins. Network size bins are 2, 3, 4, 5, 6-10, 11-15, 16-30, 31-50, 51-100, and 101 or greater. Percent identified bins are 0, 1-24, 25-49, 50-74, 75-89, 90-99, and 100.

The second contains the following information:

- Total network pairs (edges)
- Number of true positive spectral pairs (links to Identified Network Pairs view filtered to contain only true positives)
- Percent true positives, calculated as true positives divided by total positives
- Number of false positives (links to Identified Network Pairs view filtered to contain only false positives)
- Percent false positives, calculated as false positives divided by total positives
- Number of false negatives (links to Identified Network Pairs view filtered to contain only false negatives)

The third Spectral Networks table contains the following information for cosine thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9:

- Precision (proportion of detected network pairs at or above cosine threshold that are true positives)
- Recall (proportion of total correct network pairs that are at or above cosine threshold and are true positives)
- Number of pairs at or above cosine threshold
- Percent of pairs at or above cosine threshold (calculated as pairs at or above threshold divided by total pairs)

The "Groups" section contains two tables.

The first, titled "Outlier Group Counts", shows the number of unique peptide sequences (ignoring modifications), peptide variants, proteins, and unidentified clusters in each default and each user-defined group. Filtering options displayed underneath the table allow the user to specify a minimum absolute group outlier ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value) and spectrum count to display. Minimum outlier ratio options are 0, 1, 2, 4, and infinity. Minimum spectrum count options are 0, 4, and 7. Values are calculated separately for default and user-defined groups. These values take into account only the highest-count or lowest-count group for each cluster (whichever has a higher absolute outlier ratio). If there is a tie in the group outlier ratio, all outlier groups are counted.

A table titled "Top 10 Group Combinations" looks at the number of clusters, identified or not, that belong to each combination of default groups, and to each combination of user groups. The ten default and user group combinations with the most clusters are shown. Each group combination links to the All Clusters view pre-filtered for that group combination.

A section titled "Peptides" analyzes all of the peptide pairs in which one of the peptides was identified by a different algorithm than the other. These pairs are grouped into one of seven categories. This is done for each pair of algorithms (MSGF+ peptides are compared with M-SPLIT peptides, MODa peptides

are compared with M-SPLIT peptides, and MODa peptides are compared with MSGF+ peptides). The categories are as follows for a comparison of an MS-GF+ peptide with an M-SPLIT peptide (the comparisons for the other two algorithm pairs are defined analogously).

- Type 0: The MS-GF+ peptide has exact same sequence, the same number of modifications, and the same charge as an M-SPLIT peptide.
- Type 1: The MS-GF+ peptide has the exact same sequence, the same number of modifications, and a different charge from an M-SPLIT peptide.
- Type 2: The MS-GF+ peptide sequence is contained within an M-SPLIT peptide sequence.
- Type 3: An M-SPLIT peptide sequence is contained within the MS-GF+ peptide sequence.
- Type 4: An M-SPLIT peptide has the exact same sequence and more modifications than the MS-GF+ peptide.
- Type 5: An M-SPLIT peptide has the exact same sequence and fewer modifications than the MS-GF+ peptide.
- Type 6: Exactly one residue is changed to another in the MS-GF+ peptide with respect to an M-SPLIT peptide.
- Type 7: Other. (There is no M-SPLIT peptide that corresponds to the MS-GF+ peptide.)

7.2 Identified Clusters

For each cluster identified according to a spectrum-level false discovery rate, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Filename (will show cluster filename if clustering was turned on, and original filename otherwise)
- Cluster index

- Original charge
- Algorithm that identified the cluster
- Variant group (arbitrary group number assigned to each distinct peptide)
- Peptide identification
- Identification charge
- False discovery rate for cluster
- False discovery rate for peptide variant
- Unmodified peptide sequence (peptide with modifications removed)
- Protein region(s) (peptide variant regions on the protein)
- Protein(s) corresponding to peptide
- Spectral network component index for cluster, if any
- Number of network neighbors
- Number of spectra in the cluster
- Number of spectra in clusters assigned to this variant group
- Default groups to which the cluster belongs
- User-defined groups to which the cluster belongs
- Start location (amino acid) on protein
- End location (amino acid) on protein
- Protein region index(es)
- Number of modifications

- All modification offsets in peptide, rounded to the nearest integer
- Number of spectra assigned to cluster that belong to each default and user-defined group (a separate column is shown for each group)

7.3 Identified Peptide Variants

For each peptide variant group identified according to a variant-level false discovery rate, the following information is displayed in a ProteoSAFe table. Information for each protein region corresponding to the variant is combined into one row, so that a single row is shown for each variant.

- Lorikeet icon (displays the mass spectrum when clicked)
- Variant group (arbitrary group number assigned to each distinct peptide)
- Peptide identification
- Parent mass
- Original charge
- Number of modifications
- All modification offsets in peptide, rounded to the nearest integer
- Protein region indices (for all protein regions corresponding to variant)
- Protein regions (all protein regions corresponding to variant)
- Proteins (for all protein regions corresponding to variant)
- Unmodified peptide sequence (peptide with modifications removed)
- False discovery rate for variant
- All proteins corresponding to peptide

- Number of network neighbors
- Spectral network component indices for all clusters identified as variant
- Number of spectra in clusters assigned to this variant group
- Number of spectra in clusters identified as this peptide, ignoring modifications
- Number of spectra assigned to variant that belong to each default group (a separate column is shown for each group)
- Outlier default group (default group with highest or lowest variant spectrum count for given variant; see group ratio description)
- Outlier default group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)
- Number of spectra assigned to variant that belong to each user-defined group (a separate column is shown for each group)
- Outlier user group (user group with highest or lowest variant spectrum count for given variant; see group ratio description)
- Outlier user group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)

A dropdown arrow, when clicked, displays the information from the "Identified Clusters" view for each of the clusters annotated as this variant.

7.4 Identified Proteins

For each protein corresponding to one or more variants identified according to a variant-level false discovery rate, the following information is shown in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Protein
- Sum of network neighbors for all corresponding clusters
- Number of spectra identified as belonging to this protein, in total and listed separately for each default and user group
- Number of spectra identified as belonging to this protein, only taking into account peptides that unambiguously belong to/ are unique to this protein, in total and listed separately for each default and user group
- Protein outlier default group (default group with highest or lowest protein spectrum count for given protein; see group ratio description)
- Protein default group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)
- Unique protein outlier default group (default group with highest or lowest unique protein spectrum count for given protein; see group ratio description)
- Unique protein default group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)
- Protein outlier user group (user group with highest or lowest protein spectrum count for given protein; see group ratio description)
- Protein user group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)

- Unique protein outlier user group (user group with highest or lowest unique protein spectrum count for given protein; see group ratio description)
- Unique protein user group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)

A dropdown arrow, when clicked, displays the information from the "Identified Peptide Variants" view for each of the variants belonging to that protein.

7.5 All Clusters

This view contains information about all clusters, whether identified or not. For each cluster, the following columns are displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Cluster index
- Link to Network Displayer view
- Number of spectra in the cluster
- Number of files corresponding to the cluster
- Precursor mass-to-charge ratio
- Precursor charge
- Precursor intensity
- Retention time
- Default groups to which the cluster belongs
- Peptide identification, if any

- Number of spectra in each default group
- Outlier default group (default group with highest or lowest unique protein spectrum count for given protein; see group ratio description)
- Outlier default group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)
- Number of spectra in each user-defined group
- Outlier user group (default group with highest or lowest unique protein spectrum count for given protein; see group ratio description)
- Outlier user group ratio (base-2 logarithm of the ratio of the highest group count to the second-highest group count, or the lowest group count to the second-lowest group count – whichever has greater absolute value; 100 is used in place of infinity and -100 is used in place of negative infinity)

7.6 All Spectra

This view contains information about pre-identification spectra. For each spectrum, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Cluster index (cluster that given spectrum was grouped into)
- Filename
- Scan number
- Precursor intensity
- Retention time

- Parent mass
- Precursor charge
- Kullback-Leibler divergence

7.7 Identified Spectra

For each identified spectrum, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Cluster index (cluster that given spectrum was grouped into)
- Filename
- Scan number
- Cluster index
- Peptide sequence
- Protein
- Precursor intensity
- Retention time
- Parent mass
- Precursor charge
- False discovery rate for cluster
- Kullback-Leibler divergence

7.8 mzTab Result Files

Displays result files in mzTab format.

7.9 M-SPLIT Clusters

For each cluster identified by M-SPLIT, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Filename
- Cluster index
- Peptide sequence
- Protein
- All proteins containing peptide
- Charge
- Cosine(M, A+B) (where A is a spectrum corresponding to an individual peptide, B is a spectrum corresponding to a different individual peptide, and M is a mixture spectrum represented as a linear combination of A and B)
- Cosine(M, A)
- Cosine(A, B)
- Alpha (as estimated by optical cosine and by residual method)
- Residual alpha
- SVM1 score (support vector machine – this score determines whether the result is a match)

- SVM2 score (determines whether the result is a mixture match)
- Mean Kullback-Leibler divergence for spectra belonging to cluster
- Maximum Kullback-Leibler divergence for spectra belonging to cluster
- Spectral libraries containing peptide

7.10 M-SPLIT Peptides

For each unmodified peptide annotation, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Filename
- Cluster index
- Peptide sequence
- Protein
- All proteins containing peptide
- Charge
- Cosine(M, A+B) (where A is a spectrum corresponding to an individual peptide, B is a spectrum corresponding to a different individual peptide, and M is a mixture spectrum represented as a linear combination of A and B)
- Cosine(M, A)
- Cosine(A, B)
- Alpha (as estimated by optical cosine and by residual method)

- Residual alpha
- SVM1 score (support vector machine – this score determines whether the result is a match)
- SVM2 score (determines whether the result is a mixture match)
- Mean Kullback-Leibler divergence for spectra belonging to cluster
- Maximum Kullback-Leibler divergence for spectra belonging to cluster
- Spectral libraries containing peptide

A dropdown arrow, when clicked, displays the information from the M-SPLIT Clusters view, for each cluster annotated as that peptide.

7.11 M-SPLIT Proteins

For each protein corresponding to one or more peptide annotations of spectra, the following information is displayed in a ProteoSAFe table:

- Protein (links to the M-SPLIT Clusters view, sliced to show information for each cluster corresponding to this protein)
- Number of hits
- Number of unique unmodified peptides
- Number of unique modified peptides

A dropdown arrow, when clicked, displays the information from the M-SPLIT Clusters view, for each cluster corresponding to that protein.

7.12 MS-GF+ Clusters

For each cluster identified by MS-GF+, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Filename
- Cluster index
- Fragmentation method
- Peptide sequence
- Protein
- All proteins containing peptide
- Charge
- Raw score of the peptide-spectrum match
- Spectrum probability (the probability that a random peptide matched to the spectrum would produce a score at least as high as the score of the peptide-spectrum match)
- Database E-value (the probability that a random database matched to the spectrum would produce a score at least as high as the score of the peptide-spectrum match)
- False discovery rate
- Peptide-level false discovery rate
- Mean Kullback-Leibler divergence for spectra belonging to cluster
- Maximum Kullback-Leibler divergence for spectra belonging to cluster

7.13 MS-GF+ Peptides

For each peptide annotation, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)

- Filename
- Cluster index
- Fragmentation method
- Peptide sequence
- Protein
- All proteins containing peptide
- Charge
- Raw score of the peptide-spectrum match
- Spectrum probability (the probability that a random peptide matched to the spectrum would produce a score at least as high as the score of the peptide-spectrum match)
- Database E-value (the probability that a random database matched to the spectrum would produce a score at least as high as the score of the peptide-spectrum match)
- False discovery rate
- Peptide-level false discovery rate
- Mean Kullback-Leibler divergence for spectra belonging to cluster
- Maximum Kullback-Leibler divergence for spectra belonging to cluster

A dropdown arrow, when clicked, displays the information from the MS-GF+ Clusters view, for each cluster identified as that peptide.

7.14 MS-GF+ Proteins

For each protein corresponding to one or more peptide annotations of spectra, the following information is displayed in a ProteoSAFe table:

- Protein (links to the MS-GF+ Clusters view, sliced to show information for each cluster corresponding to this protein)
- Number of hits
- Number of unique unmodified peptides
- Number of unique modified peptides

A dropdown arrow, when clicked, displays the information from the MS-GF+ Clusters view, for each cluster corresponding to that protein.

7.15 MODa Clusters

For each cluster identified by MODa, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Filename
- Cluster index
- Charge
- Observed molecular weight
- Calculated molecular weight
- Mass difference
- Match score
- Match probability
- Peptide sequence
- Protein

- All proteins containing peptide
- Peptide position within the protein
- Mean Kullback-Leibler divergence for spectra belonging to cluster
- Maximum Kullback-Leibler divergence for spectra belonging to cluster

7.16 MODa Peptides

For each peptide annotation, the following information is displayed in a ProteoSAFe table:

- Lorikeet icon (displays the mass spectrum when clicked)
- Filename
- Cluster index
- Charge
- Observed molecular weight
- Calculated molecular weight
- Mass difference
- Match score
- Match probability
- Peptide sequence
- Protein
- All proteins containing peptide
- Peptide position within the protein

- Mean Kullback-Leibler divergence for spectra belonging to cluster
- Maximum Kullback-Leibler divergence for spectra belonging to cluster

A dropdown arrow, when clicked, displays the information from the MODa Clusters view, for each cluster identified as that peptide.

7.17 MODa Proteins

For each protein corresponding to one or more peptide annotations of spectra, the following information is displayed in a ProteoSAFe table:

- Protein (links to the MODa Clusters view, sliced to show information for each cluster corresponding to this protein)
- Number of hits
- Number of unique unmodified peptides
- Number of unique modified peptides

A dropdown arrow, when clicked, displays the information from the MODa Clusters view, for each cluster corresponding to that protein.

7.18 Identified Network Pairs

Pairs of peptides are displayed in a ProteoSAFe table. For each peptide, the following information is displayed:

- Spectrum index
- Peptide sequence
- Protein

- Charge

Additionally, for each peptide pair, the following information is displayed:

- Mass offsets present in both peptides, rounded to the nearest integer
- Cosine score of pair
- Mass-to-charge ratio difference of the pair
- Whether the pair is correct. 0 means the pair is incorrect, 1 means it is correct, and 2 means it is ambiguous (the pair shares over 60% of prefix residue masses).
- Whether the pair is detected (the clusters are spectral network neighbors).

A Correct value of 1 (true) combined with an Detected value of 1 indicates a true positive, a Correct value of 0 (false) and Detected value of 1 indicates a false positive, and a Correct value of 1 and Detected value of 0 indicates a false negative.

7.19 Network Pair Modifications

This table contains the number of peptide pairs at each mass offset, within each cosine bin. The cosine bins are:

- Less than 0.45
- At least 0.45 and less than 0.55
- At least 0.55 and less than 0.65
- At least 0.65 and less than 0.75
- At least 0.75 and less than 0.85
- At least 0.85 and less than 0.95
- At least 0.95

7.20 All Network Pairs

Shows the following information for each spectral network pair:

- Cluster index of first node in pair
- Cluster index of second node in pair
- Cosine score between the nodes
- Difference in mass-to-charge ratio between the nodes

7.21 Peptide Pairs

This view displays the grouping of MS-GF+ and MODa peptides into categories according to their similarity to peptides identified by earlier workflows. Peptides identified by MS-GF+ (that were not identified by M-SPLIT) are compared with peptides identified by M-SPLIT, and peptides identified by MODa (that were not identified by M-SPLIT or MS-GF+) are compared with peptides identified by M-SPLIT and with peptides identified by MS-GF+.

The possible categories for an MS-GF+-M-SPLIT peptide pair (meaning that each MS-GF+ peptide is being grouped with respect to the set of M-SPLIT peptides) are as follows:

Category 0: The MS-GF+ peptide has exact same sequence, the same number of modifications, and the same charge as an M-SPLIT peptide.

Category 1: The MS-GF+ peptide has the exact same sequence, the same number of modifications, and a different charge from an M-SPLIT peptide.

Category 2: The MS-GF+ peptide sequence is contained within an M-SPLIT peptide sequence.

Category 3: An M-SPLIT peptide sequence is contained within the MS-GF+ peptide sequence.

Category 4: An M-SPLIT peptide has the exact same sequence and more modifications than the MS-GF+ peptide.

Category 5: An M-SPLIT peptide has the exact same sequence and fewer modifications than the MS-GF+ peptide.

Category 6: Exactly one residue is changed to another in the MS-GF+ peptide with respect to an M-SPLIT peptide.

Category 7: Other. (There is no M-SPLIT peptide that corresponds to the MS-GF+ peptide.)

The categories for the other two types of peptide pairs (MODA-MSPLIT pairs and MODA-MSGF+ pairs) are defined analogously.

For each pair, this result view displays the following information in a ProteoSAFe table:

- Lorikeet icon that displays both spectra in the pair when clicked
- Description of category
- Category ID of the pair
- Workflow for each cluster in the pair
- Filename for each cluster in the pair
- Cluster index for each cluster in the pair
- Peptide sequence for each cluster in the pair
- Charge for each cluster in the pair

Peptides in Category 7 could not be paired with any other peptides, so rows containing peptides grouped into Category 7 only include data for those peptides, and not for any paired peptides.

7.22 Peptide Modifications

This view shows the number of occurrences of each mass offset at each amino acid. The table contains the following columns, for each mass offset:

- Column for each amino acid containing the number of modifications of the amino acid in the column with the mass offset in the row
- C-terminus (whether the modification occurred at the C-terminus of the peptide)

- N-terminus (whether the modification occurred at the N-terminus of the peptide)
- Sum (total modifications at the offset represented in the row)

7.23 Spectral Networks

This view is useful for analyzing the network associated with each cluster. It displays the following information in a ProteoSAFe table, for each network component:

- Index of network component
- Link to Network Displayer view containing network visualization
- Link to Cluster Info view for given component index
- Number of clusters (nodes)
- Number of identified clusters
- Percentage of clusters that are identified
- Number of spectra, in total and per default group
- Peptides for all clusters in network component
- Unmodified peptides (peptides with modifications removed) for all clusters in network component
- Cluster indices for all clusters in network component
- Default groups for all clusters in network component
- User-defined groups for all clusters in network component
- Variant groups for all clusters in network component

Chapter 8

References

[1] N. Bandeira, D. Tsur, A. Frank, and P. A. Pevzner. Protein Identification by Spectral Networks Analysis. *P. Natl. Acad. Sci. USA*, 104:6140–6145, 2007.

[2] C. Bartels. Probability-Based Protein Identification by Searching Sequence Databases using Mass Spectrometry Data. *Electrophoresis*, 20:3551–3567, 1999.

[3] J. M. Chick, D. Kolippakkam, D. P. Nusinow, B. Zhai, R. Rad, E. L. Huttlin, and S. P. Gygi. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nat. Biotechnol.*, 33:743–749, 2015.

[4] R. Craig and R. C. Beavis. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics*, 20:1466–1467, 2004.

[5] R. Craig, J. C. Cortens, D. Fenyo, and R. C. Beavis. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J. Proteome Res.*, 5:1843–1849, 2006.

[6] D. M. Creasy and J. S. Cottrell. Unimod: Protein Modifications for Mass Spectrometry. *Proteomics*, 4:1534–1536, 2004.

[7] J. K. Eng, A. L. McCormack, and J. R. Yates III. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectr.*, 5:976–989, 1994.

[8] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, 246:64–71, 1989.

[9] A. M. Frank, N. Bandeira, Z. Shen, S. Tanner, S. P. Briggs, R. D. Smith, and P. A. Pevzner. Clustering Millions of Tandem Mass Spectra. *J. Proteome Res.*, 7:113–122, 2008.

[10] M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer, and G. D. Bader. Cytoscape.js: A Graph Theory Library for Visualisation and Analysis. *Bioinformatics*, 32:309–311, 2016.

[11] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open Mass Spectrometry Search Algorithm. *J. Proteome Res.*, 3:958–964, 2004.

[12] J. Griss, A. R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G. G. Thallinger, R. M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, S. Neumann, J. Fan, F. Reisinger, Q. Xu, N. del Toro, Y. Pérez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaíno, and H. Hermjakob. The MzTab Data Exchange Format: Communicating Mass-Spectrometry-Based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Mol. Cell. Proteomics*, 13:2765–2775, 2014.

[13] A. Guthals, J. D. Watrous, P. C. Dorrestein, and N. Bandeira. The Spectral Networks Paradigm in High Throughput Mass Spectrometry. *Mol. Biosyst.*, 8:2535–2544, 2012.

[14] M. Karas and F. Hillenkamp. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10,000 Daltons. *Anal. Chem.*, 60:2299–2301, 1988.

[15] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A Draft Map of The Human Proteome. *Nature*, 509:575–581, 2014.

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

[17] S. Kim, N. Mischerikow, N. Bandeira, J. D. Navarro, L. Wich, S. Mohammed, A. J. R. Heck, and P. A. Pevzner. The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search. *Mol. Cell. Proteomics*, 9:2840–2852, 2010.

[18] S. Kim and P. A. Pevzner. MS-GF+ Makes Progress Towards a Universal Database Search Tool for Proteomics. *Nat. Commun.*, 5, 2014.

[19] A. T. Kong, F. V. Leprevost, D. N. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods*, 14:513–520, 2017.

[20] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold. Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics*, 7:655–667, 2007.

[21] S. Na, N. Bandeira, and E. Paek. Fast Multi-blind Modification Search through Tandem Mass Spectrometry. *Mol. Cell. Proteomics*, 11, 2012.

[22] S. Nahnsen, C. Bielow, K. Reinert, and O. Kohlbacher. Tools for Label-Free Peptide Quantification. *Mol. Cell. Proteomics*, 12:549–556, 2013.

[23] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. *Nat. Methods*, 4:709–712, 2007.

[24] V. Sharma, J. K. Eng, M. J. Maccoss, and M. Riffle. A Mass Spectrometry Proteomics Data Management Platform. *Mol. Cell. Proteomics*, 11:824–831, 2012.

[25] S. E. Stein. An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data. *J. Am. Soc. Mass Spectr.*, 10:770–781, 1999.

[26] J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proc. Natl. Acad. Sci. U. S. A.*, 101:9528–9533, 2004.

[27] S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P. Pevzner, and V. Bafna. InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. *Anal. Chem.*, 77:4626–4639, 2005.

[28] J. Wang, J. Perez-Santiago, J. E. Katz, P. Mallick, and N. Bandeira. Peptide Identification from Mixture Tandem Mass Spectra. *Mol. Cell. Proteomics*, 9:1476–1485, 2010.

[29] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W. T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C. C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C. C. Liaw, Y. L. Yang, H. U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. B. P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P. M. Allard, P. Phapale, L. F. Nothias, T. Alexandrov, M. Litaudon, J. L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D. T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, and N. Bandeira. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.*, 34:828–837, 2016.

[30] J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein. Mass Spectral Molecular Networking of Living Microbial Colonies. *Proc. Natl. Acad. Sci. U. S. A.*, 109:E1743–E1752, 2012.

[31] J. M. Wells and S. A. McLuckey. Collision-Induced Dissociation (CID) of Peptides and Proteins. *Methods Enzymol.*, 402:148–185, 2005.

[32] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J. H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster. Mass-Spectrometry-Based Draft of the Human Proteome. *Nature*, 509:582–587, 2014.

[33] J. R. Wiśniewski, K. Duś-Szachniewicz, P. Ostasiewicz, P. Ziółkowski, D. Rakus, and M. Mann. Absolute Proteome Analysis of Colorectal Mucosa, Adenoma, and Cancer Reveals Drastic Changes in Fatty Acid Metabolism and Plasma Membrane Transporters. *J. Proteome Res.*, 14:4005–4018, 2015.

[34] J. R. Yates III, S. F. Morgan, C. L. Gatlin, P. R. Griffin, and J. K. Eng. Method To Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis. *Anal. Chem.*, 70:3557–3565, 1998.