Journal of proteome research

Technical Note

pubs.acs.org/jpr

Param-Medic: A Tool for Improving MS/MS Database Search Yield by Optimizing Parameter Settings

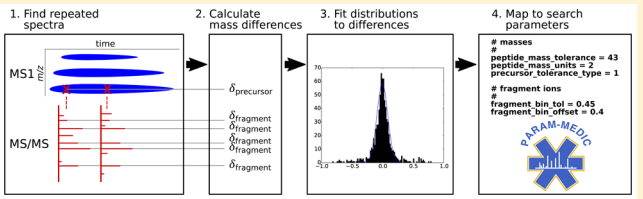
Damon H. May,[†] Kaipo Tamura,[†] and William S. Noble*[,†,‡]

[†]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States

[‡]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, United States

Supporting Information

ABSTRACT: In shotgun proteomics analysis, user-specified parameters are critical to database search performance and therefore to the yield of confident peptide-spectrum matches (PSMs). Two of the most important parameters are related to the accuracy of the mass spectrometer. Precursor mass tolerance defines the peptide candidates considered for each spectrum. Fragment mass tolerance or bin size determines how close observed and theoretical fragments must be to be considered a match. For either of these two parameters, too wide a setting yields randomly high-scoring false PSMs, whereas too narrow a setting erroneously excludes true PSMs, in both cases, lowering the yield of peptides detected at a given false discovery rate. We describe a strategy for inferring optimal search parameters by assembling and analyzing pairs of spectra that are likely to have been generated by the same peptide ion to infer precursor and fragment mass error. This strategy does not rely on a database search, making it usable in a wide variety of settings. In our experiments on data from a variety of instruments including Orbitrap and Q-TOF acquisitions, this strategy yields more high-confidence PSMs than using settings based on instrument defaults or determined by experts. Param-Medic is open-source and cross-platform. It is available as a standalone tool (http://noble.gs.washington.edu/proj/param-medic/) and has been integrated into the Crux proteomics toolkit (http://crux.ms), providing automatic parameter selection for the Comet and Tide search engines.



KEYWORDS: *database search, mass accuracy, tandem mass spectrometry*

1. INTRODUCTION

Database search algorithms such as Sequest[1] serve as the core of many shotgun analysis pipelines. Most search engines require a long list of user-supplied parameters, including cleavage enzyme, number of missed cleavages to allow, static and variable peptide modifications, and tolerances to use in matching observed precursor and fragment masses to their theoretical counterparts. Appropriate values for these parameters depend on the instrument used, the instrument settings used for a particular analysis, instrument performance at the time of acquisition, and other factors.

In this work, we focus on two of the most important search algorithm parameters. Precursor mass tolerance defines the peptide candidates considered for each spectrum. A narrower setting reduces the running time of the search algorithm by requiring it to perform fewer comparisons between peptides and spectra, but a too-narrow setting can exclude true matches. Too wide a setting can reduce sensitivity in a different way: because more candidates are considered for each spectrum, the chance of a false match randomly generating a higher score than a true match increases.[2] Similarly, fragment mass tolerance or bin size determines how small the absolute value of the difference between a pair of observed and theoretical fragment masses must be to consider them a match. A tighter setting can exclude true matches between fragments, while a loose setting can lead to false matches between fragments, leading to more high-scoring false matches.

An important goal of many proteomics workflows is to achieve high statistical power for peptide detection. A commonly used proxy for the peptide detection power of a database search is the number, or "yield," of peptide-spectrum matches (PSMs) at a set false discovery rate (FDR) such as 0.01, as estimated by target-decoy procedure.[3] We define the optimal value for precursor or fragment mass tolerance as the value that yields the most PSMs at FDR 0.01. The optimal value for either parameter may vary widely from experiment to experiment. This sensitivity to parameter settings has a real impact on experimental results because the measurement of yield can vary greatly between the best and the worst parameter settings.

Researchers adopt different strategies to arrive at the settings they use for a given analysis. Some laboratories fine-tune the optimal settings for a particular instrument by performing searches on acquired data with many different settings. Because instrument performance can change over time to cause drift in both mass accuracy and calibration, researchers most concerned with using the proper settings will periodically perform measurements solely to reassess performance. On the other

Received: January 13, 2017
Published: March 6, 2017

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824

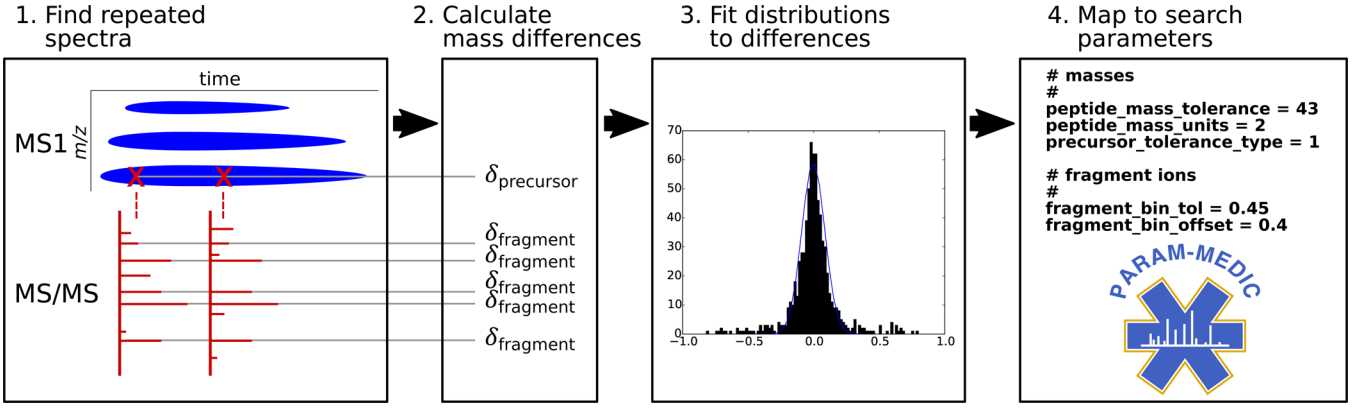


Figure 1. Param-Medic workflow. Param-Medic collects pairs of closely eluting MS/MS spectra and assembles their pairwise precursor and most intense five fragment mass differences. Precursor and fragment error are inferred by fitting a mixed Gaussian/uniform distribution to pairwise differences. Search parameter values are chosen by multiplying estimated error standard deviation by a multiplier associated with highest mean PSM yield in training data sets.

extreme, database searches are often performed by researchers other than those who ran the instrument, as when laboratories share data or when spectra are reanalyzed after being deposited in a public repository. In the absence of detailed information about how the instrument was run or how well it was performing at that time, researchers typically rely on instrument settings reported by the lab that ran the instrument or on the advertised capabilities of the instrument that was used.

Several tools have been developed to aid researchers in selecting optimal search parameter values. Many of these tools infer instrument calibration from experimental data by analyzing the observed m/z values of known ions: either spiked-in peptides or peaks confidently identified by database search.[4−7] One such tool for the Windows platform, Preview,[5] additionally assesses precursor and fragment mass error, nonspecific digestion, and sample modifications using a fast database search. However, neither Preview nor any of the other tools we surveyed provides a well-defined method for translating assessed m/z error into parameter settings for database search.

Here we describe Param-Medic: an open-source, cross-platform tool for assessing experimental m/z error and deriving parameters to search an LC−MS/MS experiment. We have trained Param-Medic to produce parameters appropriate for the Comet[8] search engine, but the same strategy could be extended to work for any algorithm. At the heart of Param-Medic is a key assumption that despite the use of so-called "dynamic exclusion" rules, LC−MS/MS experiments typically make multiple observations of many individual peptide ions. Param-Medic exploits these repeated observations to enable estimation of m/z error. Specifically, the algorithm assesses measurement precision (but not calibration) by identifying pairs of spectra likely to represent the same peptide and then analyzing the distribution of differences between those pairs' precursor and matched fragment ion m/z values. We trained Param-Medic on eight data sets from public repositories from a variety of organisms and instruments. We evaluated its performance on three additional public data sets from the same instruments as well as on a data set generated using a very different Q-TOF instrument. Param-Medic is available as a standalone tool and as a part of the Crux proteomics toolkit, providing an open, integrated platform for parameter inference and database search.

2. METHODS

2.1. Mass-to-Charge Error Estimation

Param-Medic infers both precursor and fragment m/z search parameters in a four-step procedure (Figure 1). First, it pairs closely eluting MS/MS spectra that have similar precursor and fragment m/z values. Then, it calculates the mass differences of both the paired precursors and the paired fragments. Next, it fits a separate mixed Gaussian-Uniform distribution to the error values for precursors and for fragments. Finally, it maps the standard deviation of each estimated Gaussian distribution to a value usable as a precursor tolerance or fragment bin size for database search.

Param-Medic begins by assembling pairs of measurements from spectra with an inferred charge of 2 that appear likely to represent the same precursor ion or fragment ion (Figure 1). Spectra are paired permissively to generate distributions of pairwise measurement differences with sufficient numbers of both correctly and incorrectly paired spectra so that the two component distributions can be estimated. Precursor and fragment masses are calculated from their observed m/z values and are each binned coarsely with bin size 1.0005079, corresponding to the distance between the centers of two adjacent peptide mass clusters.[9] One list of paired measurements is initialized for precursor values and another for fragments.

As Param-Medic processes each sequential MS/MS scan, the algorithm identifies the previous MS/MS scan within the last 1000 scans whose precursor falls in the same bin (if any). It then checks whether the associated precursor m/z is within 50 ppm of the precursor m/z of the new scan and whether at least 20 of the 40 most-intense binned fragments are unambiguously shared between the two spectra. If both conditions are met, then the two spectra are considered to represent the same peptide ion. In this case, the two precursor m/z values and the paired values for the five most-intense pairs of fragment m/z values are added to their respective lists. No single spectrum is included in more than one such pair, and additional measurements of the same ion are paired rather than being assembled into higher order tuples. If Param-Medic detects fewer than 200 such pairs (an arbitrary threshold that may be adjusted as desired), then the program will terminate without estimating parameter settings.

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824

Table 1. Experiments Used in the Training and Testing of Param-Medic and Their Associated Search Parameters As Adapted from Their Publications

experiment	instrument	organism	precursor tolerance (ppm)	fragment bin size (Th)
		Training Data Sets		
2014kim-kidney[14]	Orbitrap Velos	human	10	0.05
2014kim-lung[14]	Orbitrap Elite	human	10	0.05
2015clark-redefining[15]	LTQ Orbitrap	human	50	1
2015radoshevich-isg15[16]	QExactive	human	4.5	0.02
2015tanca-impact[17]	Orbitrap Velos	human gut microbiome	10	0.02
2015uszkoreit-intuitive[18]	Orbitrap Elite	mouse	5	0.4
2016mann-unpublished	QExactive	human	10	0.02
2016schittmayer-cleaning[19]	Orbitrap Velos	yeast	10	0.8
		Test Data Sets		
2016may-metapeptides[12]	Qexactive	ocean microbiome	10	0.02
2016audain-in-depth[20]	LTQ Orbitrap	yeast	25	0.5
2016zhong-quantitative[21]	Orbitrap Velos	human	20	0.5

In the second step, the ppm differences in measurement pairs are calculated from the pairs of measurements. This step and the following steps are performed separately but identically for precursor pairs and for fragment pairs. The output of this step is an empirical list of ppm differences in paired peak measurements. In practice, this list represents a mixture of differences between two correctly paired measurements of the same peak and differences between two incorrectly paired measurements of peaks that represent different ions. Below, we refer to these as "true" and "false" pairs, respectively.

In the third step, Param-Medic fits a theoretical distribution to the empirical distribution of errors from step two. Param-Medic assumes that ppm measurement error for true pairs is normally distributed. Therefore, the difference between two values drawn from the distribution of ppm measurement error is also normally distributed, with variance twice that of the measurement error. Param-Medic also assumes that differences between false pairs are uniformly distributed over the range considered. Accordingly, it models the distribution of measurement differences as a mixed Gaussian ($\mathcal{N}(y)$ for observed differences y) and uniform distribution. Expectation-maximization (EM) is used to estimate three parameters: the mean and standard deviation of the Gaussian distribution component ($\hat{\mu}_\delta$ and $\hat{\sigma}_\delta$) and the probability of membership in the Gaussian distribution (p_G). EM maximizes the log-likelihood of the observed data

$$\sum_{i=1}^{n} \ln\left(p_G \mathcal{N}(y_i; \mu_\delta, \sigma_\delta) + (1 - p_G)\frac{1}{b - a}\right) \quad (1)$$

The algorithm alternates between an E step, which estimates expectation of the log-likelihood using the current parameter estimates, and an M step, which computes new parameter values maximizing the expected log-likelihood. Once $\hat{\sigma}_\delta$ is estimated, the standard deviation of the measurement error, $\hat{\sigma}_\epsilon$, is estimated as $\hat{\sigma}_\epsilon = \frac{\hat{\sigma}_\delta}{\sqrt{2}}$.

In the final step, having estimated the standard deviation of the ppm error distributions, Param-Medic applies a scaling factor to $\hat{\sigma}_\epsilon$ to calculate the estimated optimal search parameter (either precursor tolerance or fragment bin size). This scaling factor is empirically estimated on an analysis of data from a wide variety of mass spectrometry experiments, as described in the following sections.

Many of Param-Medic's parameters are adjustable. The values mentioned above for the charge state (2), wide ppm tolerance (50 ppm), number of peaks that must be shared between spectrum pairs (20 of the most-intense 40), number of fragments per pair used for estimation (5), number of difference measurements required for estimation (200), and maximum scan distance between spectrum pairs (1000) are defaults that should be widely applicable but may be adjusted for unusual data sets. For example, a user may wish to choose a higher charge state when analyzing an experiment on tryptic peptides known to contain a very high proportion of missed tryptic cleavages or to remove the maximum scan distance constraint altogether for very long gradients.

2.2. Search of Public Data Sets with Different Parameter Values

For use in learning the scaling factors mapping $\hat{\sigma}_\delta$ to search parameter values, we collected eight training and three test data sets from the PRIDE[10] and Chorus Project (http://chorusproject.org) proteomics data repositories, representing a variety of organisms and instruments (Table 1). All database searches were performed using Comet[8] version 2015.01 rev. 2. Samples were searched against the appropriate UniProt databases for single organisms, Human Microbiome Project stool database for gut microbiome,[11] or a site-specific sequencing-derived database for ocean microbiome.[12] We used a concatenated decoy database in which peptide sequences were reversed but C-terminal amino acids left in place. Search parameters included a static modification for cysteine carbamidomethylation (57.021464) and a variable modification for methionine oxidation (15.9949). Enzyme specificity was trypsin with proline cleavage suppression, with one missed cleavage allowed. Parent ion mass tolerance was defined around five isotopic peaks. FDR was calculated by target-decoy competition using Percolator,[13] and PSMs were accepted at FDR 0.01.

The most basic method of choosing parameters is to use settings associated with the typical performance of the instrument. This method is often used when the experimental details related to a data set are unknown. In characterizing the instruments used to generate the training and test data sets, we deliberately used only the information available in the repository metadata, as would most researchers downloading the data set from the repository. In several data sets, more detailed information, such as the mass analyzer used, was not publicly available. We defined "instrument default" settings for precursor ppm error and fragment bin tolerance for each instrument represented by the training and test data sets (Table

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824

2) based on advertised instrument capabilities and literature search. We then held fragment bin tolerance for each experiment at the instrument default and performed 10 separate searches, with settings for precursor ppm error varying uniformly over the range 5−50 ppm. Similarly, we held precursor ppm error at the instrument default and performed 10 additional searches with settings for fragment bin tolerance varying uniformly over the range 0.02 to 1.0005 Da. A related parameter, fragment bin offset, should be set to roughly 0.4 when fragment bin size is near 1.0005 to ensure that the highest proportion possible of peaks associated with the same nominal mass are included in the same bin but has little effect for other bin size values. This parameter was set to 0.4 in all searches. PSM yield for each search was defined as the number of PSMs at FDR 0.01.

Table 2. Settings Used in "Instrument Default" Searches

instrument	precursor (ppm)	fragment bin (Th)
LTQ Orbitrap	20	1.005
Orbitrap Velos	20	0.05
Orbitrap Elite	20	0.02
QExactive	20	0.02

2.3. Mapping Estimated Error To Search Parameter Values

The final outputs of Param-Medic are precursor and fragment m/z tolerance values for use in a database search. To produce these estimates, we used the search results from our eight training data sets over a wide range of parameter settings, along with the empirical error standard deviations $\hat{\sigma}_e$, to estimate a multiplier that converts $\hat{\sigma}_e$ values into database search parameters that maximize PSM yield for a wide range of data sets. To this end, we normalized for differences in measurement error across the eight training data sets as follows. Separately for each parameter (precursor m/z tolerance and fragment bin size), we divided each parameter value vraw_i by the corresponding measurement error standard deviation $\hat{\sigma}_e$ for that sample and then calculated a normalized value $\hat{v}_i$ as the natural log of the result

$$\hat{v}_i = \ln\left(\frac{v_i}{\hat{\sigma}_e}\right) \tag{2}$$

We then normalized the PSM yield y_{e_i} associated with the search of an experiment e with the ith value for the parameter by dividing by the highest PSM yield observed for experiment e under any parameter setting

$$\hat{y}_{e_i} = \frac{y_{e_i}}{\max_{1\le j\le n} y_{e_j}} \tag{3}$$

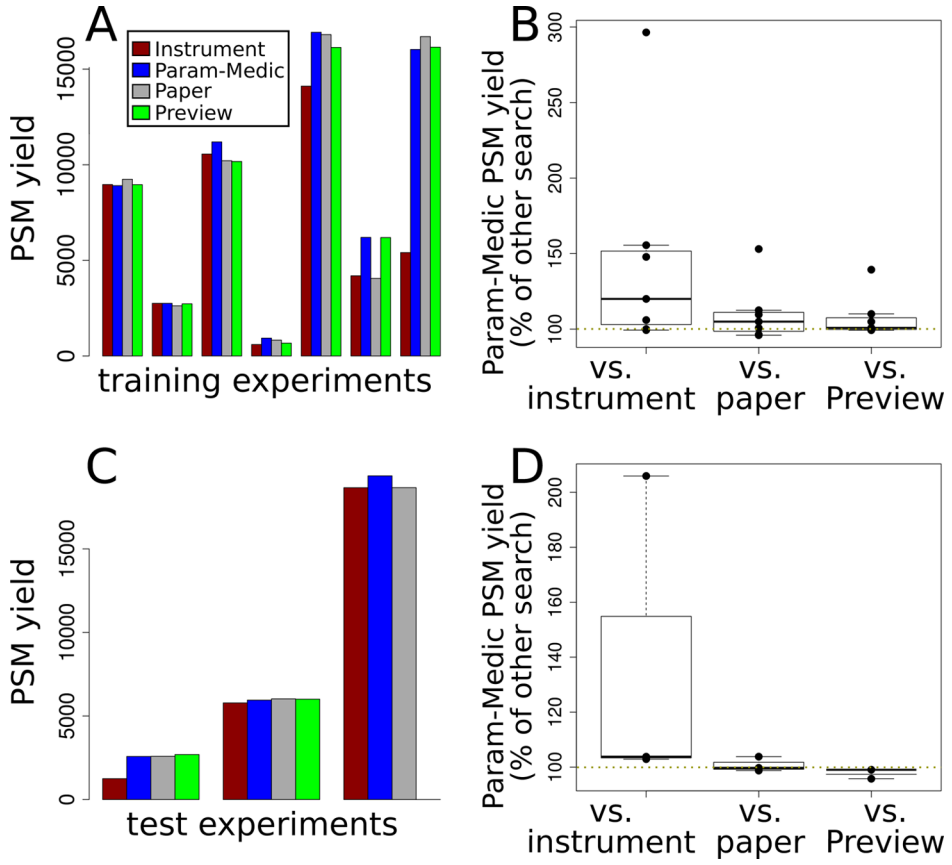


Figure 2. Comparing Param-Medic with other methods. (A) PSM yield at FDR 0.01 using parameters determined by four different methods: instrument defaults, Param-Medic, original paper settings, or Preview. Each cluster of bars represents one of the seven training experiments for which Param-Medic and Preview returned error estimates. Results are reported for the seven training data sets. (B) Box plots showing the distribution PSM yield of searches with Param-Medic parameters as a percentage of the PSM yield using instrument defaults, original paper settings, and Preview, over the same seven training experiments. (C,D) As panels A and B but showing data from the three test experiments.

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824

For each experiment, this process yielded a different set of normalized parameter setting values, each associated with a different normalized PSM yield. To estimate the value associated with the highest mean normalized PSM yield over all experiments, we segmented the range from the minimum to the maximum values of the normalized parameter setting into 200 bins. We defined the yield of experiment e in bin b, $\hat{y}_{e_b}$, as the normalized PSM yield in that experiment associated with that bin, interpolating linearly between adjacent observed measurements $\hat{y}_{e_i}$ and using the yields for the bins with highest and lowest normalized parameter values for each data set to stand in for all higher value or lower value bins not searched for that data set (Figure 3). We then chose the bin b' associated with the highest mean-normalized yield over the n experiments

$$b' = \arg\max_b \frac{1}{n} \sum_{e=1}^{n} \hat{y}_{e_b} \quad (4)$$

The center of bin b', $\overline{b}'$, is the natural log of Param-Medic's estimate of the optimal multiplier relating one of the two $\hat{\sigma}_e$ values to its corresponding search parameter value. Therefore, to calculate the optimal precursor tolerance or fragment bin size, Param-Medic multiplies the appropriate $\hat{\sigma}_e$ estimate by its associated $\exp(\overline{b}')$.

Param-Medic will refuse to estimate precursor error or fragment bin tolerance if there are fewer than 200 pairs of values that make up the mixed distribution. It will also fail if, as was the case in one of our training data sets, at least half of the values in the mixed distribution are exactly 0. This situation occurs when the values are rounded, and it is incompatible with the Param-Medic approach.

2.4. Alternative Parameter-Setting Strategies

We compared search PSM yield from settings determined by Param-Medic with PSM yield from searches using other means of determining search parameters. In addition to the instrument defaults described above, we also derived parameter settings from the publications describing the data sets (or, in the case of one as-yet-unpublished training data set, from the experimental metadata provided in the PRIDE repository for project ID PXD002854). Because the data sets were originally searched with a variety of search algorithms, the published parameter values may not map directly to Comet precursor tolerance and fragment bin size; ours is a good faith effort to represent the original searches as accurately as possible within the Comet/Percolator framework. We also used Preview to assess precursor and fragment median m/z error. To map these Preview-estimated error values to Comet search parameters, we used five times the median error, which is the "rule of thumb" suggested in the Preview user manual.

3. RESULTS

3.1. Param-Medic's Performance

We evaluated Param-Medic's performance in terms of PSM yield, comparing it with the settings used in the original papers describing our data sets, with instrument default settings, and with Preview. On seven training data sets (Figure 2), Param-Medic parameter settings yielded 96−152% as many PSMs as settings from the original papers (median: 105%) and 99−334% as many as defaults based on instrument type (median: 103%). Param-Medic failed to find a sufficient number of repeated ions for parameter estimation on one training data set because of a large proportion of exactly identical sequential values for precursor m/z, which we speculate was due to rounding of the precursor m/z values. Preview failed on the same training data set as Param-Medic due to insufficient search results for error estimation. On the remaining seven data sets, Param-Medic yielded 99−135% as many PSMs as Preview (median: 101%).

On three test data sets, Param-Medic parameter settings yielded 99−104% as many PSMs as settings from the original papers describing the experiments (median: 100%) and 103−212% as many PSMs as defaults based on instrument type (median: 104%). Preview failed on one test data set due to insufficient search results for error estimation. On the other two, Param-Medic yielded 95 and 99% as many PSMs as Preview (Figure 2).

To assess the suitability of Param-Medic for evaluating a different kind of mass spectrometry data, we used it to evaluate a human data set from a SCIEX TripleTOF 5600 (PRIDE accession number PXD000307). When we searched this data set using the parameters specified in the PRIDE submission (10 ppm precursor tolerance, 0.4 Th fragment bin size), the PSM yield was 2738. Yield with parameter values estimated by Param-Medic (10.45 ppm precursor tolerance, 0.03 Th fragment bin size) was 2949, an increase of 7.7%.

Any method for automatically estimating m/z search parameters should be fast as well as effective at optimizing PSM yield. On a 3.0 GHz Intel Core Duo processor, Param-Medic ran in a few seconds to just over a minute on all training and test data sets, while Preview ran in a few minutes to nearly 1.5 h (Table 3). Param-Medic's running time scaled with the number of spectra per experiment, while Preview's scaled with both the number of spectra and the size of the database. Preview took 88 min to run on the human gut microbiome sample, which it searched against a large gut microbiome database, even though that sample had just 10% more spectra than a human sample on which Preview ran in 14 min. The Preview running times are dominated by the database search but also include some time spent performing activities not required for inferring mass error (e.g., inferring peptide digestion and variable modifications).

Table 3. Wall-Clock Running Times for Preview and Param-Medic on Each Experiment, in Minutes[a]

experiment	organism	spectra	preview	Param-Medic
	Training Data Sets			
2014kim-kidney[14]	human	9072	2	0.07
2014kim-lung[14]	human	17 612	3	0.13
2015clark-redefining[15]	human	38 570	N/A	N/A
2015radoshevich-isg15[16]	human	63 185	14	1.03
2015tanca-impact[17]	human gut microbiome	69 685	88	0.48
2015uszkoreit-intuitive[18]	mouse	26 992	6	0.67
2016mann-unpublished	human	41 157	7	0.12
2016schittmayer-cleaning[19]	yeast	9297	1	0.19
	Test Data Sets			
2016may-metapeptides[12]	ocean microbiome	98 317	N/A	0.68
2016audain-in-depth[20]	yeast	18 175	2	0.35
2016zhong-quantitative[21]	human	14 962	3	0.27

[a]"N/A" indicates that a tool did not run successfully on a given experiment.

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824

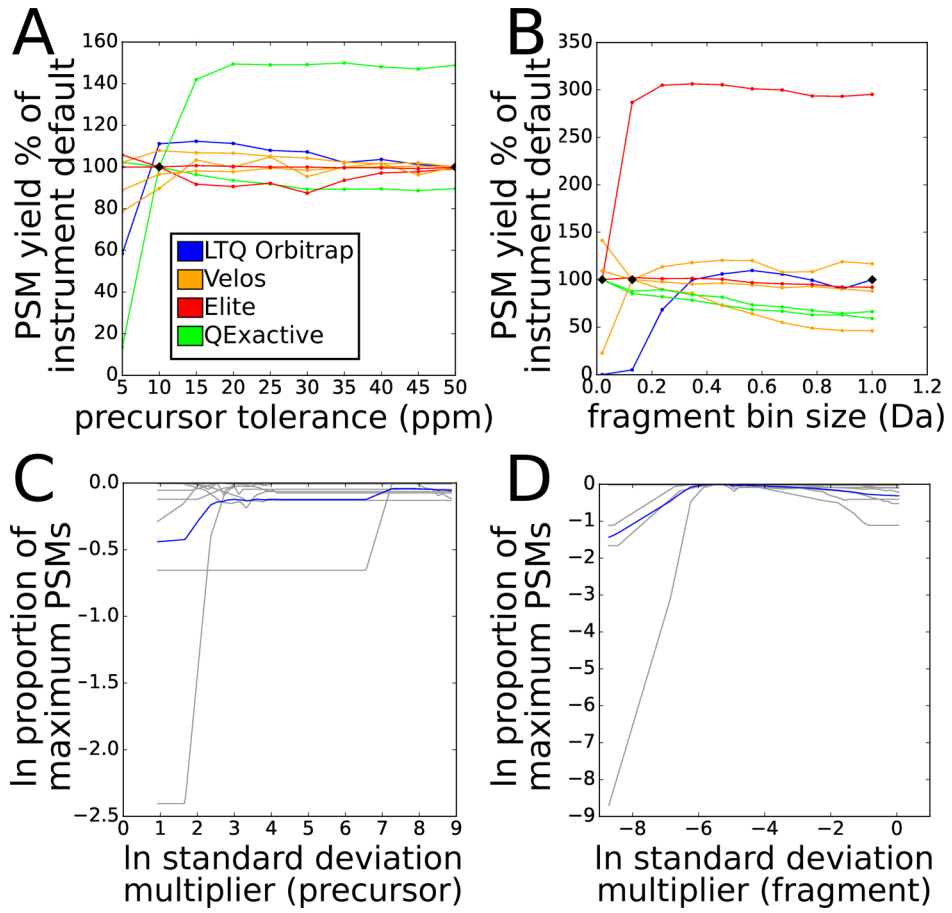


Figure 3. PSM yield versus parameter settings in training data sets. (A,B) PSM yield at FDR 0.01 as a function of the percentage of the PSM yield for that data set when searched with instrument default settings. Each line represents a different training data set, colored by instrument type. Black diamonds indicate instrument default settings. (A) Varying precursor tolerance from 5 to 50 ppm. (B) Varying fragment bin size from 0.02 to 1.005 Da. (C,D) Normalized PSM yield as a function of normalized error. Vertical axis measures normalized PSM yield at FDR 0.01: The natural log of the ratio of the yield at a given setting to the maximum yield at any setting. Horizontal axis measures parameter setting as the natural log of a multiple of the estimated standard deviation of measurement error. Gray lines represent individual experiments; blue line represents mean across all experiments. (C) Varying precursor tolerance. (D) Varying fragment bin size.

3.2. PSM Yield Variation between Parameter Settings

Some of our training experiments were much more sensitive to parameter settings than others. The extremes in difference in PSM yield between optimal and suboptimal settings for either parameter were quite high, with the worst and best parameter settings for precursor error yielding between 9 and 117% as many peptides as the instrument default settings and for fragment bin size yielding between 0 and 339% (Figure 3A,B). The relationship between parameter settings and PSM yield was not consistent within an instrument type, with, for instance, the two QExactive experiments having opposite trends in yield as a function of precursor error tolerance. These results further demonstrate that the values specified for precursor and fragment tolerances can have a sizable impact on PSM yield and that knowledge of instrument type alone is not sufficient to set those parameters optimally.

For fragment bin size, there was very close agreement between the experiments as to the optimal multiple of estimated error standard deviation (0.005). For precursor tolerance, the agreement was not as complete, with two experiments holding the most influence over the derived optimal multiple (37.40) due to their high sensitivity to changes in this parameter (Figure 3C,D). The lower level of agreement for precursor tolerance may reflect differences in the density of candidate precursor matches in the target and databases being searched against.

4. DISCUSSION

We have demonstrated that Param-Medic optimizes precursor error and fragment bin size parameter settings for LC−MS/MS search based on characteristics of the data set being searched. Param-Medic assumes that LC−MS/MS experiments are likely to make multiple observations of many peptide ions. Ironically, this phenomenon is often perceived as a chronic problem plaguing data-dependent acquisition proteomics: High-abundance peptides, in particular, will tend to trigger multiple MS/MS scans, leading to fewer acquisitions of other peptides. Accordingly, instrument makers and researchers often adjust a dynamic exclusion window to minimize these repeated measurements, but such measurements are nonetheless a constant feature of most proteomics experiments. Param-Medic exploits these repeated measurements to provide valuable information about the m/z tolerance characteristics of the experiment.

On several of our training and test data sets, Param-Medic increased PSM yield greatly over parameter settings chosen based on instrument type. Many researchers will spend time iteratively fine-tuning their search settings for a particular

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824

instrument over multiple experiments to maximize yield, a process that Param-Medic can assist with. In other circumstances, instrument-based parameter settings are used often, as when searching experimental data provided by collaborators or downloaded from a public repository, with minimal description. Param-Medic showed particularly large improvement over instrument defaults for one of the Orbitrap Elite training data sets. Neither the paper describing the data set nor the experimental metadata from PRIDE indicated whether the Orbitrap Elite was run in FT-FT mode (i.e., fragments analyzed in the orbitrap) or in FT-IT mode (i.e., fragments analyzed in the ion trap). Accordingly, we naïvely assumed the more-common (and higher-accuracy) FT-FT settings in our "instrument default" parameter settings. Upon further inspection, however, metadata in the mzML file for the acquisition indicated that the instrument was run in FT-IT mode. This setting likely accounts for the much higher yield at wider fragment bin settings and demonstrates that Param-Medic's error estimation can infer properties of the analysis that differ greatly from what might be expected from experimental metadata alone.

In our training and test data sets, Param-Medic settings yielded modestly more PSMs than settings chosen by experts for searching their own data for publication (52% more in one training data set). We do not know what criteria these authors used to choose the settings, and the settings may have behaved quite differently in their hands, using different search engines or values for parameters other than the two considered here. However, the consistency of the trend indicates that many laboratories may benefit from an approach to parameter setting that is based on the characteristics of the individual experiment being searched. The applicability of Param-Medic to Q-TOF data has particular potential to aid a subset of proteomics researchers. Some Q-TOF manufacturers write mass-corrected values in the raw data files, while others do not, leading many researchers to use a very wide and potentially suboptimal precursor tolerance in searching Q-TOF data.

In terms of PSM yield, Param-Medic performs very similarly to Preview on most data sets evaluated, with a large advantage in PSM yield in a single training experiment and nearly identical performance in our test experiments. (Supplementary Figure 1 compares the parameter estimates derived from Param-Medic and Preview on the training and test data sets.) Param-Medic and Preview each fail to assess error in different circumstances: Preview when its database search fails, Param-Medic when there are insufficient or suspicious differences in measurements available for error estimation. In our training and test data sets, Param-Medic refused to estimate error once, whereas Preview refused to estimate error on that same experiment and on one other experiment. An important difference between the tools is that Preview infers instrument calibration error in addition to measurement precision and so would presumably provide superior guidance for acquisitions with large calibration errors. On the other hand, Preview is proprietary software and runs only on Windows. Param-Medic is implemented in Python as a standalone tool and is also integrated into the Crux toolkit for streamlined parameter estimation and search with Comet and Tide search engines. In both incarnations, Param-Medic is open-source and can be run on Windows, Linux, and Mac. Furthermore, the Param-Medic running time is much shorter than that of Preview. Preview's running time scales with both the number of MS/MS spectra and the database size, whereas Param-Medic's running time scales only with the number of spectra. In practice, neither tool's running time is likely to be onerous, except possibly for Preview when the search database is large. This occurs often, for instance, in a metaproteomics context.

Although Param-Medic provides an estimate of ppm fragment error that could be used with any search engine, it currently only provides guidance for mapping this value to an appropriate fragment tolerance for search engines such as Comet, Sequest, and Tide that use fragment binning. Future work will include a reanalysis of the training data sets to provide such guidance for search engines that use fragment tolerances rather than fragment bins.

Param-Medic has been implemented as a standalone Python 2.7 tool that may be downloaded (including source code) at https://github.com/dhmay/param-medic or simply added to a Python installation with the "pip" tool. It has also been incorporated into version 3.1 of the Crux Toolkit, available at http://crux.ms. Within Crux, Param-Medic is available as a standalone tool and is also integrated into the Tide and Comet search algorithms for automatic detection of optimal parameter settings. All proteomics data sets described here, and links to all software, may be found at http://noble.gs.washington.edu/proj/param-medic/.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00028.

Supplementary Figure 1 comparing the parameter estimates derived from Param-Medic and Preview on the training and test data sets.(PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: william-noble@uw.edu.

ORCID

Damon H. May: 0000-0001-6902-3153

William S. Noble: 0000-0001-7283-4715

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program and by National Institutes of Health award P41 GM103533. We also acknowledge the contribution of a reviewer of our initial manuscript submission, who suggested the Q-TOF analysis described above.

■ REFERENCES

(1) Eng, J. K.; Mccormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(2) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73* (11), 2092−2123.

(3) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824

(4) Mortensen, P.; Gouw, J. W.; Olsen, J. V.; Ong, S.-E.; Rigbolt, K. T. G.; Bunkenborg, J.; Cox, J.; Foster, L. J.; Heck, A. J. R.; Blagoev, B.; et al. MSQuant, an Open Source Platform for Mass Spectrometry-Based Quantitative Proteomics research articles. *J. Proteome Res.* **2010**, *9*, 393−403.
(5) Kil, Y. J.; Becker, C.; Sandoval, W.; Goldberg, D.; Bern, M. Preview: A program for surveying shotgun proteomics tandem mass spectrometry data. *Anal. Chem.* **2011**, *83* (13), 5259−5267.
(6) Petyuk, V. A.; Mayampurath, A. M.; Monroe, M. E.; Polpitiya, A. D.; Purvine, S. O.; Anderson, G. A.; Camp, D. G.; Smith, R. D. DtaRefinery, a software tool for elimination of systematic errors from parent ion mass measurements in tandem mass spectra data sets. *Mol. Cell. Proteomics* **2010**, *9* (3), 486−96.
(7) Matthijnssens, J.; Ciarlet, M.; Rahman, M.; Attoui, H.; Bányai, K.; Estes, M. K.; Gentsch, J. R.; Iturriza-Gómara, M.; Kirkwood, C. D.; Martella, V.; et al. Elimination of Systematic Mass Measurement Errors in Liquid Chromatography-Mass Spectrometry Based Proteomics using Regression Models and a priori Partial Knowledge of the Sample Content. *Arch. Virol.* **2008**, *153* (8), 1621−1629.
(8) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open source tandem mass spectrometry sequence database search tool. *Proteomics* **2013**, *13* (1), 22−24.
(9) Wolski, W. E.; Farrow, M.; Emde, A.-K.; Lehrach, H.; Lalowski, M.; Reinert, K. Analytical model of peptide mass cluster centres with applications. *Proteome Sci.* **2006**, *4*, 18.
(10) Vizcaino, J. A.; Côté, R.; Reisinger, F.; Foster, J. M.; Mueller, M.; Rameseder, J.; Hermjakob, H.; Martens, L. A guide to the Identifications Database proteomics data repository. *Proteomics* **2009**, *9* (18), 4276−4283.
(11) Huttenhower, C.; Gevers, D.; Knight, R.; Abubucker, S.; Badger, J. H.; Chinwalla, A. T.; Creasy, H. H.; Earl, A. M.; Fitzgerald, M. G.; Fulton, R. S.; et al. Structure, function and diversity of the healthy human microbiome. *Nature* **2012**, *486* (7402), 207−214.
(12) May, Da. H.; Timmins-Schiffman, E.; Mikan, M. P.; Harvey, H. R.; Borenstein, E.; Nunn, B. L.; Noble, W. S. An alignment-free 'metapeptide' strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *J. Proteome Res.* **2016**, *15*, 2697.
(13) 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.
(14) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome A. *Nature* **2014**, *509* (7502), 575−581.
(15) Clark, D. J.; Fondrie, W. E.; Liao, Z.; Hanson, P. I.; Fulton, A.; Mao, L.; Yang, A. J. Redefining the Breast Cancer Exosome Proteome by Tandem Mass Tag Quantitative Proteomics and Multivariate Cluster Analysis. *Anal. Chem.* **2015**, *87* (20), 10462−10469.
(16) Radoshevich, L.; Impens, F.; Ribet, D.; Quereda, J. J.; Tham, T. N.; Nahori, M. A.; Bierne, H.; Dussurget, O.; Pizarro-Cerda, J.; Knobeloch, K. P.; et al. ISG15 counteracts Listeria monocytogenes infection. *eLife* **2015**, *4* (2015), 1−23.
(17) Tanca, A.; Palomba, A.; Pisanu, S.; Addis, M. F.; Uzzau, S. A human gut metaproteomic dataset from stool samples pretreated or not by differential centrifugation. *Data in Brief* **2015**, *4*, 559−562.
(18) Uszkoreit, J.; Maerkens, A.; Perez-Riverol, Y.; Meyer, H. E.; Marcus, K.; Stephan, C.; Kohlbacher, O.; Eisenacher, M. PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *J. Proteome Res.* **2015**, *14* (7), 2988−2997.
(19) Schittmayer, M.; Fritz, K.; Liesinger, L.; Griss, J.; Birner-Gruenberger, R. Cleaning out the Litterbox of Proteomic Scientists Favorite Pet: Optimized Data Analysis Avoiding Trypsin Artifacts. *J. Proteome Res.* **2016**, *15* (4), 1222−1229.
(20) Audain, E.; Uszkoreit, J.; Sachsenberg, T.; Pfeuffer, J.; Liang, X.; Hermjakob, H.; Sanchez, A.; Eisenacher, M.; Reinert, K.; Tabb, D. L.; et al. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J. Proteomics* **2016**, *150*, 170−182.
(21) Zhong, L.; Zhou, J.; Chen, X.; Lou, Y.; Liu, D.; Zou, X.; Yang, B.; Yin, Y.; Pan, Y. Quantitative proteomics study of the neuroprotective effects of B12 on hydrogen peroxide-induced apoptosis in SH-SY5Y cells. *Sci. Rep.* **2016**, *6*, 22635.

DOI: 10.1021/acs.jproteome.7b00028
J. Proteome Res. 2017, 16, 1817−1824