Search Databases and Statistics: Pitfalls and Best Practices in Phosphoproteomics

Jan C. Refsgaard, Stephanie Munk, and Lars J. Jensen

Abstract

Advances in mass spectrometric instrumentation in the past 15 years have resulted in an explosion in the raw data yield from typical phosphoproteomics workflows. This poses the challenge of confidently identifying peptide sequences, localizing phosphosites to proteins and quantifying these from the vast amounts of raw data. This task is tackled by computational tools implementing algorithms that match the experimental data to databases, providing the user with lists for downstream analysis. Several platforms for such automated interpretation of mass spectrometric data have been developed, each having strengths and weaknesses that must be considered for the individual needs. These are reviewed in this chapter. Equally critical for generating highly confident output datasets is the application of sound statistical criteria to limit the inclusion of incorrect peptide identifications from database searches. Additionally, careful filtering and use of appropriate statistical tests on the output datasets affects the quality of all downstream analyses and interpretation of the data. Our considerations and general practices on these aspects of phosphoproteomics data processing are presented here.

Key words Phosphoproteomics, Database Search, False Discovery Rate, Statistics, Quantitation, MaxQuant

1 Introduction

Virtually all cellular processes are regulated by posttranslational modifications (PTMs). Phosphorylation is a crucial and highly dynamic PTM that contributes to cellular physiology and pathophysiological developments [1]. Phosphoproteomics platforms are generating an ever increasing amount of data. For the experimentalist the challenge lies in transforming these vast amounts of information in the acquired MS and MS/MS scans into quantified phosphorylation sites (phosphosites) mapped to identified proteins.

There are several approaches for assigning peptide sequences to ions sequenced by mass spectrometry (MS). De novo sequencing reads out the peptide sequence from the mass differences of the ions detected in the MS/MS scan [2, 3]. In the early years of the mass

Louise von Stechow (ed.), *Phospho-Proteomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1355, DOI 10.1007/978-1-4939-3049-4_22, © Springer Science+Business Media New York 2016

spectrometry era, this daunting task was performed manually, but it is now automated. An alternative to de novo approaches is the spectral library approach, which compares each MS/MS spectrum to a reference library of previously observed MS/MS spectra [4]. Another approach is the database search strategy, which implements theoretical spectra from an in-silico digested database of all proteins in the species of interest [5]. The acquired experimental spectra can be compared to those theoretical spectra to infer and score peptide spectral matches (PSMs). The latter approach is the most widely used in phosphoproteomics analyses. This strategy has also been extended to mapping and scoring of phosphosites. As such, for every potentially phosphorylated peptide, a theoretical tandem spectrum is generated for each possibly phosphorylated version of the peptide (corresponding to each serine, threonine and tyrosine in the peptide). The platforms of phosphorylation site identification reviewed in this chapter implement this latter strategy.

There are three independent steps in the processing of raw MS data into quantified phosphosite ratios:

1. *Database search:* Raw spectra are searched against a reference peptide database, and a score is calculated for each PSM.
2. *Filtering:* In this step the PSMs are sorted and filtered down to a target False Discovery Rate (FDR) to limit false positive identifications.
3. *Quantitation:* Finally ratios are calculated for all peptides (and proteins).

This chapter gives an overview of all the important considerations which should be taken into account when processing raw data to retrieve a quantified phosphoproteomics output. The focus is on understanding the biases that are inherent to such data so that common pitfalls can be avoided.

2 MS Data Formats

Most vendors of mass spectrometric instrumentation have their own MS raw file format, and generally also provide a platform to process this raw data into quantified data. However, if the user opts for software that is unable to parse these formats, conversion will be necessary. Phosphoproteomics data generally contains two types of information: (1) full MS spectra and (2) MS/MS (tandem MS) spectra. When converting from vendor raw data format, it is important to bear in mind the information retained in the new file format. As such, the popular format MGF (Mascot Generic File) does not contain full MS information, and it therefore cannot be used for quantification based on metabolic labeling [6, 7]. For converting between MS file formats we recommend ProteoWizard MSConvert [8] or TOPPAS [9] FileConverter workflow.

ProteoWizard MSConvert [8] is available for Windows, Linux, and OS X. All versions of the software can read and write the following open formats: mzML, mzXML, MGF, ms2/cms2/bms2, and mzIdentML. The Windows version can furthermore read the following vendor formats: Agilent, Bruker FID/YEP/BAF, Thermo RAW, and Waters RAW. MSConvert has both a command line and graphical user interface, rendering it versatile and user friendly.

TOPPAS FileConverter [9] is available for Windows, Linux, and OS X, and it provides a graphical user interface to the command line tool FileConverter, which is part of TOPP (The OpenMS Proteomics Pipeline) [10, 11]. It can convert the following input file formats: mzData, mzXML, mzML, dta, dta2d, MGF, featureXML, consensusXML, ms2, fid, tsv, peplist, kroenik, and edta into mzData, mzXML, mzML, dta2d, MGF, featureXML, consensusXML, and edta format.

In our opinion the best open formats are mzML and mzXML. mzML is the de facto standard that was developed by the HUPO (Human Proteome Organization) initiative to unify the mzXML and mzData formats.

For more in-depth overview of all proteomics file formats the reader is referred to Deutsch [12].

3 Database Search

A wealth of different database search algorithms has been developed over the years. The most popular include the open source engines X!Tandem [13], OMSSA [14], MyriMatch [15], the freeware engine Andromeda [16], and the proprietary engines SEQUEST [17], PEAKS DB [3], and MASCOT [18]. All these algorithms are very mature and should produce comparable results. However, because they all use slightly different and to some extent orthogonal scoring schemes, using two in combination often yield higher confidence identifications [19].

4 Filtering

This step of the phosphosite identification workflow aims to exclude low-confidence identifications resulting from the database search. While more stringent filtering will result in a sacrifice of identifications of the total number of peptides and phosphosites, the resulting identifications are more trustworthy. This is in particular advantageous for the experimentalists using this information for hypothesis generation and downstream experiments.

There are two commonly used approaches to filter MS data, based on (1) arbitrary PSM score cutoffs and (2) FDR. The PSM score cutoff strategy depends on the size of the search database.

Larger databases will inherently yield more false positive identifications at a given cutoff. Therefore higher PSM score cutoffs are required when dealing with phosphoproteomics in comparison to proteomes due to the differences in search space. Therefore we recommend the FDR approach.

4.1 False Discovery Rate

The advantage of the FDR approach over the cutoff approach is that a given FDR is comparable across datasets, irrespective of the size of the dataset and search space. FDR is calculated using a target–decoy database approach, in which the PSMs are performed against a database of interest and a fictive database of comparable size. The most widely used decoy approach is the pseudo-reversed method proposed by Elias and Gygi [19], in which they dubbed the decoy database “reverse” and the original search database the “forward.” Pseudo-reversed is a quite fitting name, as the tryptically digested peptides are literally reversed, except for the last R/K, which is swapped. Searching against a concatenated forward and reverse database, the FDR can be calculated simply by counting the number of forward and reverse peptides/proteins above a given score cutoff. For most search engine platforms, the desired FDR is set in advance, and all peptides above the corresponding score are accepted. Additionally, most platforms allow the user to set the FDR on both peptide and protein identifications. Setting the FDR on the protein identification level will almost always result in a more stringent FDR of the peptide identifications, as multiple forward hits match the same protein. For phosphoproteomics data, however, it is recommendable to set the FDR on the peptide level, as the resulting phosphosites are identified on the peptide level and ultimately the data of interest. A 1 % FDR is commonly accepted in the phosphoproteomics community.

4.2 Phosphosite Localization Probability

The ultimate aim of phosphoproteomics investigations is to identify the exact location of the phosphorylation moiety in the sequenced phosphopeptide, and the predicted protein of origin. While no successful approach has been developed for implementing false localization rates at the phosphorylation site level, applying stringent FDR at the peptide level is beneficial in the processing of phosphoproteomics data. Peptides that pass at a higher FDR cutoff are generally identified from MS/MS scans with more peaks, thereby increasing the confidence of the phosphorylation site localization. This is illustrated in Fig. 1 where the median localization probability drops below 100 % at peptide FDR = 0.2 % (*see* Fig. 1). In addition to a stringent FDR at the peptide level, it is common practice to filter all resulting phosphosites such that none has a localization probability below 75 % [20]. While 75 % may seem very low, it is important to note that most sites can be localized with much higher accuracy, some even with 100 %

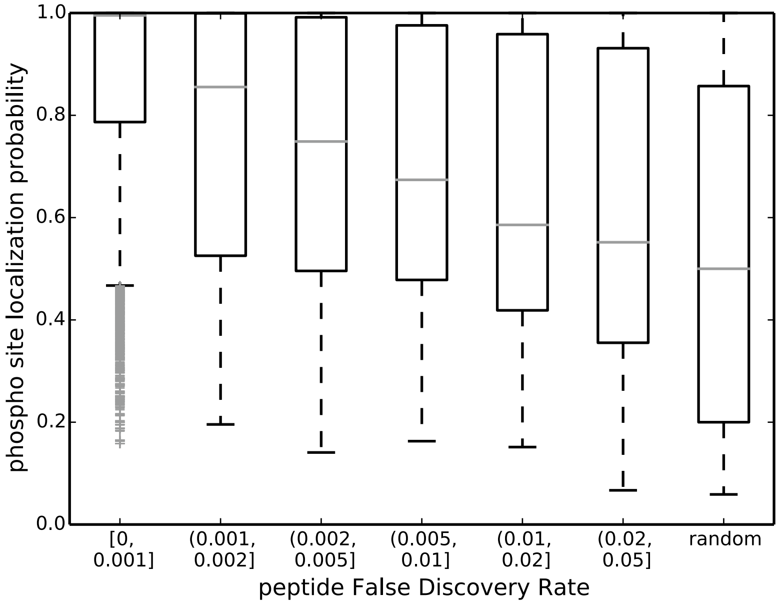


Fig. 1 Boxplots depicting the relationship between the phosphorylation localization probability and the False Discovery rate

accuracy if the peptide contains only one serine, threonine or tyrosine. The average localization probability at 1 % peptide FDR was 96.6 % for the data used to create Fig. 1.

5 Quantitation

Phosphoproteomics experiments often seek to determine the differences in the abundance of phosphosites between perturbations, tissues or other conditions, as these changes are vital to answering the biological question at hand. Quantitation strategies are either based on label-free approaches that do not require alterations to the experimental workflow, or on isotopic labeling of amino acids. Label-free quantitation (LFQ) is based on comparison of full MS scans from separate MS raw files representing the conditions of interest. Recent innovations in LFQ such as iBAQ [21] combined with improvements of MS instrumentations have rendered LFQ strategies a viable alternative to isotopic labeling. While LFQ gives hopes for clinical proteomics, labeling-based strategies remain the preferred tool in quantitative phosphoproteomics. The two most popular labeling techniques are stable isotope labeling with amino acids in cell culture (SILAC) [6, 7] and isobaric tags for relative and absolute quantitation (iTRAQ) [22]. The advantage with labeling strategies lies in the detection of all conditions within the same MS or MS/MS scan, allowing for more direct and accurate quantitation.

6 Foreground and Background for Enrichment Analyses

The goal of most phosphoproteomics experiments is to understand how phosphorylation events affect a biological system. Evidently, phosphosites that differ between biological conditions are of interest, and such phosphoproteomics experiments should ideally be performed quantitatively. These phosphosites of interest are best understood in the context of the phosphosites that are unchanged in the given experimental data. When possible, it is therefore advisable to generate a reference dataset of unchanged phosphosites (background) and a dataset of changed phosphosites (foreground) from the same data as both will have been subjected to the same experimental biases. These datasets are typically defined by applying statistical tests, such as *t*-tests, to determine the fold change cutoffs.

6.1 Sources of Experimental Biases

Various steps in the phosphoproteomics workflows will inherently introduce biases in the data that can be misinterpreted as biologically relevant if not accounted for. The easiest and most robust approach that we recommend, as previously mentioned, is to generate a reference dataset from the same original data. Common sources of bias include the lysis buffer used, fractionation methods, phosphopeptide enrichment protocols, and MS methods applied. Phosphopeptide enrichment protocols commonly display biases towards either singly or multiply phosphorylated peptides or towards hydrophobic or hydrophilic peptides. Mass spectrometric biases are specific to the fragmentation technique and/or instrument applied. In MS methods where the most intense peaks in a MS spectrum are submitted for MS/MS analysis, there is an inherent bias to sequence peptides that are more abundant. The combination of all the abovementioned biases will be present in the experimental data, stressing the importance of applying a custom reference dataset.

6.2 Example Pitfall Caused by Experimental Bias

The following example systematically dissects a fictive dataset, to illustrate the importance of using an appropriate foreground and background. For simplicity, it is assumed that every protein only gives rise to one phosphorylated peptide.

6.2.1 Example Experiment

Two experimental conditions: control and perturbation.

Our fictive organism has 40,000 different proteins, of which 100 are ribosomal.

4000 SILAC phosphopeptide pairs are identified, of which 50 are ribosomal.

400 of them have a perturbation:control ratio above 5, of which 5 are ribosomal.

The aim is to calculate whether this perturbation of interest enriches ribosomal phosphosites. Using the above data we explore the importance of using a custom background.

6.2.2 Faulty Approach: Global Background

It is common in proteomics data analysis to compare an experimental foreground dataset to an entire proteome database of interest to search for significantly perturbed pathways, sequence motifs, and more. However, this approach assumes that the preparation and analysis of the sample does not introduce any biases. With this approach, when calculating the enrichment factor of ribosomal protein phosphorylation in your data compared to a complete proteome for this fictive dataset (n denotes "number of"):

Expectation: $n_{\text{ribosomal proteins}}/n_{\text{total proteins}} = 100/40{,}000 = 0.25\ \%$

Observation: $n_{\text{ribosomal proteins above cutoff}}/n_{\text{total proteins above cutoff}} = 5/400 = 1.25\ \%$

Enrichment: Observation/Expectation = 1.25 %/0.25 % = 5

Significance: A two-sided binomial test with 400 trials, 5 successes and an expected frequency of 0.25 % gives a *p*-value of 3.6 %, and the enrichment would thus be deemed significant.

Conclusion: The applied experimental perturbation increases phosphorylation on ribosomal proteins fivefold.

However, our fictive dataset has a bias, which is common to most phosphoprotemics workflows, namely an enrichment for abundant proteins. Therefore, 1.25 % of the identified phosphosites reside in ribosomal proteins, compared to only 0.25 % in the total proteome. We therefore encourage using the unperturbed experimental data itself as background when applying this custom background from the experimental data:

Expectation: $n_{\text{ribosomal proteins}}/n_{\text{ total proteins}} = 50/4000 = 1.25\ \%$

Observation: $n_{\text{ribosomal proteins above cutoff}}/n_{\text{total proteins above cutoff}} = 5/400 = 1.25\ \%$

Enrichment: Observation/Expectation = 1.25 %/1.25 % = 1

Significance: A Fisher's exact test using a 2 × 2 contingency table with the values (5, 400) vs. (50, 4000) yields a *p*-value of 100 %, meaning there is a 100 % chance there is no difference between the number of regulated ribosomal proteins and other regulated proteins.

Conclusion: The applied experimental perturbation does not affect phosphorylation on ribosomal proteins.

6.3 Statistical Test Used to Define Foreground and Background Datasets

Foreground and background datasets are typically defined using statistical tests to determine appropriate cutoffs to judge whether a phosphosite is changed or unchanged between experimental conditions. Here we discuss common statistical approaches that can be applied to phosphoproteomics data and when those different tests are appropriate to apply.

Many statistical tests assume that the data conforms to the normal distribution and that the data has the same mean and variance across datasets in experiments with many conditions. It is therefore important when analyzing quantitative data, to perform all statistical tests on log transformed data (usually $\log_2$ or $\log_{10}$), as fold change ratios in a linear scale are not normally distributed. Most proteomics software packages also normalize the output data to compensate for this.

Cutoffs are usually set using *p*-values, the motivation for which is to ensure that the entries in the foreground are significantly different from the background. However, when searching for biologically meaningful regulatory events, it is important to bear in mind that *p*-values only define whether an entity is significantly different from another. Thus, *p*-values are very susceptible to the number of replica and sample variance, which in the correct combination can lead to very small fold change being considered significant.

The nature of the phosphoproteomics data will determine the ideal statistical test, which should be used to calculate *p*-values. In cases where many replicate experiments have been performed, a student's *t*-test is advisable. This test requires many data points (at least three ratios) for each phosphorylation site, as this test is based on both the fold changes and the variance of each ratio across measurements. As such, the student's *t*-test also compensates for experimental and instrumental imposed variance, and should be used when those are expected. However, most phosphoproteomics screens are not performed with enough replicate experiments to perform student's *t*-test, in which case, a test based on the distribution of the entire dataset is beneficial. For this we recommend the Significance A test.

6.3.1 Student's t-tests

There are two types of *t*-tests that in principle could be applied to phosphoproteomics data: the related and the independent *t*-test. Both tests determine whether a phosphorylation site differs between conditions or not, and both are performed on the abundance values of the phosphosites rather than their ratios.

Related t-test: This test compares the abundance of phosphosites between conditions within each replicate experiment. This means that the abundances in condition 1 and 2 are compared within the first replicate, then within the second replicate and so forth. In theory this test has good statistical power, but it is limited by the fact that MS data often has many missing values, in which case the replicate data-points are lost, and the test loses its power.

Independent t-test: The independent *t*-test determines the statistical difference between the mean intensities of a phosphosite across conditions. As this test uses the mean intensities, missing values are tolerable, and it is therefore more robust for MS data compared to the related *t*-test.

Both of these *t*-tests can be performed assuming either equal or unequal variance. Equal variance calculates a variance of the abundance of the phosphosite across conditions and replicates, and the test will therefore use all data points. Unequal variance assumes that the variance is not comparable between conditions, and the test will therefore calculate a variance across replicates but separately for each condition. This will therefore include less data points per variance calculation, and as missing values are common in MS data, applying equal variance is more robust.

6.3.2 Significance A

Significance A is a statistical test that determines for each ratio whether it differs significantly from the distribution of the whole dataset [23]. This test is particularly applicable for phosphoproteomics data that is heavily perturbed. This test takes advantage of the fact that the middle 68.26 % of the ratio distribution conforms better to a normal distribution.

Here the 15.87 %, 50 % and 84.13 % percentiles are assumed to correspond to left standard deviation r_{-1}, the mean r_0 and the right standard deviation r_1 respectively. The distance z (analogous to the standard deviation in t-test) is then calculated and applied to determine a p-value for the ratio of each phosphorylation site:

$$\text{if}\, r > r_0\, \text{then} : z = \frac{r - r_0}{r_1 - r_0}$$

$$\text{if}\, r < r_0\, \text{then} : z = \frac{r - r_0}{r_{-1} - r_0}$$

$$\text{Significance A} = \frac{1}{2} erfc\left(\frac{z}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}} \int_z^{\infty} e^{-0.5t^2}\, dt$$

In the publication introducing Significance A, the authors also proposed the option of correcting the p-values for multiple testing. This can be done with a Benjamini–Hochberg correction.

To get an intuitive idea of how Significance A works, let us imagine data ratios that are distributed as follows:

9000 peptides are unregulated; average log ratio = 0 and a standard deviation = 1

1000 peptides are regulated; average log ratio = 0 and standard deviation = 5

Table 1 outlines the difference between using the “real” standard deviation (which we know for the used model dataset) and using Significance A or RMSD (Root Mean Square Deviation) to calculate the standard deviation (*see* Table 1). First it should be noted, for the above data, significance A overestimates the standard deviation by 21 % whereas RMSD overestimates it by 84 %. Using RMSD only 269 of the 1000 regulated proteins would be found above 3 standard deviations (used as cutoff) as estimated by RMSD. There against 469 proteins would pass this criterion when using significance A, which is much closer to what we would have gotten had we known the “real” standard deviation (549).

Table 1
Table showing how Significance A compares to root mean square distance (RMSD) when estimating standard deviation (SD) on a set of peptides where 9000 are unregulated and have a average log ratio of 0 and a SD of 1 and 1000 are regulated and have a average log ratio of 0 and a SD of 5

	Real	Significance A	RMSD
Standard deviation	1.00	1.21	1.84
Regulated peptides (of 1000) at 3 SD	549 (54.9 %)	469 (46.9 %)	269 (26.9 %)
Unregulated peptides (of 9000) at 3 SD	24 (0.270 %)	3 (0.029 %)	0 (3.2×10^{-6}%)

6.4 Approaches to Applying Cutoffs to Define Foreground and Background Dataset

There are several common approaches to define foreground or background datasets. These typically apply cutoffs for fold change, *p*-values or a combination thereof based on the quantitative phosphoproteome or the proteome from the same experiment. Choosing a good cutoff requires several considerations that we review here.

6.4.1 Setting Cutoffs Based on the Experimental Proteome vs. Phosphoproteome

If the proteome in the given data is expected to be largely unperturbed, for example in short term treatment conditions, the distribution of the proteome can be used to define a cutoff for the phosphoproteomics data. Such proteomics data can be acquired by sampling the experiment prior to phosphopeptide enrichment or be extracted from the non-modified peptides that are sequenced after phosphopeptide enrichment. This is particularly advantageous when comparing datasets with very different amounts of perturbation on the phosphoproteome, as most statistical tests will in this case require more stringent cutoffs for significance in the more perturbed datasets. In a mildly or unperturbed proteome, however, the quantitative data is more comparable across datasets, and will therefore result in more similar stringency when acquiring a cutoff.

Therefore you can perform a statistical test, such as Significance A, on your proteome data and use the identified cutoffs to define regulated events in your phosphoproteomics data. However, caution has to be applied when using proteomics data which is not directly extracted from the phosphoproteomics experiment itself. Due to less complex sample preparation protocols, data distribution can be narrower, leading to an overestimation of regulated phosphopeptides compared to the proteome.

6.4.2 Additional Considerations

Defining a cutoff will deem all entries above the cutoff to be regulated, while everything below is unregulated. However, biological systems are not binary and events falling right below the cutoff might still be regulated and biologically relevant. Therefore it is often advisable to have cutoff for the background as well as for the foreground. Events falling into the “grey zone” between those two cutoffs will be considered neither regulated nor unregulated and therefore not used for the analysis. Commonly only the ratios

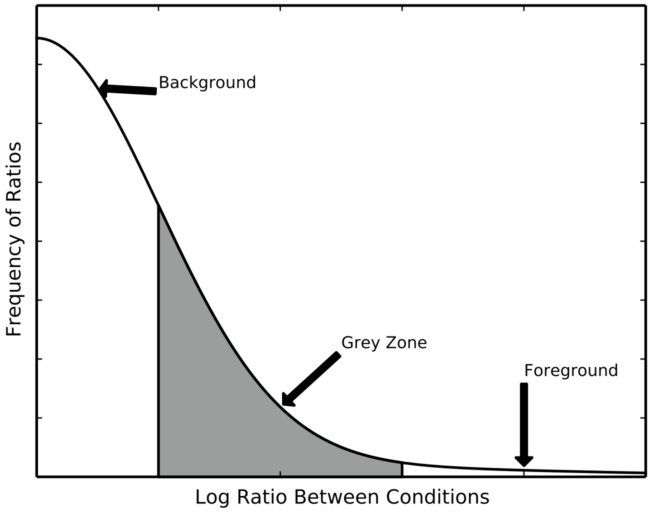


Fig. 2 Data separated into a Foreground and Background, the Grey Zone data is not used for further analysis, as it contains a mixture of regulated and unregulated peptides

within 1 standard deviation of the sample mean are used as the background cutoff, thus if a cutoff of 3 standard deviations has been chosen, then peptides with a standard deviation between 0 and 1 will become the background dataset and everything with a standard deviation above 3 will become the foreground dataset, as visualized in Fig. 2.

Considering the biological question of a specific analysis type, the user may benefit from performing all downstream analysis twice: once for the upregulated and once for the downregulated phosphosites. This is of particular relevance if a given dataset is skewed towards upregulation or downregulation, e.g., in the context of kinase or phosphatase inhibitors.

7 Platforms for Phosphoproteomics Analysis

Many software packages have been developed to search, filter and/or quantify (phospho)peptides. Below we present a few popular software packages that can perform all three steps of the analysis. All presented software is very mature, and the choice of software package is therefore usually based on user preferences. Generally the software can be classified into two different categories:

1. Pipeline/workflow oriented tools offer very high levels of flexibility and automation and are capable of creating workflows that can be reused for different types of follow-up analyses. However, generation of the workflow requires time and familiarity with the software, and can be difficult for first-time users.

2. Conventional software packages are generally equally powerful and tend to be easier to learn; however, they offer a lower degree of automation and flexibility. For those software options we put an emphasis on MaxQuant, which is freely available and very user friendly.

7.1 Workflow- and Pipeline-Based Software

Workflow-based software packages allow the users to generate a workflow, in a drag-and-drop manner. Once a workflow has been generated, it can be saved and shared among coworkers ensuring that everybody analyzes the data in the exact same way. Pipeline software packages work like workflows, but are generally made up of small command-line programs that can be chained together to form a pipeline by scripting.

7.1.1 Trans-Proteomic Pipeline

The Trans-Proteomic Pipeline [24] (TPP) is an open-source project from the Seattle Proteome Center (SPC). TPP is a web-based front end to a large collection of command-line tools. It can be used for almost any combination of MS instruments and labeling techniques.

A typical approach for using this pipeline in processing of quantitative phosphoproteomics data would include the following steps:

1. Standard input: Convert vendor format to mzXML (mzML or mzData)
2. Peptide assignment: Search data against one of the following databases: SEQUEST [17], MASCOT [18], COMET [25], ProbID [26] X!Tandem [13], or any other database of interest.
3. Validation:
   (a) Rank (phospho)peptides based on scores and filter based on a user-specified FDR using PeptideProphet [27].
   (b) Optionally: assemble peptides into proteins using ProteinProphet [27]
4. Quantification: use ASAPRatio [28] to calculate peptide ratios.

The software is very easy to install for Windows. While, Linux is officially supported it requires editing the *make* file to compile the source code. The advantage of this platform is that it covers all steps in phosphoproteomics data processing, from format conversion to quantification. This platform could be useful for inexperienced users in phosphoproteomics data analysis, as it offers pipelines that guide the user through the workflow.

7.1.2 TOPP and TOPPAS

The OpenMS [10] Proteomics Pipeline [11] (TOPP) is a large collection of programs with a command-line interface that can be chained together, much like the TPP. TOPPAS [9] is a workflow-based graphical front end to TOPP. TOPP/TOPPAS can use the following search databases: Mascot [18], MyriMatch [15], OMSSA [14] and X!Tandem [13]. TOPP and TOPPAS are available for Windows, Linux, and OS X. For working with this platform you can follow the same four steps as described above for TPP. However, it allows for a greater level of flexibility in setting up the workflow.

7.1.3 Proteome Discoverer™

Proteome Discover is commercially available Windows software developed by Thermo Scientific. Its main focus is on data generated with Thermo Scientific Orbitrap instruments. Like TOPPAS it is also used in a workflow manner and can supports most of the popular search databases and labeling techniques.

7.2 Conventional Software Packages

Conventional software packages offer a workflow, which cannot be edited. However, these platforms generally allow a great level of flexibility within this scheme. This approach is more readily accessible as the platforms are very user-friendly and recommendable for beginners.

7.2.1 Mascot

Mascot [18] was developed by Matrix Science and is one of the oldest and most well-established database search engines. Mascot refers to the core database search algorithm, but complete data processing requires two main products: Mascot Server, which does the database search, and Mascot Distiller, which can do validation and quantitation. Mascot is only available for Windows.

7.2.2 MaxQuant

MaxQuant [23] is freeware developed at the Max Planck Institute of Biochemistry. MaxQuant uses a search database, Andromeda, developed specifically for this platform [16]. In recent years, the development of this platform has focused on including features that allow for compatibility with many different MS instruments (Thermo *.Raw, Brucker *.d, Sciex *.wiff and mzXML) and labeling techniques. This particular advantage renders MaxQuant an attractive software for the broad MS user community.

MaxQuant offers the option to group your input data, so that the chosen parameters can be applied specifically for a group or globally for all the data analyzed. As such, the user can analyze proteomics and phosphoproteomics data in parallel by applying specific parameters to some of the data (such as the search for phosphorylation in the phosphoproteome) while still using shared parameters, such as FDR cutoffs and the FASTA file being queried. Additionally, MaxQuant allows the option to configure the Andromeda database search engine, for example to include new modifications and FASTA files of interest. MaxQuant is only available for Windows.

8 Protocol: Phosphoproteomics Analysis with MaxQuant

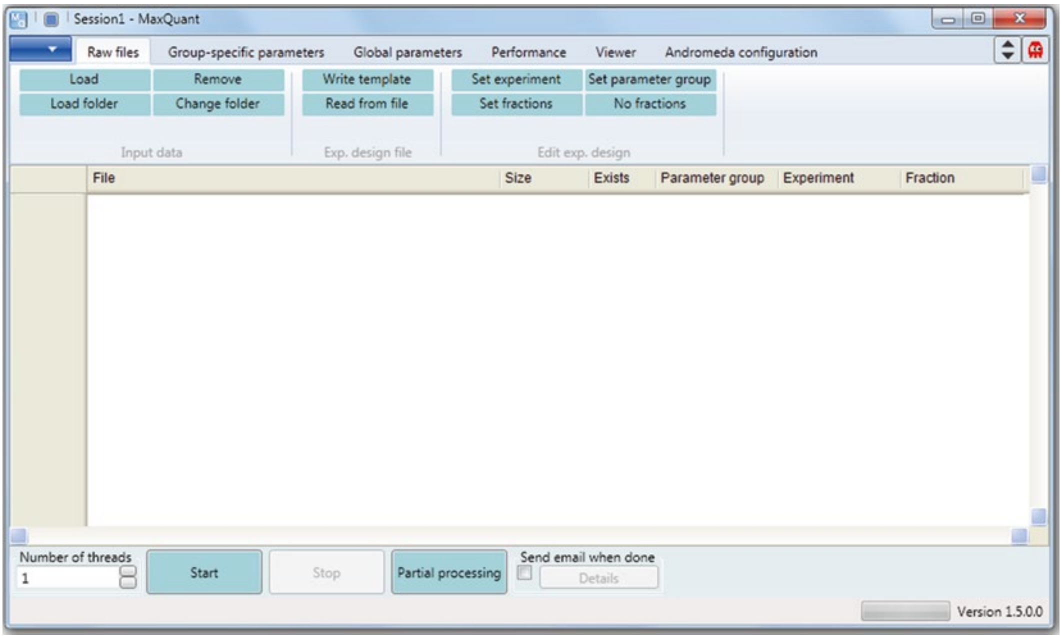
8.1 Materials

To use MaxQuant, download the latest version from http://www.maxquant.org/. This protocol is designed for v. 1.5.0.0 but is readily transferrable to other versions.

Download a relevant proteome in FASTA format (*see* **Note 1**), for example from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes.

8.2 Method

1. Launch MaxQuant. The user is presented with the interface seen in Fig. 3.
2. Navigate to "Raw files" tab.
3. Under *Input data* click "Load" (or Load folder) to import MS data files (or complete folders containing raw MS data) into MaxQuant.
4. Highlight all files that belong to the same experiment (usually also grouping replicates), and under *Edit exp. Design* click "Set Experiment." Write a descriptive name for the given experiment in the popup menu. All files should be assigned to an experiment.
5. If the files include different experimental workflows that require applying different parameters, the files must be grouped accordingly (*see* **Note 2**). Highlight the files to be grouped



Number of threads 1

Version 1.5.0.0

Fig. 3 Screen shot of MaxQuant version 1.5.0.0

together, by clicking "Set parameter group" and indicate a group number.

6. Navigate to *Group-specific parameters* tab.
7. For every parameter group repeat the following steps.
8. Under *Parameter section* select "General."
9. Under *Type* change the multiplicity (number of labels) to reflect the numbers of conditions in your experiment and select labels appropriately (*see* **Note 3**).
10. In the *Variable modifications* section keep default settings and scroll down to select modifications of interest if applicable. For phosphoproteomics experiments chose "Phospho (STY)" then click the ">" button.
11. In the *Digestion mode* section select the relevant enzyme using the ">" and "<" buttons (*see* **Note 4**).
12. Under *Parameter section:* select Instrument, and change values to reflect the quality and settings of the instrument that generated the MS data. Default parameters are provided for Orbitrap, Brucker TOF and AB Sciex TOF.
13. If applicable select "LFQ" in the drop-down menu, under Label-free quantification in the *Parameter section.*
14. Navigate to *Global Parameters* tab.
15. Under Parameter section select "General."
16. In the *FASTA files* section click "Add file" to important the proteome FASTA file.
17. Under *Identification* change PSM FDR (FDR at the spectrum level), Protein FDR and Site decoy fraction (modified peptides FDR) (*see* **Note 5**).
18. Navigate to *Performance* tab.
19. In the footer of the program change the number of threads (cores) to be used for processing. This can be up to the number of cores on the computer (*see* **Note 6**).
20. In the footer of the program click "Start."

9 Notes

1. Not all FASTA files are configured in Andromeda. Under the "Andromeda configuration" tab, this can be checked, and file of interest can be configured for use in MaxQuant.
2. Examples of situations where grouping files to apply group specific parameters would be useful:
   (a) Some of the files contain data with two labels while others have three.

(b) Different variable modifications are desired for different files, as some are to be used for phosphoproteome determination and others for proteome.

(c) The MS data files were generated using different MS-instruments or instrumental settings.

3. A triple SILAC experiments requires a multiplicity of 3. These should be specified: labels-1: nothing indicated, labels-2: Arg6 and Lys4, labels-3: Arg10 and Lys8.
4. Here, indicate the enzymes used for digestion in the experimental workflow. This is typically Lys-C and/or trypsin for most phosphoproteomics workflows.
5. We suggest setting all FDRs to 1 %.
6. If your system has less than 1GB ram per core, set the number of threads to the same number of GB ram available. Regardless of processing power, if all cores are used the processing capacity of the computer will be consumed.

Acknowledgments

This work was in part funded by the Novo Nordisk Foundation Center for Protein Research [NNF14CC0001]

References

1. Cohen P (2002) The origins of protein phosphorylation. Nat Cell Biol 4(5):E127–E130
2. Hughes C, Ma B, Lajoie GA (2010) De novo sequencing methods in proteomics. Methods Mol Biol 604:105–121
3. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B (2012) PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 11(4), M111.010587
4. Lam H (2011) Building and searching tandem mass spectral libraries for peptide identification. Mol Cell Proteomics 10(12) R111.008565
5. Eng JK, Searle BC, Clauser KR, Tabb DL (2011) A face in the crowd: recognizing peptides through database search. Mol Cell Proteomics 10(11) R111.009522
6. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1(5):376–386
7. Ong S-E, Mann M (2006) A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat Protoc 1(6):2650–2660
8. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30(10):918–920
9. Junker J, Bielow C, Bertsch A, Sturm M, Reinert K, Kohlbacher O (2012) TOPPAS: A Graphical Work flow Editor for the Analysis of High-Throughput Proteomics Data. J Proteome Res 11(7):3914–3920
10. Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-

Trieglaff O, Zerck A, Reinert K, Kohlbacher O (2008) OpenMS—an open-source software framework for mass spectrometry. BMC Bioinformatics 9:163

11. Kohlbacher O, Reinert K, Gröpl C, Lange E, Pfeifer N, Schulz-Trieglaff O, Sturm M (2007) TOPP–the OpenMS proteomics pipeline. Bioinformatics 23(2):e191–e197
12. Deutsch EW (2012) File formats commonly used in mass spectrometry proteomics. Mol Cell Proteomics 11(12):1612–1621
13. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20(9):1466–1467
14. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH (2004) Open mass spectrometry search algorithm. J Proteome Res 3(5):958–964
15. Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis research articles. J Proteome Res 6(2):654–661
16. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10(4):1794–1805
17. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5(11):976–989
18. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20(18):3551–3567
19. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4(3):207–214
20. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127(3):635–648
21. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. Nature 473(7347): 337–342
22. Wiese S, Reidegeld KA, Meyer HE, Warscheid B (2007) Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. Proteomics 7(3):340–350
23. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26(12):1367–1372
24. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R (2010) A guided tour of the trans-proteomic pipeline. Proteomics 10(6):1150–1159
25. Eng JK, Jahan TA, Hoopmann MR (2013) Comet: an open-source MS/MS sequence database search tool. Proteomics 13(1):22–24
26. Zhang N, Aebersold R, Schwikowski B (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics 2(10):1406–1412
27. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74(20):5383–5392
28. Li X-J, Zhang H, Ranish JA, Aebersold R (2003) Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal Chem 75(23):6648–6657