Published in final edited form as:

Rapid Commun Mass Spectrom. 2015 October 15; 29(19): 1841–1848. doi:10.1002/rcm.7272.

MZDASoft-A Software Architecture that Enables Large Scale Comparison of Protein Expression Levels over Multiple Samples Based on LC-MS/MS

Mehrab Ghanat Bari¹, Nelson Ramirez², Zhiwei Wang², and Jianqiu (Michelle) Zhang^{1,*}

¹Department of Electrical and Computer Engineering, Univ. of Texas at San Antonio, One UTSA Circle, San Antonio, TX 782

²Computational Biology Initiative, Univ. of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

Abstract

RATIONALE—Without accurate peak linking/alignment, only the expression levels of a small percentage of proteins can be compared across multiple samples in Liquid-Chromatography Mass Spectrometry/Tandem Mass Spectrometry (LC-MS/MS) due to the selective nature of tandem MS peptide identification. This greatly hampers biomedical research that aims at finding biomarkers for disease diagnosis, treatment, and the understanding of disease mechanisms. A recent algorithm, PeakLink, has allowed the accurate linking of LC-MS peaks without tandem MS identifications to their corresponding ones with identifications across multiple samples collected from different instruments, tissues and labs, which greatly enhanced the ability of comparing proteins. However, PeakLink cannot be implemented practically for large number of samples based on existing software architectures, because it requires access to peak elution profiles from multiple LC-MS/MS samples simultaneously.

METHODS—We propose a new architecture based on parallel processing, which extracts LC-MS peak features, and saves them in database files to enable the implementation of PeakLink for multiple samples. The software has been deployed in High Performance Computing (HPC) environments. The core part of the software, MZDASoft Parallel Peak Extractor (PPE), can be downloaded with users and developer's guide, and it can be run on HPC centers directly. The quantification applications, MZDASoft TandemQuant and MZDASoft PeakLink are written in Matlab, which are compiled with Matlab runtime compiler. A sample script that incorporates all necessary processing steps of MZDASoft for LC-MS/MS quantification in a parallel processing environment is available. The project webpage is <http://compgenomics.utsa.edu/zgroup/MZDASoft>.

RESULTS—The proposed architecture enables the implementation of PeakLink for multiple samples. Significantly more (100%–500%) proteins can be compared over multiple samples with better quantification accuracy in test cases.

*To whom correspondence should be addressed, michelle.zhang@utsa.edu.

CONCLUSION—MZDASoft enables large scale comparison of protein expression levels over multiple samples with much larger protein comparison coverage and better quantification accuracy. It is an efficient implementation based on parallel processing which can be used to process large amount of data.

INTRODUCTION

Protein level analysis is important for understanding the underlying biological processes in biomedical and clinical research. [1,2] Such research demands the ability to compare large number of protein expression levels over multiple samples, which is still a challenge in Liquid Chromatography-Mass Spectrometry/Tandem Mass Spectrometry (LC-MS/MS) based methods despite their wide application. Tandem MS provides reliable elution time and sequence information of identified peptides, [3] based on which, one can find their corresponding peaks in LC-MS for quantification. However, since tandem MS only samples and identifies a small fraction of total peptide species in a run, only a small fraction of peptides can be quantified with reliable identification per run, [4] and most peptides cannot be identified simultaneously in multiple runs. Consequently, only a small number of commonly identified proteins are quantified and can be compared across multiple samples. In this paper, we define *protein comparison coverage* as the total number of proteins quantified across all samples in an experiment.

Chemical isobaric tagging (including iTRAQ and TMT) solves this problem partially by allowing 8 or more multiplexing channels, such that 8 or more samples can be compared simultaneously. However, isobaric tagging suffers from severe dynamic-range compression and reduced quantitative accuracy due to precursor interference. [5,6] In addition, when the multiplexing capacity is reached, protein comparison coverage will reach its limit. Solutions based on spectral counting in label free tandem MS, [7,8] suffer from large quantification variations. [9,7,10] Recently, Data-independent acquisition (DIA) [11,12,13,14] has received a lot of interest, which fragments all peptides and generates reproducible peptide fragment maps. Although the total number of peptide identifications of DIA is similar to that of tandem MS per run, [15] the protein comparison coverage over several runs can be increase significantly. However, DIA is still limited by the intersection of all identified peptides across multiple runs.

LC-MS registers the signal of all sufficiently ionized peptides, and theoretically, it is possible to link peptide LC-MS peaks that have tandem MS identifications in some runs to their corresponding ones without identifications in other runs, so that the union instead of the intersection of all identified peptides in an experiment can be quantified and compared over multiple samples. In the past, this idea has been explored extensively. Earlier reviews that describe different algorithms for peak linking are available in [16,17]. Since then, there has been more publications including XCMS, [18] MZmine2, [19] MultiAlign, [20] SuperHirn, [21] ChromAlign, [22] OpenMS, [23,24] SIMA, [25] and SMFM/SMFM-g, [26]. Most of these algorithms are warping based correspondence algorithm (WCA), in which, the elution time shifts between two LC-MS/MS runs are corrected by finding a warping (alignment) function, and peaks are linked based on their closeness in retention

time. However, very few of these algorithms utilize MS/MS information for generating warping functions, [17] and none of them uses peak shape information for peak linking.

Most recently PeakLink [27], has been developed for peak linking using peak shape information and SVM classification. Wavelet transform, [28,29] is used to obtain peak shape correlation, which it is coupled with time information and composed the input of the SVM classifier. Its main advantage is that it provide consistent and better performance in challenging cases when samples collected from different tissues, labs and instruments are required for comparison. WCAs on the other hand can only perform well in some cases. PeakLink could increase protein comparison coverage significantly.

However, given that LC-MS datasets are usually in the giga- or tera- byte (GB or TB) range, and there exists millions of LC-MS peaks in each dataset, no existing software architecture can accommodate the practical implementation PeakLink, which requires access to LC-MS peak elution profiles from multiple samples simultaneously. The promised gain in protein comparison coverage by PeakLink cannot be realized practically. For example, the popular software package MaxQuant, [30] and Trans-Proteomics Pipeline (TPP), [31] do not output LC-MS peak features required by PeakLink. Recently, OpenMS has been updated to allow the output of peak features for feature alignment. [24] However, OpenMS only outputs very few peak parameters, and does not output elution profiles required by PeakLink. Another parallel processing architecture, PP-VLAM, [32] was not designed for linking peptides, and does not output enough peak features.

We propose a novel and flexible software architecture, MZDASoft, to address the key challenges in implementing PeakLink for multiple sample comparison. The core of the architecture, which is shown in Figure 1, is a stand-alone software module called “MZDASoft Parallel LC-MS Peak Extractor (PPE)” that runs on Texas Advanced Computing Center (TACC) (<http://www.tacc.utexas.edu>), or on local computer clusters (for example, the Cheetah Cluster at <http://www.cbi.utsa.edu>). The architecture is designed to be flexible in supporting a variety of High Performance Computing (HPC) resource managers. MZDASoft has the following characteristics.

1. **Parallelism and scalability:** MZDASoft supports efficient processing because the time consuming step, feature extraction, is processed in parallel, which can utilize as many systems and cores as available. Users can specify the total number of cores. It uses a shared file system to support multiple threads in parallel processing. It finishes the processing of 134G of data within 8 hours in a tested case.
2. **Easy data access:** MZDASoft PPE saves extracted features with enough information to reconstruct 2D LC-MS peaks in range query optimized database files (SQLite3). [33] Various applications can query these database files directly without repeating the feature extraction process, and thus it supports fast and recursive research cycles. SQLite3 can be accessed from a plethora of operating systems and programming languages/environments. In particular, we provide an access methodology from Matlab in both Windows and Linux operating systems. None of the existing software provide easily accessible and complete peak feature information. They either do not output extracted features (MaxQuant, TPP), or the

extracted features are stored in XML formats (OpenMS), which requires special parsing programs that slow down the process.

3. **Generality:** LC-MS feature extraction is not application specific. PPE extracts features without requiring knowledge of the analyte, and it can be readily used in various applications based on LC-MS. In contrast, features are extracted after deisotoping in, [32] and the pre-processing part needs to be readjusted for different labeling methods.
4. **High Performance Computing (HPC) support:** The architecture has been implemented on a HPC center, TACC, which is accessible to most academic users in Texas. Similar centers exist in other states. Through implementation on central and large computing grids, the cost related to establishing and maintaining local computing clusters is avoided.
5. **Novel algorithm development support:** The architecture adopts a layered structure and employs independent modules. It also provides visualization tools for algorithm development. Thus the architecture can be easily updated and modified for novel applications. In contrast, in past parallel implementation, [32] features are summarized after deisotoping, and important feature information such as elution profile are not provided. Thus we cannot create new applications by developing new modules. We have developed novel application modules based on the MZDASoft architecture: MZDASoftQuant which includes two parts: MZDASoft TandemQuant and MZDASoft PeakLink for stable isotope labeling by amino acids (SILAC) data. [34] MZDASoft TandemQuant quantifies tandem MS identified peptides. Quantification results based on different tandem MS search engines can be integrated to boost the total protein comparison coverage. TandemQuant also provides a series of statistics required by MZDASoft PeakLink.

The novel aspect of MZDASoft is that it supports data integration over multiple LC-MS/MS samples. Through such an architecture, PeakLink becomes implementable, which provides better protein comparison coverage and quantification accuracy. Without the proposed architecture and parallel implementation, clinical and biomedical researchers would have limited ability in comparing samples collected from different disease conditions.

MATERIAS

The Algorithm Design of MZDASoft Parallel Peak Extractor (PPE)

PPE conducts data processing at three levels. At level one, raw LC-MS files are partitioned into multiple sections given rt and m/z window sizes specified by users. Subsequently, each section is further partitioned into Pre-LC region of interests (PLCROI). At level 2, LC peak candidates (LCC)s are extracted from each PLCROI, and at level three, all LCCs are merged together and saved in database files. PLCROI processing are explained in detail in the supplementary file.

PPE output data structure

For each LC-MS peak, MZDASoft PPE extracts several features as listed in Table 1. Based on these features, we can reconstruct the 2D LC-MS peaks. Among these features, elution profile is highly valuable because elution profiles of a peptide at different charge states and isotope positions are identical, and by matching elution profiles, we can eliminate interfering signals when quantifying peptides. In addition, elution profile shape is influenced by physiochemical properties of peptides, which is consistent between different LC-MS runs, which is used for linking peptide peaks. [35] Note that most other pipelines, such as OpenMS, [24] do not retain elution profile information.

C++ interface for querying PPE database files from Matlab

The SQLite database files, [33] store all features extracted from mzXML files together with a special index that allows fast retrieval of data using range queries based on m/z and rt bounding boxes. SQLite is an in-process library that implements a self-contained, server-less, zero-configuration, and transactional SQL database engine. Access to SQLite database files is possible from a variety of high level programming languages. Since the example applications are developed in Matlab, the matlab-sqlite3-driver open source project is used which is available at <https://github.com/kyamagu/matlab-sqlite3-driver/tree/master/src>. We further develop a custom tool that converts MZDASoft .dat text files to standard range query optimized SQLite3 database files.

MZDASoftQuant for SILAC Data

To achieve the goal of boosting protein comparison coverage significantly, we developed MZDASoftQuant, which interacts with MZDASoft PPE and tandem MS search results. MZDASoftQuant consists three parts: Time scale conversion, MZDASoft TandemQuant and MZDASoft PeakLink. In time scale conversion, a common time scale is established, and warping functions between the common and local time scales in each LC-MS files are extracted. MZDASoft TandemQuant aims at quantifying tandem MS identified peptides over multiple files simultaneously. By using tandem MS identified peptides as the ground truth, it extracts statistical information on elution time shifts, elution peak shape correlations, and the isotope distribution matches between corresponding and non-corresponding peptide features in different LC-MS runs. Subsequently, MZDASoft Peaklink links peptide features unidentified by tandem MS in certain LC-MS runs to their corresponding ones in identified runs, which is the key for boosting protein comparison coverage. We implemented the quantification modules for Stable Isotope Labeling of Amino Acid in Cell Cultures (SILAC) [36] because it's high quantification accuracy. Quantification of other labeling methods can be realized by slight modification of the quantification module.

Time scale conversion

Due to changes in the elution gradient, the retention time scales are always different in different LC-MS runs. This creates difficulties in linking peaks of identical peptides in different runs. We have to establish a common time scale for LC-MS/MS runs in the same fraction, and find forward and backward warping functions for the conversion of peptide retention times between the local and common scales. Given a peptide with tandem MS

identifications in some runs, the goal of this step is to provide estimated retention time of the peptide in runs without tandem MS identifications, and estimate the 99% confidence window.

To accomplish this goal, the average of tandem MS reported peptide retention time in the common time scale will be set as the predicted retention time. Then this predicted retention time will be converted to predicted local retention times. A flow diagram of this process and the details of are given in the supplementary file.

MZDASoft TandemQuant

The goal of MZDASoft TandemQuant is to find and quantify LC-MS peaks of tandem MS identified peptides, and provide statistics such as retention time shift and peak shape correlations between tandem MS identified corresponding features. Tandem MS identified corresponding features are LC-MS peak features linked to the same peptide sequence reported by tandem MS in different LC-MS runs. The flow diagram of TandemQuant and the detailed explanations are given the supplementary file.

Through MZDASoft TandemQuant, not only we obtain quantification of tandem MS identified peptides, but also elution time signal templates of peptides; samples of elution time shifts and peak shape correlations between corresponding/non-corresponding features; and the matching of isotope distributions of corresponding/non-corresponding features from measured ones to their theoretically predicted ones. Subsequently, these are as inputs for the next stage of processing called MZDASoft PeakLink. During quantification processing, we performs boundary detection, matching of theoretically predicted isotope patterns, and the matching of elution profiles to ensure quantification quality.

MZDASoft PeakLink

In MZDASoft PeakLink, peptide LC-MS peaks that are not identified by tandem MS in some runs can be linked to their corresponding signal templates with tandem MS identifications. It is extended from the PeakLink algorithm. [27] While PeakLink is designed for pairwise peak linking, MZDASoft PeakLink considers multiple LC-MS samples. The flow diagram and details are provided in the supplementary file. At the end of the process, a peptide quantification map will be produced for each fraction. The rows of the map represent peptides in the union list, and the columns of the map represent different LC-MS runs.

Testing Datasets

We use two datasets for performance testing. Testing dataset one (Yeast Dataset), is obtained via private communication with Dr. J. Coon's lab. These data are labeled with pre-mixed heavy to light ratios (HLR) at 10:1, 5:1, 2:1, 1:2, and 1:1. The 5:1, 2:1, and the 1:1 samples are provided with two technical replicates. [37] There are a total of 8 files. This dataset provides the chance to evaluate quantification coverage and accuracy, because the measured HLRs can be compared to the pre-mixed HLRs directly.

Testing dataset two contains 5 sets of 6 files (SuperSILAC Dataset) in mzxml format (134G in total) gathered from breast cancer research using Super-SILAC labeling. [34] The dataset was originally uploaded to the Tranche database, and it has been moved to <ftp://MSV000074502:a@massive.ucsd.edu>. We downloaded the dataset which is heavy labeled with a super-SILAC mix, and mixed with: normal tissues, ductal tumors, lobular tumors, Astrocytoma protoplasmaticum, and breast tumor. The filenames of the dataset is provided in the supplementary file. These represent a set of typical samples collected in cancer studies. Please refer to the original paper for detailed descriptions of these datasets. Six fractions are collected for this dataset, and within each fraction, there are technical replicates. We selected one technical replicate for the testing of quantification coverage, because with PeakLink, the union of peptide identification is very large, and one technical replicate provides significant amount of peptide identifications already. This dataset is valuable in providing an estimation of protein comparison coverage in a realistic setting and the processing speed.

Tandem MS Search and Tandem MS Results Format Conversion

The proposed architecture can be used for the quantification of identified peptides from many search tandem engines. For evaluation purpose, we select the widely available and free search engine X!Tandem [38] embedded in the trans-proteomics-pipeline (TPP) [31] for tandem MS search. The tandem MS search results are subsequently reorganized into a uniform data structure in the Matlab .mat file format. The details of the data structure and tandem MS search parameters can be found in the supplementary file.

Given these tandem MS search results, which are organized in the specified format, we can import them and use Matlab to send queries to the database files generated by PPE for quantification.

Combination and reporting of quantification results from different fractions

MZDASoftQuant consists of two main processing steps: MZDASoft TandemQuant and MZDASoft PeakLink. The two steps will produce two quantification maps for tandem MS identified peptides and peptides that are quantified after PeakLink. The procedure are repeated for each fraction, and multiple peptide quantification maps from the two steps are combined to generate an overall quantification map.

At the beginning of this combination process, a union peptide list across all samples and fractions is generated by combining the peptide list in each fraction and sample. Duplicated peptide ids are eliminated. Then a MZDASoft TandemQuant map is generated for each fraction by quantifying all tandem MS identified peptides. The rows of the map correspond to different peptides, and the columns of the map corresponds to different samples. Subsequently, MZDASoft PeakLink is applied to link all corresponding peptides in each fraction. Linked peptides are then quantified in a map with the same number rows and columns as that of the MZDASoft TandemQuant map. These two maps have no overlapping entries and they are unionized to generate the overall peptide quantification map for each fraction.

Protein quantification is derived based on the unionized peptide list. If a protein is identified through multiple peptide entries scattered in different fractions, the median value of the HLRs of these peptides are reported as the HLR of the protein. The software also reports protein quantification results based on unique peptides. Protein and peptide quantification results are printed in two comma separated files. One file reports protein quantification results. Each protein is listed as an entry, and is followed by the number of peptides/unique peptide identified in all files, then the HLR of the protein in each technical replicate is reported. The second report is an expanded version of the first one, in which, peptide sequences and their quantification information associated with each protein are listed.

RESULTS

We test MZDASoftQuant in terms of protein comparison coverage, quantification accuracy, processing speed, and data storage requirements. MaxQuant is selected as the benchmark software because of its superior performance in quantification accuracy and coverage and its wide usage in the field. [7]. While there exists many software packages that perform peak linking such as IDEAL-Q, [39] and Open-MS, they are mostly designed for label-free data, and are not directly comparable to MZDASoft and MaxQuant in quantification coverage and accuracy for SILAC data. Those software packages can be compared with PeakLink in terms of peak linking accuracy, and the the superior performance of PeakLink has been established before.

MaxQuant does not provide the functionality of linking LC-MS peak in SILAC data processing, and only common tandem MS identified peptides can be compared across multiple runs. The results are organized and described as the following.

Protein Comparison Coverage

The goal of building MZDASoft is to increase protein comparison coverage. We compare the total number of proteins that are quantified across 8 Yeast dataset between MZDASoftQuant and MaxQuant to demonstrate the efficacy of peak linking. To ensure fair comparison, we count the number of proteins in the same way in both algorithms. We use the yeast protein name in the fasta file as the index for counting without grouping isoforms. Secondly, we evaluated the false positive rate (FPR) of peptide identification after combining 8 Yeast datasets based on decoy peptides, since MaxQuant uses decoys to estimate the FPR.

Theoretically, unionizing tandem MS identification results across multiple files could increase the FPR slightly. We find a total of 3 decoys out of 2835 peptides in the merged peptide list across 8 files when the PeptideProphet probability threshold is set at 90%. We subsequently counted the total number of decoyed peptides without the threshold on PeptideProphet score. A total of 1560 decoy peptides is counted, and the FPR is estimated as 0.19% by dividing the total number of decoys in the union list with the total number of decoys when no filtering of peptides is performed based on TPP search results. This FPR is lower than the FPR of MaxQuant on a per-file basis that we set at 1% when we run MaxQuant. In our implementation, out of 2835 tandem MS identified peptides across 8 datasets, 816 or 28.78% is not quantifiable in any dataset using our quantification standards,

because they have isotope patterns that mismatch with the theoretically predicted ones. In MaxQuant, the matching to the theoretically predicted isotope pattern is not performed, and most identified peptides are quantified. MaxQuant have a higher per file FPR, and it quantifies most peptides that it detects, thus our comparison in protein comparison coverage is more favorable to MaxQuant. In Figure 2, we show the total number of proteins quantified across 1 to 8 samples in the Yeast dataset of MZDASoftQuant based on X!Tandem search results. We can see that protein comparison coverage is increased from 178 to 362 by MZDASoftQuant. We also show the protein comparison coverage of MZDASoft TandemQuant and MZDASoft PeakLink separately. We can see that by combining MZDASoft PeakLink and MZDASoft TandemQuant, the comparison coverage improves significantly over using MZDASoft TandemQuant alone. We further analyzed the case when both the TPP and Andromeda search engines are used to identify peptides. Identified peptides are subsequently linked through PeakLink before quantification. This time 768 overlapping proteins has been quantified in all 8 files which is a **4.3** fold increase in overlapping protein quantification coverage.

The Super-SILAC dataset has 30 files in total, and we performed a test run on TACC using the search engine, X!Tandem. In Figure 3, we plot the total number of proteins that can be compared across and across 1 to 5 samples. We can see that MZDASoftQuant dramatically increases protein comparison coverage from under 1000 to more than 6000 proteins across 5 samples comparing to MaxQuant. This shows that PeakLink can significantly improve protein quantification coverage in this case.

Quantification Accuracy

Our main goal is to allow higher protein comparison coverage across multiple samples. In this section, we investigate if the increase of the coverage leads to the degradation in quantification accuracy. The boxplot of measured HLRs in the Yeast dataset are presented in Figure 4 for all measured proteins as well as for proteins that has been quantified in all 8 samples. We can see that the measured ratios of MZDASoft are closer to the pre-defined ratios than that measured by MaxQuant in both cases. However, there is an increased variance in measured ratios. (Note that MZDASoft quantifies more proteins than MaxQuant, and there also exists more outliers in the box plot.) This brought up an important question: what constitutes a fair comparison when different algorithms report different bias and variance? Which statistical measure counts? Unfortunately, various statistical measures that are correlated with bias and variance exists in the field and there is a lot of confusion in which statistics should be used for algorithm comparison.

Obviously, both bias and variance matters in quantification algorithms, especially, consider that the most important application of these algorithms is to distinguish differentially expressed proteins. The higher bias in MaxQuant reduces the difference between different ratios, while the higher variance in MZDASoft also leads to the same effect. Based on any single bias or variance measure, one cannot conclude if one algorithm is better than the other. Thus, we further compare the mean square error (MSE) between the log of measured ratios and the log of predefined ratios based on the Yeast dataset. The mean square error is the summation of the squared bias and variance, and the results are summarized in the Table

2. We can see that MZDASoftQuant reduces MSE significantly especially for proteins that are quantified across all 8 samples using one search engine with a MSE of 0.30, and a protein comparison coverage of 362. In contrast, MaxQuant has a MSE of 0.43 while the protein comparison coverage is 178. When two search engines are used in MZDASoftQuant, the protein comparison coverage is increased to 768, while the MSE is at 0.41, which is still lower than that of MaxQuant. The smaller MSE leads to a better separation between different measured ratios, and it shall be taken as the standard statistics for comparing different algorithms.

These testing results show that protein comparison coverage over MaxQuant is not gained at a cost of quantification accuracy, and there is an overall improvement instead. The improvement of MZDASoft over MaxQuant is gained by using more stringent quantification criteria and employing boundary detection techniques. [40]

Processing Speed and Data Storage Requirement

There are many possible scenarios in using MZDASoft. In the tested scenario, mzXML files are first submitted to TACC installed with TPP for peptide identification. Then the mzXML files are uploaded to TACC to be processed by PPE. Finally, MZDASoftQuant, including MZDASoft TandemQuant and MZDASoft PeakLink are employed to perform quantification either at TACC or on a local PC. The total time for processing the SuperSILAC, and the Yeast dataset is listed in Table 3.

A direct and fair comparison of running time to the parallel version of MaxQuant is not possible because MaxQuant does not provide the peak linking functionality. The non-parallel version of MaxQuant requires around 1–3 hours to process each of the raw data files, and if it is employed for processing the SILAC dataset, it requires around 90 hours of total processing time on a local Dell computer T7500.

CONCLUSION

In this paper, we propose a novel LC-MS/MS processing architecture which separates the most time consuming task of feature extraction from other application dependent steps. The core of the architecture: MZDASoft ParallelPeakExtractor is implemented in parallel which enables fast processing of large scale LC-MS/MS based studies with hundreds Giga bytes of data from multiple patients/time points/conditions. The new architecture not only allows fast processing, it also stores extracted LC-MS features in database files that can be queried easily and repetitively. Consequently, the most accurate peak linking algorithm PeakLink can be implemented which significantly boosts the total number of proteins that can be compared across multiple samples. In the tested cases, the expression levels of 100%–500% more proteins can be compared across multiple runs with better accuracy than that of MaxQuant. The example application is developed for SILAC, which can be easily modified for other LC-MS labeling/label-free methods. It is expected that the new architecture will allow accelerated research and development of fast algorithms that consider multiple LC-MS/MS samples together.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was supported by a grant from the National Institute on Minority Health and Health Disparities (G12MD007591) from the National Institutes of Health. We thank the Computational Biology Initiative (UTSA/UTHSCSA) for providing access and training to the analysis software used.

References

1. Tyers, Mike; Mann, Matthias. From genomics to proteomics. *Nature*. 2003; 422(6928):193–197. [PubMed: 12634792]
2. Vaudel, Marc; Sickmann, Albert; Martens, Lennart. Current methods for global proteome identification. 2012
3. Nesvizhskii, Alexey I.; Aebersold, Ruedi. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug discovery today*. 2004; 9(4):173–181. [PubMed: 14960397]
4. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*. 2007; 389(4):1017–1031. [PubMed: 17668192]
5. Ow, Saw Yen, et al. iTRAQ underestimation in simple and complex mixtures. *Journal of proteome research*. 2009; 8(11):5347–5355. [PubMed: 19754192]
6. Christoforou, Andy; Lilley, Kathryn S. Taming the isobaric tagging elephant in the room in quantitative proteomics. *Nature methods*. 2011; 8(11):911–913. [PubMed: 22036744]
7. Trudgian, David C., et al. Comparative evaluation of label-free SING normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics*. 2011; 11(14):2790–2797. [PubMed: 21656681]
8. Egelhofer, Volker; Hoehenwarter, Wolfgang; Lyon, David; Weckwerth, Wolfram; Wienkoop, Stefanie. Using ProtMAX to create high-mass-accuracy precursor alignments from label-free quantitative mass spectrometry data generated in shotgun proteomics experiments. *Nature protocols*. 2013; 8(3):595–601. [PubMed: 23449253]
9. Li, Zhou, et al. Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. *Journal of proteome research*. 2012; 11(3):1582–1590. [PubMed: 22188275]
10. Zhang Z. Automated precursor ion exclusion during LC-MS/MS data acquisition for optimal ion identification. *Journal of The American Society for Mass Spectrometry*. 2012; 23(8):1400–1407. [PubMed: 22669759]
11. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods*. 2004; 1(1):39–45. [PubMed: 15782151]
12. Egertson JD, et al. Multiplexed MS/MS for improved data-independent acquisition. *Nature methods*. 2013; 10(8):744–746. [PubMed: 23793237]
13. Doerr, Allison. DIA mass spectrometry. *Nature methods*. 2015:35–35.
14. Avtonomov CC, Tsou D, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods*. 2015; 12(3):258–264. [PubMed: 25599550]
15. Blackburn, Kevin; Mbeunkui, Flaubert; Mitra, Srijeet K.; Mentzel, Tobias; Goshe, Michael B. Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. *Journal of proteome research*. 2010; 9(7):3621–3637. [PubMed: 20450226]
16. Lange E, Tautenhahn R, Neumann S, Gröpl C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC bioinformatics*. 2008; 9(1):375. [PubMed: 18793413]

17. Vandenbogaert M, et al. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics*. 2008; 8(4):650. [PubMed: 18297649]
18. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry*. 2006; 78(3):779–787. [PubMed: 16448051]
19. Pluskal T, Castillo S, Villar-Briones A, Orešić M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*. 2010; 11(1):395. [PubMed: 20650010]
20. LaMarche, Brian L., et al. MultiAlign: a multiple LC-MS analysis tool for targeted omics analysis. *BMC bioinformatics*. 2013; 14(1):49. [PubMed: 23398735]
21. Mueller LN, et al. SuperHirn—a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*. 2007; 7(19):3470–3480. [PubMed: 17726677]
22. Sadygov RG, Martin Maroto F, Hühner AF. ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Analytical chemistry*. 2006; 78(24):8207–8217. [PubMed: 17165809]
23. Sturm M, Bertsch A, et al. OpenMS -- An open-source software framework for mass spectrometry. *BMC bioinformatics*. 2008; 9(1):163. [PubMed: 18366760]
24. Weisser, Hendrik, et al. An automated pipeline for high-throughput label-free quantitative proteomics. *Journal of proteome research*. 2013; 12(4):1628–1644. [PubMed: 23391308]
25. Voss B, et al. SIMA: Simultaneous Multiple Alignment of LC/MS Peak Lists. *Bioinformatics*. 2011
26. Lin, Hao; He, Lin; Ma, Bin. A combinatorial approach to the peptide feature matching problem for label-free quantification. *Bioinformatics*. 2013; 29(14):1768–1775. [PubMed: 23665772]
27. Bari, Mehrab Ghanat; Ma, Xuepo; Zhang, Jianqiu. PeakLink: A new peptide peak linking method in LC-MS/MS using wavelet and SVM. *Bioinformatics*. 2014:btu299.
28. Ghanatbary M, Mehridehnavi AR, Rabbani H, Mahoori AR, Mehrjoo M. A comparative study of the output correlations between wavelet transform, neural and neuro fuzzy networks and BIS index for depth of anesthesia. *Industrial Electronics & Applications (ISIEA)*. 2010:655–659.
29. Rabbani, H.; Dehnavi, AM.; Ghanatbary, M. Artificial Neural Networks-Methodological Advances and Biomedical Applications. Rijeka, Croatia: intech; 2011. Estimation the Depth of Anesthesia by the Use of Artificial; p. 283-302.
30. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*. 2008; 26(12):1367–1372.
31. Deutsch EW, et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics*. 2010; 10(6):1150–1159. [PubMed: 20101611]
32. Vander Burgt YEM, et al. Parallel processing of large datasets from NanoLC-FTICR-MS measurements. *Journal of the American Society for Mass Spectrometry*. 2007; 18(1):152–161. [PubMed: 17055738]
33. Newman, Chris. SQLite (Developer's Library). Sams; 2004.
34. Geiger, Tamar; Cox, Juergen; Ostasiewicz, Pawel; Wisniewski, Jacek R.; Mann, Matthias. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature methods*. 2010; 7(5):383–385. [PubMed: 20364148]
35. Cui J, Ma X, Chen L, Zhang J. SCFIA: a statistical corresponding feature identification algorithm for LC/MS. *BMC bioinformatics*. 2011; 12(1):439. [PubMed: 22078262]
36. Ong, Shao-En, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics*. 2002; 1(5):376–386. [PubMed: 12118079]
37. Hebert, Alexander S., et al. Neutron-encoded mass signatures for multiplexed proteome quantification. *Nature methods*. 2013; 10(4):332–334. [PubMed: 23435260]
38. Craig, Robertson; Beavis, Ronald C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 2004; 20(9):1466–1467. [PubMed: 14976030]

39. Tsou CC, et al. IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Molecular & Cellular Proteomics*. 2010; 9(1):131–144. [PubMed: 19752006]
40. Cui, Jian, et al. Accurate LC Peak Boundary Detection for ¹⁶O/¹⁸O Labeled LC-MS Data. *PloS one*. 2013; 8(10):e72951. [PubMed: 24115998]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
The MZDASoft architecture with Parallel Peak Extractor (PPE) at its core.

Figure 2.

Protein comparison coverage of the Yeast dataset based on X!Tandem and Andromeda.

Figure 3.
Protein comparison coverage of the Super-SILAC dataset based on X!Tandem.

Figure 4.

Table 1

Data structure of extracted LC-MS peak features

Field Name	Meaning
<i>recordid</i>	id for database search
<i>centermz</i>	m/z value at the peak apex
<i>centerrt</i>	rt value at the peak apex
<i>startmz</i>	peak boundary: start in m/z dimension
<i>endmz</i>	peak boundary: end in m/z dimension
<i>deltamz</i>	m/z grid size
<i>startrt</i>	peak boundary: start in rt dimension
<i>endrt</i>	peak boundary: end in rt dimension
<i>deltart</i>	rt grid size
<i>mzprofile</i>	m/z dimension peak intensities
<i>elutionprofile</i>	rt dimension peak intensities
<i>mzlength</i>	number of samples in m/z dimension
<i>rtlength</i>	number of samples in rt dimension

Table 2

Mean Square Error of log measured ratios

MSE at Ratio	10:1	2:1	1:1	1:2	1:5	mean	#Proteins
MaxQuant	2.17	0.61	0.51	0.43	0.45	0.83	1103
MZDASoft	1.98	0.43	0.30	0.44	0.84	0.80	751
MZDASoft(Two Search Engines)	2.03	0.54	0.28	0.46	0.90	0.84	1245
MaxQuant Overlapping	1.07	0.43	0.25	0.23	0.19	0.43	178
MZDASoft Overlapping (One Search Engine)	0.87	0.13	0.14	0.11	0.27	0.30	362
MZDASoft Overlapping(Two Search Engines)	1.05	0.23	0.12	0.15	0.46	0.41	768

Table 3

Summary of Running Time (Hours) and Data size

Processing step	Yeast		SuperSILAC	
	TACC	PC	TACC	PC
PPE	0.65	4	2	24
TandemQuant/PeakLink	2	7	6	168
Total time	2.65	11	8	192
Data size	8Gb		134Gb	