Journal of proteome research

Article

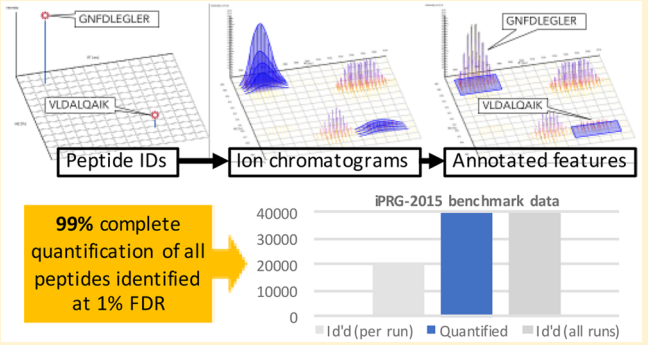
pubs.acs.org/jpr

Targeted Feature Detection for Data-Dependent Shotgun Proteomics

Hendrik Weisser[†] and Jyoti S. Choudhary*

Proteomic Mass Spectrometry, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, United Kingdom

ABSTRACT: Label-free quantification of shotgun LC−MS/MS data is the prevailing approach in quantitative proteomics but remains computationally nontrivial. The central data analysis step is the detection of peptide-specific signal patterns, called features. Peptide quantification is facilitated by associating signal intensities in features with peptide sequences derived from MS2 spectra; however, missing values due to imperfect feature detection are a common problem. A feature detection approach that directly targets identified peptides (minimizing missing values) but also offers robustness against false-positive features (by assigning meaningful confidence scores) would thus be highly desirable. We developed a new feature detection algorithm within the OpenMS software framework, leveraging ideas and algorithms from the OpenSWATH toolset for DIA/SRM data analysis. Our software, FeatureFinderIdentification ("FFId"), implements a targeted approach to feature detection based on information from identified peptides. This information is encoded in an MS1 assay library, based on which ion chromatogram extraction and detection of feature candidates are carried out. Significantly, when analyzing data from experiments comprising multiple samples, our approach distinguishes between "internal" and "external" (inferred) peptide identifications (IDs) for each sample. On the basis of internal IDs, two sets of positive (true) and negative (decoy) feature candidates are defined. A support vector machine (SVM) classifier is then trained to discriminate between the sets and is subsequently applied to the "uncertain" feature candidates from external IDs, facilitating selection and confidence scoring of the best feature candidate for each peptide. This approach also enables our algorithm to estimate the false discovery rate (FDR) of the feature selection step. We validated FFId based on a public benchmark data set, comprising a yeast cell lysate spiked with protein standards that provide a known ground-truth. The algorithm reached almost complete (>99%) quantification coverage for the full set of peptides identified at 1% FDR (PSM level). Compared with other software solutions for label-free quantification, this is an outstanding result, which was achieved at competitive quantification accuracy and reproducibility across replicates. The FDR for the feature selection was estimated at a low 1.5% on average per sample (3% for features inferred from external peptide IDs). The FFId software is open-source and freely available as part of OpenMS (www.openms.org).



KEYWORDS: *bioinformatics, machine learning, shotgun proteomics, label-free quantification, feature detection*

INTRODUCTION

Shotgun proteomics enables the identification and quantification of proteins in complex mixtures in a high-throughput fashion.[1] After enzymatic digestion of the proteins, resulting peptides are analyzed by liquid chromatography coupled to tandem mass spectrometry (LC−MS/MS). In data-dependent acquisition (DDA) mode, the MS instrument acquires precursor ion spectra (MS1) and selects candidate precursors based on charge and intensity for fragmentation and fragment ion spectra (MS2) acquisition. To quantify peptides in LC−MS/MS data and thereby enable the inference of protein abundances, the identities of peptides and the corresponding quantitative measures both have to be available. The quantitative information is contained in the MS1 spectra, and sequence information required for identification is captured in the MS2 spectra. DDA thus allows for a discovery-driven approach to protein identification and quantification that requires only minimal preparation or a priori knowledge of the sample.

Peptide identification based on MS2 spectra is now a routine operation supported by many effective software tools for sequence database searching[2−4] or spectral library searching.[5] Importantly, target-decoy search strategies and mature algorithms for postprocessing of search results[6,7] allow the statistical validation and the assignment of meaningful confidence values for peptide identifications. Nonetheless, the area continues to evolve and improve.[8]

One of the main strategies for quantifying the peptides and proteins observed in shotgun LC−MS/MS experiments is label-free quantification. This is based on the principle that signal intensities in mass-spectrometry data are proportional to

Received: April 25, 2017
Published: July 4, 2017

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

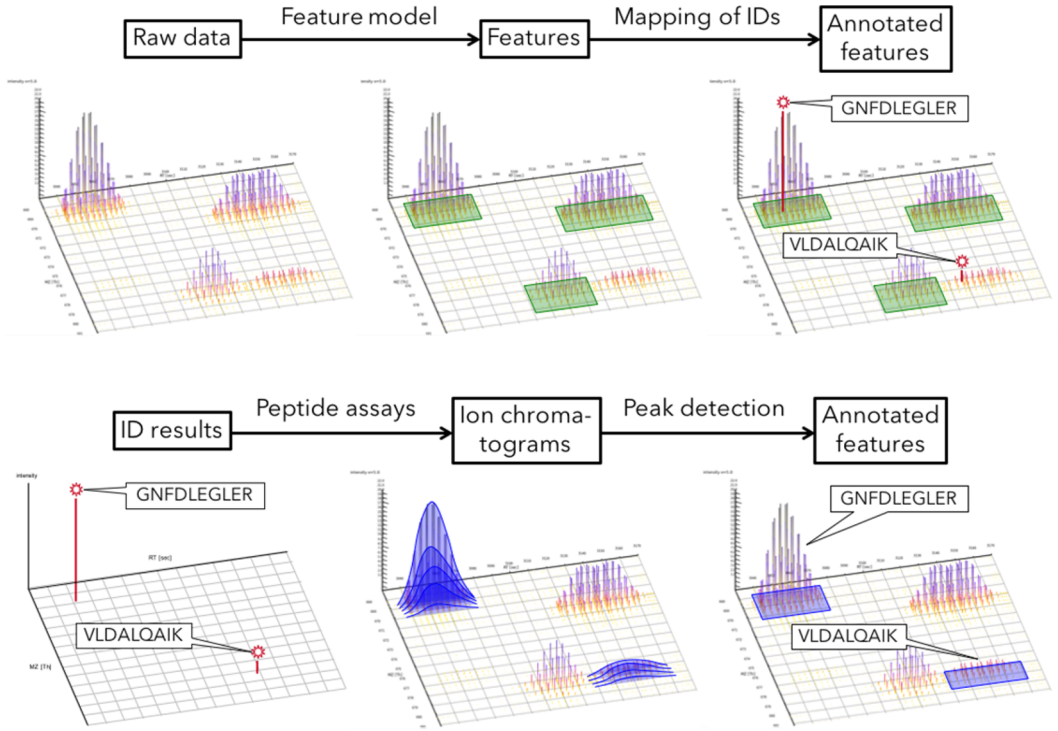


Figure 1. Schematic overview of the two principal approaches to feature detection in label-free LC−MS/MS data. Top: "De novo" feature detection, for example, using FeatureFinderCentroided and IDMapper in OpenMS. Bottom: Targeted feature detection using FeatureFinderIdentification.

the amounts of ions that generated the signals (within the dynamic range limits of the instrument). A peptide can thus be quantified by finding all, or a representative fraction, of the signals that were caused by its ions and integrating the signal intensities. In an LC−MS/MS experiment, ions of each detected peptide are observed at multiple time points, as molecules of the peptide elute from the chromatography column over time. Ions are also observed at multiple locations in the mass-to-charge (m/z) dimension, prominently due to multiple charge states, but independent of that due to different isotopic compositions of the detected peptide molecules. Each peptide with distinct charge thus causes a characteristic signal pattern, called "feature". A feature is defined by a roughly Gaussian-shaped peak in the retention time (RT) dimension and multiple mass traces for the most prevalent isotopologues (variants of the peptide with different isotopic compositions) in the m/z dimension. Software tools can detect peptide features in LC−MS data based on these properties, in what we call the "de novo" approach to feature detection (illustrated in Figure 1, top). This classical approach is used by many important software tools for label-free quantification, including SuperHirn,[9] MaxQuant,[10] and OpenMS (e.g., FeatureFinderCentroided).[11] However, decoupling feature detection from peptide identification in the "de novo" approach leads to a common problem: Some peptides that were identified with high confidence may not be quantifiable, because no corresponding feature was found. This causes missing values in the peptide or protein expression matrix, hindering downstream analyses. (Less problematically, "de novo" approaches typically also produce an overabundance of features that cannot be annotated with peptide sequences. These unannotated features could be used to guide MS2 acquisition in subsequent LC−MS/MS runs using inclusion lists.)

To overcome this "missing value problem" that is widely associated with label-free quantification, an alternative to the "de novo" approach has to be used for feature detection. In what we call the "targeted" approach, a set of high-confidence (e.g., 1% false discovery rate (FDR)) peptide identifications from an experiment is taken as "given" and the information therein is utilized for feature detection. Software tools for label-free quantification that implement a targeted approach include Skyline[12] and moFF.[13] The DeMix-Q workflow[14] contains a hybrid algorithm that uses both "de novo" and targeted feature detection phases.

We present here a new algorithm for targeted feature detection, termed FeatureFinderIdentification (FFId), that we developed within the OpenMS software framework.[15] The FFId algorithm builds on concepts from the targeted analysis of multiplexed MS2 data that are utilized in data-independent acquisition (DIA) experiments.[16] Inspired by selected reaction monitoring (SRM), targeted protein quantification of DIA data relies on predefined "in silico" assays for proteins. Each protein assay consists of so-called transitions, which correspond to selected fragment ions of proteotypic peptides of the protein. The transitions capture information about precursor and fragment mass-to-charge ratios, expected retention times, and relative intensities. To quantify, software extracts data corresponding to the assays from the MS2 spectra and estimates protein abundances based on the resulting intensities. We translated these principles to the analysis of MS1 data from DDA experiments. Our approach, illustrated in Figure 1 (bottom), generates an "in silico" MS1 assay library for the target peptides. Each assay captures information about different isotopologues of its target peptide, in what we call "MS1 transitions" by analogy to the transitions that make up SRM/DIA assays. The MS1 transitions combine measured data (peptide retention times, charge states) with theoretical

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

properties inferred from the peptide sequences (mass-to-charge ratios, isotopic distributions). On the basis of the assay library, we detect features corresponding to the identified peptides by applying methods from the OpenSWATH toolset[17] that is part of OpenMS.

Significantly, we implemented a sophisticated approach for utilizing peptide identifications from multiple LC−MS/MS runs. In shotgun mass spectrometry, the number of MS2 spectra that can be acquired in a given time interval is limited. In analyses of reasonably complex samples, this leads to undersampling of the available precursor ions by MS2. Thus, not all peptides that are present in the sample and that would be detectable on the MS1 level can be identified based on MS2 spectra in a single shotgun LC−MS/MS run. However, these runs are typically not acquired and analyzed in isolation but in the context of multiple related runs (e.g., different conditions, biological/technical replicates) that together make up an experiment. Indeed, the relative quantification afforded by label-free approaches presupposes a comparison of different samples. If multiple related LC−MS/MS runs in an experiment are acquired using DDA, then the stochasticity of the MS2 precursor selection means that different subsets of the available peptides can be identified in each run. The set of peptides identified over all runs will thus be larger than the set of peptides identified in any individual run. Naturally, it would be desirable to quantify all peptides from the larger overall set in each run. To this end, approaches for label-free quantification usually facilitate the inference of peptide IDs across related LC−MS/MS runs. In our FFId algorithm, we incorporate peptide identifications from related runs using a machine learning-based approach, which enables the assignment of meaningful confidence scores and the estimation of a FDR for the detected features.

With FFId, we aim for essentially complete coverage of the given set of peptides in terms of detected features while providing state-of-the-art quantitative accuracy. We will show, based on the evaluation of a benchmark data set, that both of these goals are achieved. This makes FFId an excellent choice for feature detection in a label-free quantification workflow and a valuable addition to the suite of data analysis tools provided by OpenMS.

TERMINOLOGY

To avoid misunderstandings, we define here certain terms that we will use consistently throughout the text. First, we distinguish between an "identified peptide" and a "peptide identification (ID)". An identified peptide is characterized solely by its sequence and modifications. (We consider peptides with the same sequence but different modifications as different peptides.) A peptide ID is synonymous with a peptide-spectrum match (PSM); it comprises the identified peptide together with the retention time and precursor m/z value of the originating spectrum as well as a quality score from the search engine (or statistical validation tool).

Consider one LC−MS/MS run "A" out of multiple runs that belong to the same experiment. We refer to the set of peptide IDs derived directly from run "A" as its "internal peptide identifications" ("internal IDs") and to the peptide IDs from other runs of the experiment as "A's" "external peptide identifications" ("external IDs"). Accordingly, "internal/external peptides" denote the sets of distinct peptides (including modifications) identified by internal/external IDs, and "internal/external features" mean the sets of features detected based on internal/external IDs in the respective LC−MS/MS run.

MATERIALS AND METHODS

Software

We developed our feature detection algorithm in C++ within the OpenMS open-source framework for computational mass spectrometry,[15] version 2.1. Internally, our software makes use of the OpenSWATH toolset for DIA data analysis[17] and of the LIBSVM C library,[18] version 3.12. The two libraries are included in or come bundled with OpenMS, respectively. In the analysis of the benchmark data set (see below), we used several of the executable tools provided by OpenMS for various data processing tasks. For downstream data analysis we used the R environment for statistical computing,[19] versions 3.3.0 and 3.3.2., and especially the MSstats R package,[20] version 3.6.0, for quantification on the protein level.

Benchmark Data Set

We evaluated our algorithm based on data from the iPRG-2015 study.[21] The data set consists of four biological samples, analyzed in three technical replicates (12 LC−MS/MS runs in total). The samples contain yeast cell lysate with spike-ins of six different purified proteins at varying concentrations. LC−MS/MS data were acquired on a Thermo Scientific Q Exactive orbitrap instrument, then centroided and converted to mzML using the ProteoWizard software. We downloaded the data in mzML format from the study repository (ftp://iprg_study:ABRF329@ftp.peptideatlas.org).

This data set had been analyzed by Zhang and colleagues for their publication on the DeMix-Q workflow, which included a benchmark of multiple other software tools for label-free quantification.[14] Information about the spike-in proteins and the amounts used was taken from supplementary table 1 of this publication. The authors also kindly made their peptide/protein identification results available, which were generated using the DeMix search pipeline.[22] The results were prefiltered to 1% FDR on the PSM level. We performed the peptide-to-protein mapping and removed nonmatching (decoy) peptides using the OpenMS tools PeptideIndexer and IDFilter. Afterward the data contained around 28 300 (±2200) peptide IDs per file, corresponding to around 21 100 (±1800) distinct peptide sequences (including modifications) per file. We used this data in our analysis to enable a comparison with DeMix-Q and, by extension, with the other software tools tested by Zhang et al.

FeatureFinderIdentification Analyses. In analyses incorporating external peptide IDs, prior to feature detection with FFId, retention time alignments of those IDs were performed using OpenMS tools. For each LC−MS/MS run, its internal IDs were used as reference to align external IDs from other runs, using MapAlignerIdentification[11] with the "lowess" smoothing option. Subsequently external IDs from multiple runs were merged using IDMerger. In all FFId analyses, two isotopologues per peptide (monoisotopic and ^{13}C peak) were used in the generation of assays. The expected elution peak width was set to 40 s. The m/z window size for chromatogram extraction was 10 ppm.

Variability of Peptide Quantification. To compute coefficients of variation (CVs) of peptide abundances, FFId feature detection results were first processed with ProteinQuantifier to obtain peptide quantities. The peptide data was loaded into R and normalized using simple scaling to equal medians to correct for global differences between the LC−MS/

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

MS runs. The following data processing choices were made to allow comparisons to the results reported for DeMix-Q and other software tools:[14] Abundances of differently modified variants of the same peptide sequence were summed. Only peptides that were quantified in all 12 runs were considered. Peptides from spike-in proteins were not excluded from the analysis. (Although their abundances are not expected to be constant, the influence on median CVs is negligible.) Finally, CVs for the resulting peptide abundances were calculated.

Protein Quantification. To perform protein quantification, FFId feature detection results were exported using OpenMS' ProteinQuantifier with the "filter_charge" option enabled, generating separate results for peptides in different charge states. The exported data were loaded into R and filtered to remove peptides mapping to more than one protein or with missing values in more than half of the samples. The results were formatted to fit the requirements for MSstats input data and analyzed using MSstats functions (with default options, unless otherwise noted).[20] Initial processing, including normalization, was performed using the "dataProcess" function, selecting the top 10 features for quantification. Protein abundances for the four biological samples were calculated using the "quantification" function. Protein fold changes for all sample pairs were calculated using the "groupComparison" function.

Feature Selection for External IDs. The following procedure was used to assess whether our algorithm successfully selected the correct features for external peptide IDs, when separate sets of internal and external IDs were given as input (see Strategy 2: Machine Learning section). For each LC−MS/MS run from the iPRG-2015 data set, the FFId feature detection result based on only the internal peptide IDs was regarded as the gold standard feature set for that run, determining which features were considered as correct for the corresponding charged peptides. The test queried whether the same features were selected if some of the internal IDs were replaced by external IDs (with different RT values) for the same peptides/charges.

In more detail, a suitable test data set was generated as follows: One LC−MS/MS run from the iPRG-2015 data set and its internal peptides was considered at a time. All external ID sets originating from other runs were filtered to retain only IDs matching the internal peptides (using OpenMS' IDFilter tool with the "whitelist:peptides" option). The set of internal peptides was randomly split into two halves (using a custom Python script), and two corresponding, disjoint sets of internal IDs were generated (again using IDFilter). After these preparations, two FFId analyses were carried out, using each of the two halves of the internal IDs together with the filtered external IDs. (The FFId algorithm requires some overlap in the peptides of internal and external IDs to calculate an RT alignment between the two ID sets, so not all internal IDs could be replaced at the same time.) As a consequence, in each FFId result, features for one-half of the internal peptides were detected based on external IDs. To evaluate how well these features agreed with the gold standard, pairs with matching RT and m/z positions in both feature sets were detected. This was done using the OpenMS tool FeatureLinkerUnlabeledQT,[11] with very small tolerances (1 s in RT, 0.0001 Da in m/z) and the requirement of matching IDs during the grouping of features. To contrast the number of paired features, the number of possible pairs was determined based on which peptides and charge states yielded features in both the test and gold standard FFId analyses.

RESULTS AND DISCUSSION

Targeted Feature Detection

We developed an algorithm to detect peptide features in MS1 data in a targeted manner based on peptide identifications derived from MS2 spectra. The algorithm was developed within the OpenMS open-source framework for computational mass spectrometry. A corresponding executable tool, named "FeatureFinderIdentification" (FFId) in adherence to OpenMS naming conventions, is made freely available as part of OpenMS.

In the most basic case, our algorithm is used to detect features for peptide quantification in a single LC−MS/MS run. Figure 1 (bottom) gives a schematic overview of this. The inputs for the FFId tool then consist of the raw MS1 data (in mzML format) and a set of high-confidence peptide identifications (in OpenMS' idXML format) derived from the MS2 spectra. OpenMS supports a range of options for generating or importing suitable identification results. The goal of our algorithm is to detect features for all identified peptides in the high-confidence set to enable their quantification. Data processing happens in several phases, which we describe below.

Generation of "in Silico" Peptide Assays. In the context of targeted feature detection, assays encode the targets for which we attempt to detect features. Given the input set of high-confidence peptide identifications, we generate (at least) one assay per peptide and distinct charge state that was identified. To generate the assay or assays for a specific peptide, we consider all available IDs for that peptide. We first determine one or more RT ranges in which we expect the peptide to have eluted from the chromatographic column. (Because elution happens before peptides are ionized in the mass spectrometer, we consider IDs from all charge states together for a more complete picture.) Intuitively, the ID or IDs of a peptide indicates one or several time points at which the peptide was observed; however, we want to capture the whole time range over which the peptide eluted. We thus define a "window of interest" around the RT of every peptide ID based on a parameter for the expected width (in RT) of an elution peak and then merge overlapping windows. Typically, this results in one RT range per peptide that contains all corresponding IDs. However, multiple disjoint ranges can arise for some peptides and must be handled.

Next, we calculate the isotopic distribution, up to a configurable number N of isotopologues (default: $N = 2$), of the peptide based on its sequence and modifications, yielding the theoretical relative intensities of the respective isotopic peaks. For each charge state of the peptide, we then calculate the theoretical m/z values of those isotopic peaks.

Finally, on the basis of these intermediate results we generate one assay per charge state of the peptide and RT range provided that an ID for the charge was observed in that range. An assay contains N "MS1 transitions" for the isotopologues, each defined by the RT range and by the theoretical m/z value and relative abundance of the isotopic peak.

Chromatogram Extraction and Peak Group Detection. Processing the peptide assays in sequence, OpenSWATH[17] is used to extract chromatograms for the MS1 transitions in the assays. Each extracted ion chromatogram (XIC) is a rectangular

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

(RT by m/z) slice of the LC−MS data based on the RT range of the assay and a small window (e.g., 10 ppm) around the m/z value of the corresponding transition. The XIC is meant to capture the elution profile of ions of the corresponding isotopologue (among other ions of approximately the same m/z value).

The next step is to detect peak groups in the XICs of each assay. Again, OpenSWATH provides the algorithms for this. The process starts by detecting elution peaks in an individual XIC at a time. To this end, the signal in the XIC is smoothed using a Gaussian filter, then local maxima are detected and extended downward on both sides to find the limits of the peaks. Subsequently peaks with corresponding RT ranges are grouped across XICs in order of decreasing peak size. This produces groups of RT-aligned peaks in the XICs corresponding to isotopologues of a peptide. These peak groups are considered as feature candidates in the following steps. The sums of the peak areas in the groups give "raw" feature intensities that could already be used for quantification.

ID Mapping and Feature Selection. From the feature candidates (OpenSWATH peak groups) generated in the previous step, we now select one as "the" feature for each peptide and charge state. The best feature candidate is the one that is supported by the highest number of peptide IDs, and thus we map the peptide IDs of appropriate sequence and charge to the candidates based on retention time. The candidate with the highest number of matching IDs is kept, and others are removed; ties are broken according to higher feature intensity. We allow only one feature per peptide/charge even if there are multiple assays, based on multiple RT ranges, to ensure the robustness of the method across different samples. If there is no feature candidate with a matching ID, then we do not report a feature for the peptide/charge. This can happen if OpenSWATH fails to detect a suitable peak group in the data; however, this is a very rare occurrence, as our benchmark below will show.

Elution Model Fitting. We add a final processing phase that helps to increase the robustness of our feature intensity estimation. Instead of using "raw" feature intensities calculated by OpenSWATH for quantification, we apply a model-based approach, essentially smoothing out fluctuations in the measured data. Our model for peptide elution over time is a Gaussian function ("bell curve"), which we fit for each feature to the XIC data within the corresponding RT range. Data points from all MS1 transitions/isotopologues are used at the same time, weighted according to their theoretical relative intensities. A least-squares model fit is performed using the Levenberg−Marquardt algorithm. If successful, we set the feature intensity to the area under the curve, calculated from the model parameters. If model fitting does not converge or if the estimated model parameters fail consistency checks, then we revert to the "raw" OpenSWATH intensity estimate but adjust it to be on the same scale as the model-based intensities.

Combining Multiple LC−MS/MS Runs

In a shotgun experiment with multiple similar LC−MS/MS runs, we aim to detect features to enable quantification of the total set of peptides identified (with high confidence) across all runs. However, we still process the MS1 raw data one run at a time; only the peptide IDs are carried over across runs, giving rise to the notion of internal and external IDs (see Terminology section). We have already discussed how FFId detects features for the internal IDs of an LC−MS/MS run. There are two ways to incorporate external IDs in the analysis, which we present in the following.

Strategy 1: Merging Internal and External Peptide IDs. The simplest approach for combining information from multiple LC−MS/MS runs is to make no distinction between internal and external IDs. The sets of IDs can simply be merged (OpenMS offers the IDMerger tool for this purpose) and altogether treated as internal IDs, as described above. However, the drawback is that our assumptions about internal peptide IDs may not hold for external ones. We regard internal IDs as evidence for the detection of the respective peptide at a certain location in the RT-by-m/z space of the MS1 data. We rely on this information to pinpoint what we consider to be the correct feature for the peptide. Because external IDs come from a different LC−MS/MS run, we face two issues in particular: First, variations in the chromatography between different runs will affect the RT coordinates of external peptide IDs and potentially make them unreliable for feature selection. This can be addressed by computing a retention time alignment between different LC−MS/MS runs and adjusting the RT scales accordingly. The OpenMS tool MapAlignerIdentification[11] is ideally suited for this because it directly uses the RT values of peptide IDs to calculate the alignment and it enables nonlinear correction of chromatographic differences. MapAlignerIdentification and similar solutions will generally work well to correct monotonic deviations between chromatographic runs; however, if the chromatography suffers from low reproducibility, for example, if the order in which peptides elute changes, then RT differences between runs may become impossible to resolve. To increase the peptide coverage while minimizing the potential for unreliable RT values, we can include external IDs only for peptides that are not already represented among the internal IDs. This is supported in OpenMS by the "add_to" option in the IDMerger tool.

The second issue with treating external IDs as internal is overimputation: If different biological samples are analyzed, then a peptide that was observed in one sample may not necessarily be present in another. An external ID for such a peptide could lead to the detection of a false-positive feature if a corresponding peak group is found by OpenSWATH. (However, provided that the RT value of the external peptide ID is reliable, this is due to a general limitation of our targeted feature detection approach: We have no way of distinguishing "true" signals from interference caused by unrelated ions if they overlap in RT and m/z.)

These caveats notwithstanding, aligning and merging internal and external IDs can be a very effective way of combining data from multiple runs. This is especially true if the samples in question are highly similar and the chromatography is stable, for example, if a low number of runs are analyzed in direct succession on the same column. In such cases the ability to rely on the RT coordinates of peptide IDs to determine the "correct" feature among a set of candidates is difficult to surpass with other approaches.

Strategy 2: Machine Learning. FFId offers an alternative, novel mode of operation in which features for external IDs can be detected without depending on the accuracy of the external RT values. The basic idea is to use internal IDs to generate a training data set for a classifier, which can then be applied to predict the correct features for external IDs. This approach is similar to that used by Percolator[7] and mProphet[23] for the statistical validation of PSMs and SRM results, respectively. An overview is shown in Figure 2.

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

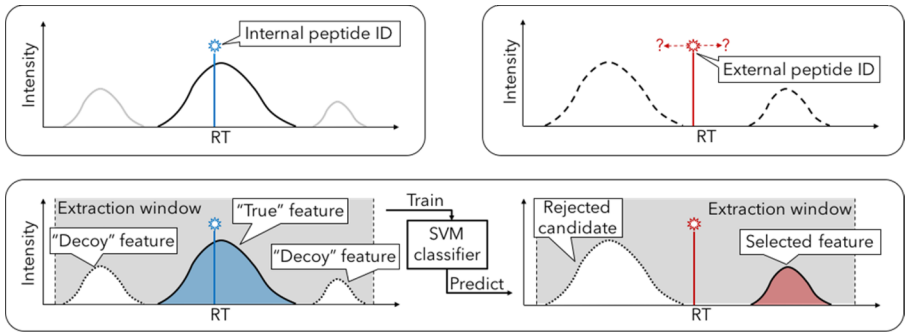


Figure 2. Overview of using machine learning to combine internal and external peptide IDs for feature detection. Each subplot shows a simplified view of an ion chromatogram, containing multiple elution peaks, for the mass-to-charge value that corresponds to the peptide ID of interest. Top left: Given an internal peptide ID, selecting the correct peak (feature) that allows quantification of the ID is simple because the retention time (RT) value of the ID can be relied upon. Top right: For an external peptide ID, the exact RT, relative to the "internal" chromatographic time scale, is uncertain. In addition, because the ID is only inferred, the peptide is not necessarily present in the sample, so there might not be a correct peak. Bottom: Our solution is to consider feature candidates in an RT window around the locations of internal and external peptide IDs. For internal IDs the correct candidates are known. This gives rise to a data set of positive and negative examples ("true" and "decoy" features), with which we train a support vector machine (SVM) classifier. The classifier predicts correctness probabilities for feature candidates, which we use to score and select features for external peptide IDs.

In more detail, the process works as follows: Before analyzing any particular LC–MS/MS run with FFId, the external IDs from all related runs should be aligned to the internal IDs (e.g., using MapAlignerIdentification), just like in "Strategy 1" explained above. However, now only the aligned sets of external IDs are merged, and internal and external IDs are passed separately to the FFId program. Internally, FFId realigns the RT scales of both sets of IDs to estimate how well they agree. The algorithm then aggregates all IDs and uses them, without discriminating between internal and external ones, to generate MS1-level assays for peptides. The procedure is the same as described in the Generation of "in Silico" Peptide Assays section, with only one difference: The RT region around each peptide ID is increased to account for uncertainty in the RT coordinates of external IDs based on the realignment. Chromatogram extraction and peak group detection then proceed as previously described. Importantly, during the detection step, OpenSWATH calculates a variety of quality scores for the peak groups, measuring, for example, signal-to-noise ratio, m/z deviation between transitions and XICs, overlap between peaks in different XICs, and agreement of peak intensities with theoretical isotopologue intensities. FFId adds a special RT deviation score, which measures the distance between a feature candidate and the closest external ID of the matching peptide. (If the assay is based only on internal IDs, then their RTs are transformed to the external scale using the alignment computed in the beginning.) This RT deviation score allows us to capture RT information without fully relying on the RT values of IDs for feature selection. We thus obtain a selection of roughly 15 scores for each feature candidate (OpenSWATH peak group), which will serve as predictor variables for our classifier.

In the feature selection phase, we generate a training data set of positive and negative instances. Positive instances are feature candidates that we consider as "correct" based on matching internal IDs (see ID Mapping and Feature Selection section). Negative instances (decoys) are feature candidates without any matches but where internal IDs are available in the assay; in such cases we can be confident that the "correct" feature must be a different one. (Candidates with some but not the highest number of matching IDs for the peptide/charge are considered as ambiguous and are not used for training.) Because the positive and negative instances are selected using internal IDs but the classifier will be applied to feature candidates of external IDs, there is a risk of introducing biases in the training data. This must be avoided to ensure maximum classification performance. The detection of feature candidates works in exactly the same way for peptides with internal IDs, external IDs, or a mixture of both; however, a potential biasing factor is MS1 signal intensity: DDA preferentially triggers MS2 acquisition for precursors of high intensity, so feature candidates in our positive set, supported by internal IDs derived from MS2, will tend to have high intensities. However, we are interested in predictions for peptides *without* internal IDs, which conversely may generate features of lower intensity. To avoid an intensity bias, we select positive and negative instances with the same intensity distribution, in roughly equal numbers, for the training data set. This whole approach allows us to generate an optimal classifier for the data set at hand, in which the weights for different predictors are chosen according to the characteristics of the data. For example, the influence of the RT deviation score can be larger or smaller depending on the reliability of the RT alignment.

Our classification algorithm of choice is the support vector machine (SVM), a robust and versatile machine learning technique.[24] We utilize the SVM implementation in the LIBSVM C library,[18] for which we created a C++ wrapper class for use in FFId. Given the data set of feature candidate scores and positive/negative labels for the training instances, we train an SVM classifier according to the guidelines provided by the LIBSVM authors.[25] To this end, we first scale each predictor to the range 0 to 1. Next, we convert the data into the sparse representation required by LIBSVM. We then set up an SVM model, by default with a radial basis function (RBF) kernel. The parameters of the model (the misclassification cost, C, and the kernel bandwidth, γ) must be optimized; we do this by performing a grid search over predefined sets of parameter values and running a cross-validation for each parameter combination. After evaluating the cross-validation performances we select the best parameters, then retrain the model on the whole training data set. Subsequently we apply this final model to all feature candidates, predicting the probability of each being the "correct" feature for its peptide and charge.

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964–2974

After the SVM scoring we select which feature candidates to retain (followed by elution model fitting as described above). For peptides/charges with only external IDs, we select the candidate with the highest SVM probability. For peptides/charges with internal IDs, we apply the same ID-based selection as before; that is, we keep the positive instances from the training data set. However, in these cases we also note which feature candidate was scored highest by the SVM, and we record whether this candidate is the correct one according to the "gold-standard" ID-based selection. In the end, the overall fraction of incorrect cases, where we would have chosen the wrong candidate if we had relied on the SVM score, gives an estimate of the FDR for the SVM-based feature selection. However, selection based on the SVM score is only applied to peptide assays derived entirely from external IDs, and thus the FDR estimate has to be scaled according to the fraction of such assays in the data set to arrive at a meaningful FDR value for the whole FFId analysis.

Probabilistic scoring and FDR estimation for "inferred" features enable us to overcome one of the inherent difficulties in combining data from multiple LC−MS/MS runs: By filtering to a specific FDR threshold we can control the risk of false-positive features among our results for cases where peptides referenced by external IDs are not actually present in the sample under consideration. Feature candidates that are not scored highly enough by the SVM classifier can be excluded, thus allowing us to make an informed decision to *not* quantify certain peptides.

Evaluation

We applied our FFId algorithm to the iPRG-2015 data set,[21] comparing the three strategies described above: First, we used only the internal IDs from each run as targets for quantification ("internal IDs only"). Second, we used both internal and external IDs together after aligning and merging them ("internal/external IDs, merged"). Third, we used internal and external IDs together in the machine learning approach ("internal/external IDs, machine learning"). The results of these analyses are available in the PRIDE repository[26] under accession PXD006336. A workflow for the KNIME Analytics Platform (freely available at www.knime.org) that reproduces the FFId analyses can be downloaded from www.openms.de/workflow/targeted-feature-detection-for-lfq.

Usability and Runtime. The FFId algorithm is very easy to use. While a variety of parameters can be tuned for expert usage, there is only one main parameter that may have to be adapted on a per-data set basis—the expected chromatographic peak width. The appropriate value can be estimated by visual inspection of the LC−MS raw data, for example, in the OpenMS viewer, TOPPView.[27] Furthermore, FFId is robust to reasonable variation in the choice of this parameter (for example, using 60 s instead of 40 would give very similar results for the benchmark data set used here).

FFId is compatible with any peptide identification pipeline that can produce one of the many ID file formats supported by OpenMS (e.g., mzIdentML, pepXML, idXML). In our benchmark analysis, we have combined FFId with other OpenMS tools (MapAlignerIdentification, ProteinQuantifier) and with the R package MSstats to create a powerful data processing pipeline for protein quantification. However, alternative solutions for retention time alignment, normalization, or protein-level quantification could be used just as well.

To date we did not invest much time into optimizing the runtime of the FFId algorithm. A full analysis of the iPRG-2015 data set, using the machine learning approach for combining internal and external IDs and processing all 12 LC−MS/MS runs in parallel on a single server, was completed in under 8 h. In this time 17 GB of mzML files was processed, >337 000 peptide IDs were considered in each LC−MS/MS run, and ∼44 300 features were detected per run (over 532 000 in total). The especially thorough data analysis performed by FFId, combined with the unbiased way of handling internal and external peptide IDs that enables our use of machine learning, implies a significant overhead in data processing that accounts for much of the runtime. As such, on average 2.3 feature candidates were detected and scored for every feature in the final results. Using six parameter combinations (a restricted set based on prior knowledge) and 5-fold cross-validation, 30 SVM models were trained and evaluated per LC−MS/MS run to find the optimal SVM parameters.

Quantification Coverage. Our algorithm achieved exceptional quantification coverage with all three strategies (see Figure 3). In the first instance, targeting internal IDs only,

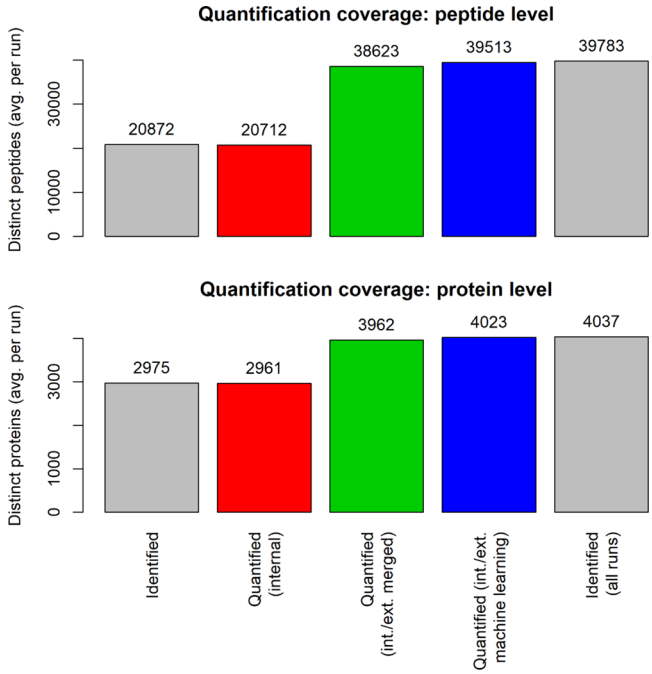


Figure 3. Top: Comparison of the numbers of (modified) peptides identified with high confidence (1% FDR, gray bars) in the iPRG-2015 data set and the numbers of quantified peptides using FFId with different data analysis strategies (red/green/blue bars). Bottom: Analogous comparison for identified proteins. For simplicity, only proteins identified by uniquely matching (proteotypic) peptides were counted.

>99% of the (modified, high-confidence) peptides identified in each LC−MS/MS run could be quantified in that run. Second, after aligning and merging internal and external peptide IDs, >95% of the high-confidence peptides identified over *all* runs could be quantified in each run. Finally, using the machine learning approach to combine internal and external IDs, >99% of the full set of high-confidence peptides was quantified in each LC−MS/MS run.

The same relations hold for the numbers of identified and quantified proteins (Figure 3, bottom); however, as expected

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

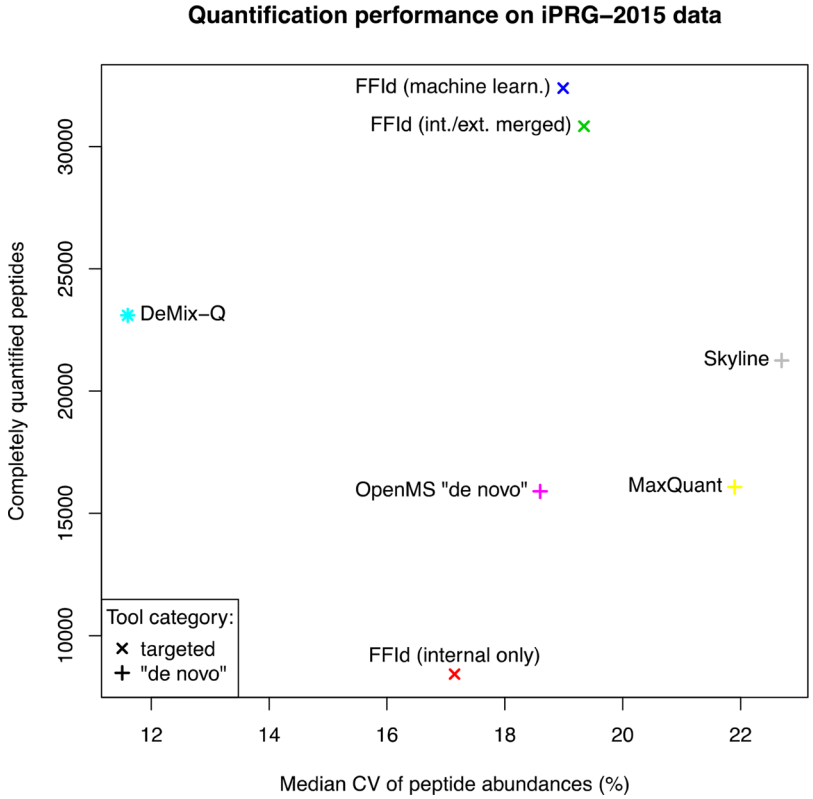


Figure 4. Quantification performance of different software tools for label-free quantification on the iPRG-2015 data set. We compare the median coefficient of variation (CV) for peptide abundances and the number of peptides (ignoring modifications) quantified in all 12 runs. Data for DeMix-Q, OpenMS "de novo", MaxQuant, and Skyline was taken from Zhang et al.[14]

there is a smaller gap on the protein level between (a) results on individual LC−MS/MS runs and (b) results obtained by integrating data from multiple runs. Compared with a single LC−MS/MS run, the number of identified peptides almost doubled when all runs were taken into account, while the number of identified proteins increased by only about a third. The improvements in coverage achieved by the two "integrative" FFId approaches are thus less dramatic on the protein level. However, beyond the number of quantifiable proteins, a more exhaustive quantification of peptides also has the potential to improve the *quality* of quantification on the protein level (see Protein Quantification Accuracy section).

Reproducibility and Completeness of Quantification. To assess the reproducibility of quantification across replicate measurements, we calculated coefficients of variation (CVs) for peptide abundances, as described in the Methods section. Because most peptides come from constant background proteins, lower CVs are better. The results are shown in Figure 4. Our method achieved median CVs of 17.1% ("internal IDs only"), 19.3% ("internal/external IDs, merged"), and 19.0% ("internal/external IDs, machine learning") using the three different strategies for incorporating external peptide IDs. These values are worse than the 11.6% median CV reported for DeMix-Q, but they are on par with the results for the OpenMS "de novo" approach (18.6%), MaxQuant (21.9%), and Skyline (22.7%) (Bo Zhang, personal communication, 11 January 2017; see also Figure 6 in the DeMix-Q paper[14]). Note that the DeMix-Q workflow performs additional steps besides the feature detection to minimize deviations, for example, quality scoring that includes a penalty term for intensity variation, imputation of missing values, and intensity recalibration.

Using only internal IDs for quantification, FFId reached the lowest number of fully quantified peptides, 8426, among the tools compared. As the only approach in the comparison that did not infer IDs across runs, it was limited by the number of peptides that were consistently identified in all 12 runs. However, using the two strategies that combine internal and external IDs, FFId on the DeMix search results provided complete quantification for far more peptides than any other software in this comparison: Out of a total of 33 198 distinct peptides (ignoring modifications) that were identified, 30 832 ("merging" strategy) and 32 387 ("machine learning" strategy) were quantified without any missing values—33 and 40% more than for DeMix-Q, respectively.

When peptide abundances from technical replicates were averaged, median CVs for peptides with complete quantification across the four different samples, this time excluding spike-in proteins, dropped to 11.4% ("internal IDs only"), 10.8% ("internal/external IDs, merged") and 10.4% ("internal/external IDs, machine learning") compared with 6.0% for DeMix-Q. At the same time the numbers of fully quantified peptides rose to 19 204, 32 756, and 33 153 for the three FFId strategies. Significantly, once technical replicates were aggregated, the additional coverage gained by including external IDs reduced variation below the level of quantifying only internal IDs, while increasing the number of completely quantified peptides by 70%.

Protein Quantification Accuracy. The iPRG-2015 data set enables us to evaluate the quantification accuracy of our method by comparing abundance estimates to the known amounts of the six spike-in proteins. To this end, we performed protein quantification using MSstats[20] on the basis of FFId

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

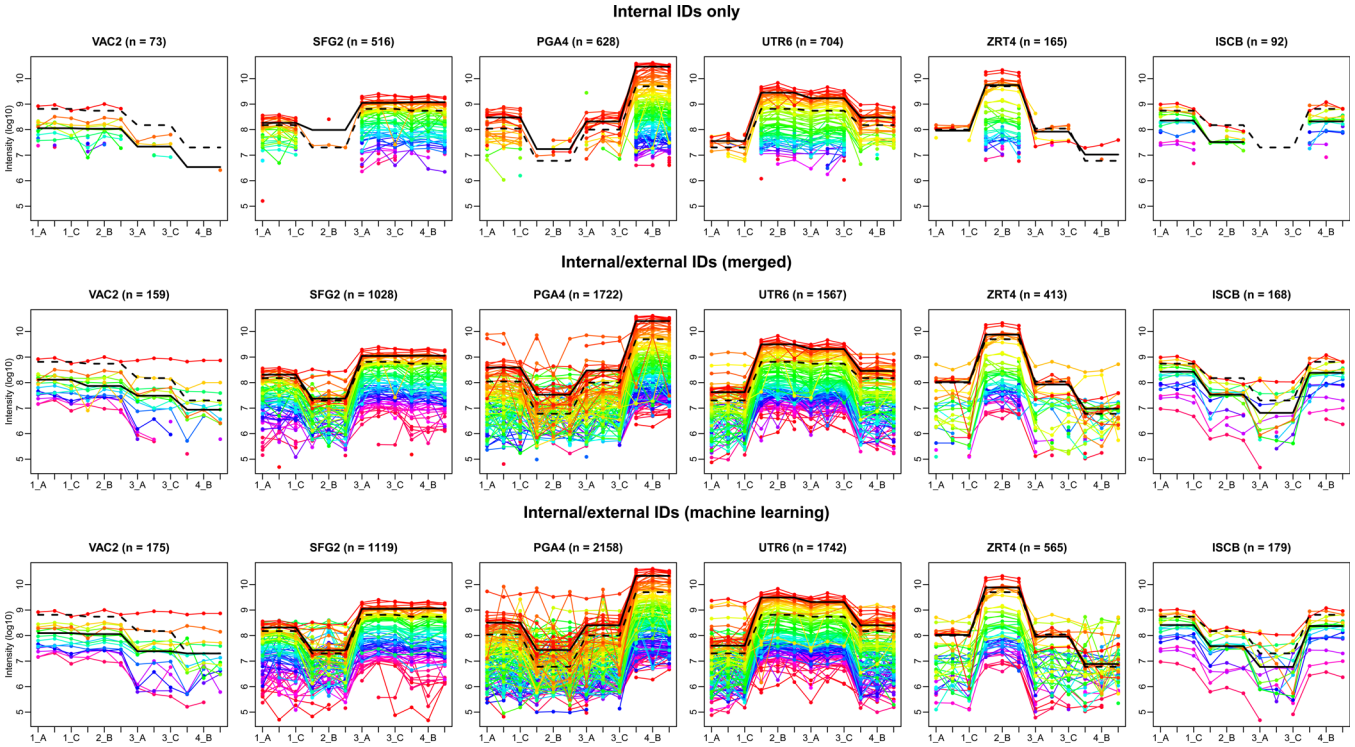


Figure 5. Peptides of spike-in proteins from the iPRG-2015 data set quantified using FFId. Each subplot corresponds to one protein and one quantification strategy, with the 12 LC−MS/MS runs on the *x* axis and intensity (log-scale) on the *y* axis. Each colorful line in a subplot shows one quantified peptide of the corresponding protein (the "*n* = ..." in the subplot title gives their number). The solid black line represents the protein-level quantification result calculated by MSstats. For reference, the dashed black line indicates the true amount of the spiked-in protein (value in fmol multiplied by 10^7). The scaling of the true protein amount is arbitrary and has been done for visualization purposes. Because we do not aim for absolute protein quantification, constant differences between true and estimated protein amounts can be ignored.

feature detection results, as described in the Materials and Methods section. Figure 5 shows the quantification results for the six spike-in proteins and their peptides. Comparing the dashed and solid black lines, we can appreciate that the estimated (relative) protein abundances closely match the trends of the true amounts. We can also observe that many values for peptides at lower concentrations are missing in the "internal IDs only" strategy (due to lower intensity precursors not being selected for MS2) and that this deficiency is overcome by the strategies for merging internal and external peptide IDs. On the basis of the high degree of completeness that these strategies achieve for quantification on the peptide level, interesting avenues open up for optimizing the selection of peptides that are used for protein quantification. MSstats already contains an option for this ("dataProcess(..., featureSubset = "highQuality")"), which unfortunately was not yet supported in the version we used.

To look at the accuracy of relative protein quantification in a more quantitative fashion, we calculated the fold changes in protein abundance between any two of the four samples in the data set. The results for all three strategies of handling external peptide IDs are shown in Figure 6, comparing expected to estimated fold changes. All strategies achieved very good agreement between true values and estimates, with Pearson correlations around 0.98. For comparison, the correlation for DeMix-Q results was 0.96.[14]

Classification Performance and FDR Estimation. When FFId analyses are performed with the "machine learning" strategy for combining internal and external peptide IDs, an important consideration is whether the classification approach succeeds in selecting the correct features for the external IDs. This is a necessary requirement for accurate quantification, albeit not sufficient (e.g., the intensity estimation of features has to be accurate as well). One indicator for reliability is the prediction accuracy (fraction of correctly classified instances) of the SVM model, which is determined by cross-validation on the training data set during the parameter optimization of the classifier. In our analysis of the 12 iPRG-2015 samples, we performed five-fold cross-validation and found the prediction accuracy to be 95% on average (range: 94.4 to 95.7%).

When internal and external peptide IDs are combined using the "machine learning" approach, FFId calculates an FDR estimate for the feature selection, that is, a measure for the fraction of cases in which the wrong feature candidate was selected for a peptide assay. In our analysis of the benchmark data set, we estimated an average FDR of 3% (range: 2.7 to 3.3%) for features derived from external IDs. Those external features accounted for roughly 48% of all features detected per LC−MS/MS run. The remaining 52% were features supported by internal IDs, which we always consider as "correctly selected" as per our definition. Consequently, FDRs for the full feature set per LC−MS/MS run were estimated to be 1.5% on average (range: 1.3 to 1.7%).

To validate our FDR estimates, we performed an experiment trying to answer the question, "If only external IDs are available for a given peptide and charge, do we still detect the same feature that we would detect given internal IDs?" Thus features detected for internal IDs were considered as the "gold standard", against which features detected for external IDs of the same peptides were compared (see Feature Selection for

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

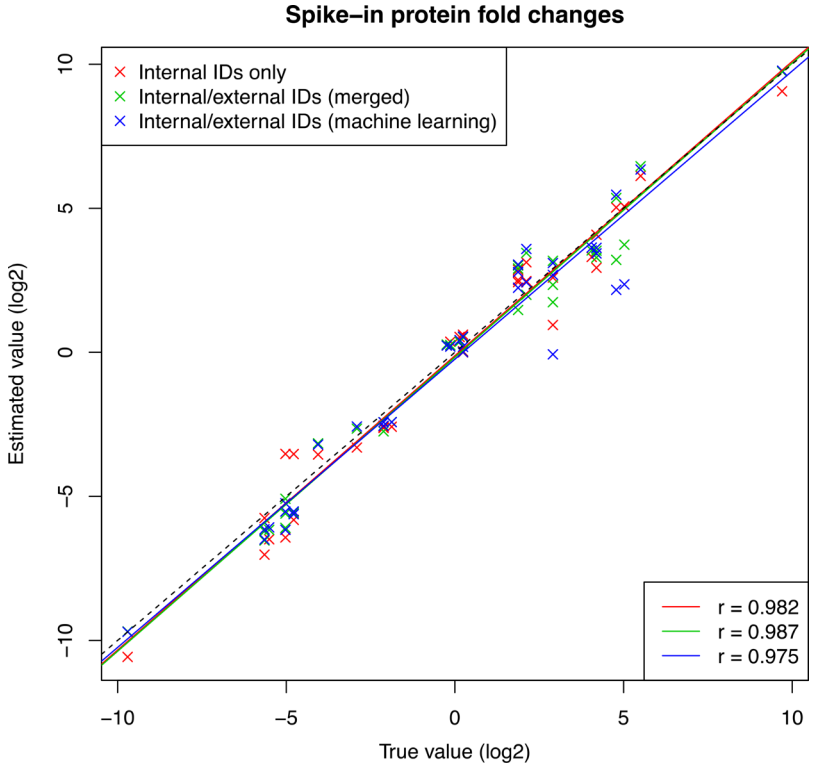


Figure 6. Comparison of true and estimated fold-changes of spike-in proteins from the iPRG-2015 data set. Each data point ("×") marks a fold-change between two samples for one of the six spike-in proteins, color-coded according to the FFId quantification strategy used. The solid lines show the corresponding least-squares linear fits; the dashed line is the diagonal. The box on the bottom right gives the Pearson correlations between true and estimated fold changes for the three quantification strategies.

External IDs section for details). The estimated FDRs for the full feature sets in the corresponding FFId analyses averaged 1.6% (range: 1.4 to 1.8%). The actual FDRs, calculated as the fraction of external features that did not match the "gold standard" internal features, were found to be slightly lower, averaging 1.4% (range: 1.2 to 1.8%). This shows that our FDR estimates are very close to the true values and in most cases are slightly conservative. However, as a limitation of this analysis, only peptides with internal IDs could be considered, which excludes many peptides of low abundance.

CONCLUSIONS

We have presented a novel feature detection algorithm that realizes the core data analysis step for the quantification of peptides and proteins in label-free shotgun LC−MS/MS experiments. Shotgun proteomics based on DDA is a very mature technology and the method of choice for unbiased, discovery-driven studies of the proteome. For quantitative proteomics, a label-free shotgun experiment provides the simplest possible experimental workflow. However, until recently the full potential of this approach could not be realized due to limitations of the algorithms used for data analysis; they were marred by a tendency to generate abundance data sets with substantial fractions of missing values. Now, in part inspired by advances in computational methods for processing SRM and DIA data, there is a new generation of software tools for label-free quantification, including our own FeatureFinderIdentification (FFId). Although these tools cannot provide the increased sensitivity of SRM and DIA (achieved by detecting and quantifying peptides on the fragment ion level),[16] by implementing targeted approaches, they overcome the persistent "missing value problem" that has limited the appeal of label-free methods. FFId, in particular, provides outstanding quantification coverage combined with good reproducibility and accuracy, as we have shown in our analysis of a benchmark data set. This superior performance derives in part from our use of machine learning to combine information from multiple LC−MS/MS runs for feature scoring and selection. Altogether, our method shows that reliable and highly sensitive protein quantification based on label-free shotgun data is possible.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jc4@sanger.ac.uk. Tel: +44 1223 494986.

ORCID

Hendrik Weisser: 0000-0001-9723-7503

Present Address

†H.W.: Storm Therapeutics, Ltd., Cambridge, United Kingdom.

Author Contributions

H.W. developed the software, performed the data analysis, and wrote the manuscript. J.S.C. supervised the project and contributed to the manuscript.

Notes

The authors declare no competing financial interest.

The results of the analyses are available in the PRIDE repository[26] under accession PXD006336.

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974

■ ACKNOWLEDGMENTS

We thank Bo Zhang for providing the DeMix search results used in our analysis and for explaining details of the DeMix-Q benchmark. We also thank Lu Yu for generating an LC−MS/MS data set used in the development of our algorithm, Hannes Röst for useful discussions and help with OpenSWATH, and the OpenMS developer community for their work on that project. We gratefully acknowledge funding from the Wellcome Trust (grant WT098051).

■ REFERENCES

(1) Aebersold, R.; Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **2016**, *537*, 347−355.
(2) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.
(3) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.
(4) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7* (8), 3354−3363.
(5) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655−667.
(6) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.
(7) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923−925.
(8) Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; Del-Toro, N.; Rurik, M.; Walzer, M. W.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcaíno, J. A. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* **2016**, *13* (8), 651−656.
(9) Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.-Y.; Vitek, O.; Aebersold, R.; Müller, M. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **2007**, *7*, 3470−3480.
(10) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, *13*, 2513−2526.
(11) Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; Brauer, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; Malmström, L. An automated pipeline for high-throughput label-free quantitative proteomics. *J. Proteome Res.* **2013**, *12* (4), 1628−1644.
(12) Schilling, B.; Rardin, M. J.; MacLean, B. X.; Zawadzka, A. M.; Frewen, B. E.; Cusack, M. P.; Sorensen, D. J.; Bereman, M. S.; Jing, E.; Wu, C. C.; Verdin, E.; Kahn, C. R.; Maccoss, M. J.; Gibson, B. W. Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in Skyline: application to protein acetylation and phosphorylation. *Mol. Cell. Proteomics* **2012**, *11*, 202−214.
(13) Argentini, A.; Goeminne, L. J. E.; Verheggen, K.; Hulstaert, N.; Staes, A.; Clement, L.; Martens, L. moFF: a robust and automated approach to extract peptide ion intensities. *Nat. Methods* **2016**, *13*, 964−966.
(14) Zhang, B.; Käll, L.; Zubarev, R. A. DeMix-Q: Quantification-Centered Data Processing Workflow. *Mol. Cell. Proteomics* **2016**, *15*, 1467−1478.
(15) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **2016**, *13* (9), 741−748.
(16) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **2012**, *11*, O111.016717.
(17) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; Aebersold, R. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **2014**, *32*, 219−223.
(18) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (27), 1−27:27.
(19) R Core Team. *R: a Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2016. https://www.R-project.org.
(20) Choi, M.; Chang, C.-Y.; Clough, T.; Broudy, D.; Killeen, T.; MacLean, B.; Vitek, O. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **2014**, *30*, 2524−2526.
(21) Association of Biomolecular Resource Facilities. *iPRG-2015 Proteomics Informatics Research Group Study: Differential Abundance Analysis in Label-Free Quantitative Proteomics*; ABRF: Bethesda, MD, 2015. https://abrf.org/research-group/proteome-informatics-research-group-iprg.
(22) Zhang, B.; Pirmoradian, M.; Chernobrovkin, A.; Zubarev, R. A. DeMix workflow for efficient identification of cofragmented peptides in high resolution data-dependent tandem mass spectrometry. *Mol. Cell. Proteomics* **2014**, *13*, 3211−3223.
(23) Reiter, L.; Rinner, O.; Picotti, P.; Hüttenhain, R.; Beck, M.; Brusniak, M.-Y.; Hengartner, M. O.; Aebersold, R. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods* **2011**, *8*, 430−435.
(24) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20* (3), 273−297.
(25) Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; National Taiwan University: Taipei, Taiwan, 2016. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
(26) Vizcaíno, J. A.; Csordas, A.; del Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.-W.; Wang, R.; Hermjakob, H. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44* (D1), D447−D456.
(27) Sturm, M.; Kohlbacher, O. TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.* **2009**, *8*, 3760−3763.

DOI: 10.1021/acs.jproteome.7b00248
J. Proteome Res. 2017, 16, 2964−2974