Journal of proteome research

ARTICLE

pubs.acs.org/jpr

MSSimulator: Simulation of Mass Spectrometry Data

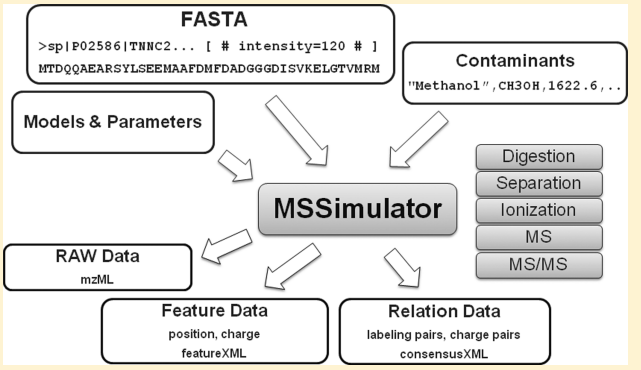
Chris Bielow,[*,†,‡,§] Stephan Aiche,[†,*,‡,§] Sandro Andreotti,[‡,§] and Knut Reinert[‡]

[‡]Institute of Computer Science, Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany

[§]International Max Planck Research School for Computational Biology and Scientific Computing, Berlin, Germany

Supporting Information

ABSTRACT: Mass spectrometry coupled to liquid chromatography (LC−MS and LC−MS/MS) is commonly used to analyze the protein content of biological samples in large scale studies, enabling quantitation and identification of proteins and peptides using a wide range of experimental protocols, algorithms, and statistical models to analyze the data. Currently it is difficult to compare the plethora of algorithms for these tasks. So far, curated benchmark data exists for peptide identification algorithms but data that represents a ground truth for the evaluation of LC−MS data is limited. Hence there have been attempts to simulate such data in a controlled fashion to evaluate and compare algorithms. We present MSSimulator, a simulation software for LC−MS and LC−MS/MS experiments. Starting from a list of proteins from a FASTA file, the simulation will perform in-silico digestion, retention time prediction, ionization filtering, and raw signal simulation (including MS/MS), while providing many options to change the properties of the resulting data like elution profile shape, resolution and sampling rate. Several protocols for SILAC, iTRAQ or MS^E are available, in addition to the usual label-free approach, making MSSimulator the most comprehensive simulator for LC−MS and LC−MS/MS data.



KEYWORDS: Mass Spectrometry, Simulation, Benchmarking, Ground Truth, SILAC, iTRAQ, MS^E

INTRODUCTION

In mass spectrometry (MS) based proteomics, often proteins in a sample are digested and the resulting peptides are separated by high-performance liquid chromatography (LC) before injecting them into the mass spectrometer. Subsequently, data can be obtained in two modes, the LC−MS mode, in which continuous sampling over the whole mass range occurs and which is used solely for quantitation, and the LC−MS/MS mode where a fragmentation of selected sample compounds is performed to obtain ion ladders that can be used for the identification of the compound.

Modern mass spectrometers can easily generate thousands of mass spectra in a short time. This ability has been an incentive for the development of new experimental protocols and new fully automated methods to analyze the resulting data. For the creation of efficient and robust methods, developers of new algorithms need benchmark data to compare their approach to existing ones or to assess the robustness of their algorithm to different kinds of data (e.g., another MS machine, more background noise, or more complex samples). This is a difficult task, since carefully compiled databases of annotated test data are scarce in mass spectrometry-based proteomics. An ideal LC−MS data set for the evaluation of feature detection, alignment and quantitation algorithms would contain annotations with the positions of all peptide ion signals, their charge states, monoisotopic masses and abundances. Only this information would allow meaningful comparisons between different methods and fair benchmark studies. If in addition one could alter certain parameters (like the background noise or the type of MS machine), it would be possible to assess the robustness of the newly developed algorithm.

One approach to overcome this problem is using simulated data sets. This idea is not new and was already used by Morris et al.[1] in 2005 to benchmark their new approach for feature extraction and quantitation by validating it on a data set simulated by the Cromwell software presented by Coombes et al.[2] In 2008 Schulz-Trieglaff et al.[3] presented a comprehensive approach to simulate LC−MS data and used it to benchmark different feature detection approaches. Renard et al.[4] implemented a quite simple simulation approach to validate the NITPICK feature finding algorithm. In 2009 Yang et al.[5] used a simulated data set from Morris et al.[1] to benchmark different peak picking algorithms.

However, those approaches were focusing mostly on one particular aspect of a simulation. In this paper we present MSSimulator, a comprehensive simulator for LC−MS and LC−MS/MS data that includes all functionalities of the so far most comprehensive tool LC−MSsim and extends it in many respects. In the following part we will describe the basic steps which can be simulated with MSSimulator and the underlying theoretical models. Then we give some examples of how

Received: February 22, 2011
Published: April 28, 2011

MSSimulator can be used to benchmark algorithms or conduct an experimental robustness analysis.

METHODS

MSSimulator is written in C++ as part of the OpenMS[6] framework and is integrated into The OpenMS Proteomics Pipeline (TOPP).[7] The simulator is configurable via a parameter file, which can be edited using a dedicated GUI shipped with OpenMS. As input we use FASTA files, in addition to the parameter file containing the configuration. The FASTA file provides the protein or peptide sequences including modifications (we support all modifications contained in UNIMOD.[8]) and can also be used to provide protein/peptide specific information like the abundance or a specific retention time.

The user can also include contaminants into the simulation. For a detailed description of the format see Section II, Supporting Information. The simulation is divided into several submodules, accounting for the different steps carried out in a classical LC−MS experiment, which will be explained in detail in the following sections.

Digestion

Digestion can be performed in two modes or can be switched off. The first mode does a complete in-silico digest, also modeling missed cleavages. Note that when missed cleavages are used, also the completely cleaved peptides will be contained in the sample. To add another level of realism, the second mode uses a model from Siepen et al.,[9] which was reimplemented in OpenMS to predict missed cleavages. The current model is based on trypsin data but can be easily adapted simply by substituting a text file containing the model parameters. To extend the model to other enzymes, the log likelihood ratio data matrix described in the original paper needs to be computed.

Peptide Separation

As prefractionation techniques, two widely used approaches are available in MSSimulator: capillary electrophoresis (CE) and high performance liquid chromatography (HPLC). Both techniques yield separation of peptides according to different properties, therefore complementing each other. In CE mode, MSSimulator will predict a migration time based on a theoretical linear model described below, whereas for HPLC simulation we use a machine learning approach based on support vector regression.

A Model for Capillary Electrophoresis. In a strong electric field, molecules are separated based on their physicochemical properties that determine their migration time, which is further dependent on the background electrolyte and its properties, for example, ionic strength, pH, type of ions.

Our migration time model concentrates on simulating the electrophoretic mobility (μ_{ep}) of analytes, while electroosmotic flow (μ_{eo}), which is mainly governed by the viscosity of the buffer and the capillary itself, is a parameter provided by the user.

Electrophoretic mobilities and separations are predicted from physicochemical properties of the peptide species, namely net charge and mass. A common model for electrophoretic mobility is

$$\mu_{ep} = q/\mathrm{MW}^{\alpha} \tag{1}$$

where q is the net charge of the ion, MW is its molecular weight, and α is some constant. In a vacuum, an ions speed is proportional to its net charge when an electric field is applied. In a medium, however, we need to correct for frictional drag (MW^{α} term). The choice of α has been the topic of extensive discussion. The most common values include 1/3, 1/2, 2/3, which all relate to theoretical models. For details on choices of α and charge determination, see Section III, Supporting Information.

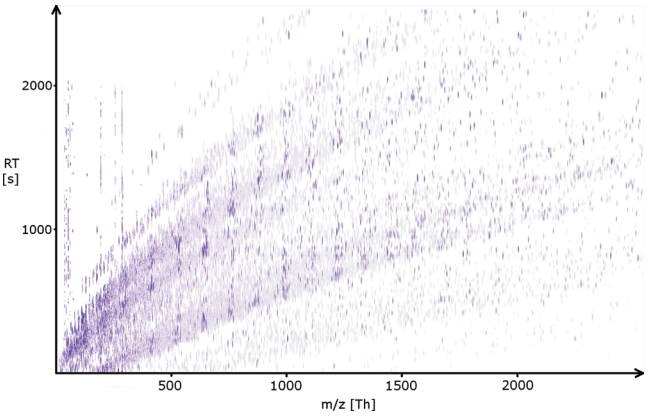


Figure 1. Raw CE/MS map of 100 proteins using default CE settings.

To determine the migration time we compute:

$$t = \frac{L_d L_t}{(\mu_{ep}+\mu_{eo})V} \tag{2}$$

where L_d is the distance between injection site and detector, L_t is the total capillary length and V is the applied voltage (see Laughlin et al.[10]). Peptides with negative migration times will be discarded (but mentioned in a summary statistic).

In contrast to HPLC where elution profiles remain constant across the RT dimension, in CE the peak width increases as a function of migration time due to dispersion factors and decreased mobility. We use a linear model to account for this effect. Figure 1 shows an exemplary CE/MS map using our CE model. The typical charge bands can be observed easily.

Prediction of Retention Times in Liquid Chromatography. Schulz-Trieglaff et al. already applied the Paired Oligo-Border Kernel (POBK) presented by Pfeifer et al.[11] to accurately predict the retention times for peptides in their simulation. We use the same approach in MSSimulator. A trained model is provided with our software but training a custom model using MS/MS identifications is easy using the RTModel tool which is part of TOPP.

A Model for Elution Profile Shape. Since peptides will not elute from the chromatography column at a single time point but over a period of time, we need to model the shape of the signal in the retention time dimension. In most cases this will be a Gaussian-like shape but can also have an asymmetric character. To be as flexible as possible, we have chosen the exponential Gaussian hybrid[12] (EGH) function

$$f_{egh}(t) = \begin{cases} H\exp\left(\dfrac{-(t-t_R)^2}{2\sigma_g^2+\tau(t-t_R)}\right), & 2\sigma_g^2+\tau(t-t_R) > 0 \\ 0, & 2\sigma_g^2+\tau(t-t_R) \le 0 \end{cases} \tag{3}$$

where t is the retention time, t_R is the center of the chromatographic peak, H is the peak height, σ_g is the standard deviation of the peak, and τ is the time constant of the exponential decay.

dx.doi.org/10.1021/pr200155f |J. Proteome Res. 2011, 10, 2922–2929

MSSimulator comes with a set of default values for σ_g and τ and the possibility to vary them using a Lorentzian distribution. For details on the motivation for the Lorentzian distribution and the determination of the parameters, see Section IV, Supporting Information.

To reflect poor chromatographic conditions, the user can also customize the quality of the generated elution profiles, by adding uniformly distributed noise.

Peptide Detectability Filter

Although detectability and ionization are closely coupled, we treat them as separate steps during simulation. To account for the effect that not necessarily all peptides ionize with the same efficiency, we include the peptide detectability filter presented by Schulz-Trieglaff et al.[3] It uses a support vector machine combined with a paired oligo-border kernel to compute the likelihood of each peptide to create a signal in a mass spectrum. The user can define a threshold value—every peptide below the threshold will be discarded. MSSimulator is shipped with a trained model. Customized models can be trained using TOPP's PTModel.

Ionization

We support the two common ionization methods electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). For ESI, we sample charge states for each peptide entity from a binomial distribution $B(n,p)$ where n is equal to the number of basic residues and p is set to 0.8 by default. We also support custom adducts like Na^+or K^+.

For MALDI we have chosen a discrete distribution of the charge states, with default probability values of $P(q = 1) = 0.9$ for charge 1 and $P(q = 2) = 0.1$ for charge 2. The user can customize the charge probabilities according to his own needs.

Modeling Peptide Signals in the Mass Spectrum

At this point, a list of peptides annotated with charge, retention time and an elution profile shape was generated. Based on this list MSSimulator computes the signals for each peptide ion. Each signal has two components, that is, the shape in the retention time dimension, which was already defined during the simulation of the chromatographic column, and the signal in m/z dimension.

To compute the complete isotopic envelope MSSimulator uses a fast algorithm[13] implemented in OpenMS. The shape of each individual isotopic peak is a topic of discussion in the literature[14] and can therefore be modeled during the simulation by either a truncated Gaussian or Lorentzian distribution. The width of the peaks can be controlled by the user in terms of the resolution. We additionally provide three models of resolution behavior, which are present in common instruments. Resolution is constant in time-of-flight (TOF) instruments; in Fourier transform ion cyclotron resonance (FTICR) instruments, it is known to degrade linearly with m/z, whereas in Orbitrap mass spectrometers it degrades with the square root of m/z.[15]

MS/MS Sampling

The prediction of fragment peak intensity in MS/MS spectra comprises a difficult task since the rules governing the fragmentation are not yet fully understood. For the most commonly used fragmentation method, the collision induced dissociation (CID), several approaches to predict the intensity pattern and identify important features with a strong influence on the fragmentation exist. Zhang[16] proposed a kinetic model to predict fragmentation for low energy CID spectra. Other approaches apply machine learning techniques like neural networks,[17] Bayesian neural networks,[18] probabilistic decision trees[19] or RankBoosting.[20]

In MSSimulator, the user can choose between three modes to simulate MS/MS spectra.

The naïve simulator generates peaks for all selected ion types (including neutral loss ions and charge variants) with a user defined intensity.

In the second mode, a support vector machine (SVM) is trained as a classifier to predict the abundance or absence of the primary ion types (b- and y-ions for CID spectra). For every peptide bond the fragment ions are encoded with the complete set of 35 descriptors used by Zhou et al.[18] (see Section VII, Supporting Information). If no peak is found within a certain user-defined interval around the expected m/z value, the ion is counted as missing. The classifier is trained on a class-balanced set of positive (abundant) and negative (missing) training samples, and suitable values for the SVM are obtained by grid search.

The third mode uses support vector regression (SVR) to predict the intensity of fragment ion peaks. The target value for a training feature vector is not a class label but the normalized (see ref 18) intensity of the observed peak. For computational efficiency reason, we use this model only to predict the b- and y-ion intensities. For the prediction of neutral loss ion intensities, we use a simple Bayesian approach where we learn the probability of observing a certain loss ion with a certain intensity, given the predicted intensity of the corresponding primary ion. As this approach requires discrete intensity levels, we apply intensity binning.

The latter two modes are currently only supported for peptides with a maximum charge of three. Customized models can be trained on user defined data sets.

In Section VII of the Supporting Information, we present a comparison of our SVR prediction approach with the two models by Zhang[16] and Zhou et al.,[18] which reveals that the performance is comparable to the other two approaches.

Simulating MS^E Data. Concurrent peptide fragmentation (i.e., MS^E) is an emerging technique that could revolutionize the way peptides are identified and quantified. Currently there are very few algorithms capable of analyzing MS^E data, for example, ETISEQ.[21] By providing simulated data, we hope to facilitate algorithm development, as the simulator provides an easy means to benchmark the results. MS^E data is generated by alternatively recording data in MS and MS/MS mode, where the latter has no restriction on the precursor mass, thus all ions are fragmented simultaneously. This has the advantage that suboptimal precursor selection is no longer an issue but leads to congested MS/MS spectra that need to be disentangled for proper peptide identification. The simulator will create MS/MS spectra for each peptide currently eluting from the HPLC/CE column according to our fragmentation model, scaled by their respective intensity, such that MS and MS/MS spectra will display proper elution profiles, which can be used to correlate MS/MS peaks with MS features. Subsequently, the single spectra are merged to form the final MS^E spectrum. An example can be seen in Figure 2. The peaks are color coded by precursor.

Labeled Experiments

The simulator contains a framework which allows the easy and fast incorporation of any labeling technique used in mass spectrometry. We currently provide three widely used techniques, namely iTRAQ (isobaric tag for relative and absolute quantitation),[22] SILAC (stable isotope labeling by amino acids in cell culture)[23]

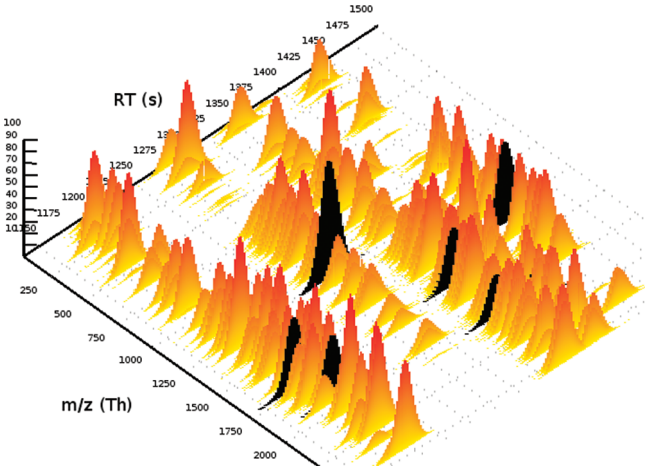


Figure 2. Color coded detail of MS^E spectrum containing seven precursor species (black). Intensities are scaled to 100% for MS and MS/MS spectra.

and ^{18}O labeling,[24] in addition to the usual label-free setup. For each labeled channel, a FASTA input file must be given. This allows to model different protein/peptide sets but also abundances and modification states separately for each channel.

iTRAQ Labeling. The software can be used to simulate iTRAQ MS/MS spectra with arbitrary channel allocation (using 4plex or 8plex) and customizable isotope correction matrices (the default being the matrix provided by Applied Biosystems). The labeling efficiency of tyrosine residues can be changed as desired, with a default of 30%. A peptide containing a Y residue will be split into two sibling peptides with different masses, each with an abundance reflecting the labeling efficiency. N-terminus and lysine residues are assumed to be fully labeled. The MS/MS spectra generated in iTRAQ mode differ from normal MS/MS spectra in that they contain the reporter ions in the m/z range from 113 to 121 Th and that the fragment ions are 145 Da heavier for every iTRAQ modified amino acid they contain. Fragment ions with partially or even completely cleaved iTRAQ tags seem to be missing from the iTRAQ spectra we examined.

Stable Isotope Labeling by Amino Acids in Cell Culture. SILAC is a prominent approach in quantitative proteomics based on the incubation of cell lines with an isotopically labeled form of an amino acid (e.g., deuterated leucine).

MSSimulator currently supports two channel SILAC labeling with a user defined modification. The default is a modified lysine and arginine introducing a mass shift of ~6.02 Da. We assume complete incorporation of the label into the labeled channel, but incomplete incorporation could be easily implemented as well.

^{18}O Stable Isotope Labeling. Labeling peptides with stable ^{18}O isotopes is a widely used technique in quantitative proteomics and therefore also supported by MSSimulator. Labeling peptides with ^{18}O tags is achieved by digesting the proteins with an endoprotease (usually trypsin) in the presence of $H_2{}^{18}O$. This reaction exchanges two C-terminal oxygen molecules by ^{18}O and thereby introduces a mass shift of 4 Da. Since the labeling reaction is not always complete, also monolabeled peptides (mass shift of 2 Da) and unlabeled peptides will occur in the labeled channel.

To account for the labeling efficiency MSSimulator splits the total peptide amount given in the labeled channel B on the three different states: unlabeled B_0, mono- B_1 and dilabeled B_2. To compute the quantities depending on the labeling efficiency f the kinetic model of Ramos-Fernández et al.[25] is used.

$$B_0 = B(1-f)^2 \quad (4)$$

$$B_1 = B_2 f(1-f) \quad (5)$$

$$B_2 = Bf^2 \quad (6)$$

Output

The user can specify multiple output files, which provide different layers of ground truth. The most important one and the only mandatory is the output file for the raw MS data in mzML[26] format. If the user requires another PSI format, the OpenMS FileConverter can be used for conversion.

The second optional output file is a feature map (as featureXML) containing all simulated peptides, annotated with charge, charge adducts, and sequence. The featureXML file can easily be converted to an Excel sheet or csv (comma-separated values) file. Also a list of features describing the contaminants in the data set can be requested by the user.

Additionally, MSSimulator can provide files containing the correct associations between the different charge variants of a single peptide and the correct associations between the labeled and unlabeled versions of the simulated peptides. The files are in the OpenMS specific consensusXML format which also can be easily converted to an Excel sheet or csv file.

Availability

The presented software is included in v1.8 of the open source C++ software library OpenMS, running on all major platforms, available at http://www.OpenMS.de. This also applies to all TOPP tools used in this publication.

RESULTS

Due to the wide feature range of MSSimulator, it can be easily adapted to mimic certain instrument types. We provide exemplary preset configurations for a QTOF and an FT instrument; other instruments can be created easily. To asses the level of realism of the simulated data, when using a similar setup (in terms of protein mix, instrument settings etc.), we used data sets from the Standard Protein Mix Database[27] (Mix 3, low-res QTOF and high-res Fourier Transform (FT) data) for comparison to simulated data. After applying the same analysis pipeline (centroiding, feature finding) to both data sets, we find that the number of peptide signals, charge distribution, intensity range are highly comparable. For a detailed comparison and configuration files, see Section V, Supporting Information. For a visual comparison, see Figure 3.

As the focus of the simulator is benchmarking of algorithms, we will apply MSSimulator to a wide range of tasks in the following subsections.

Algorithm Verification for MS^E Data

We used MSSimulator in MS^E mode to benchmark the ETISEQ software, which to our knowledge is the only software publicly available for this task. A very simple data set was generated, consisting of one protein (P62739, bovine actin), yielding 44 peptide signals in different charge and cleavage states. We disabled simulation of contaminants to make the spectra as clean as possible. MS and MS^E spectra were generated alternating. Additionally the simulator was configured to create the

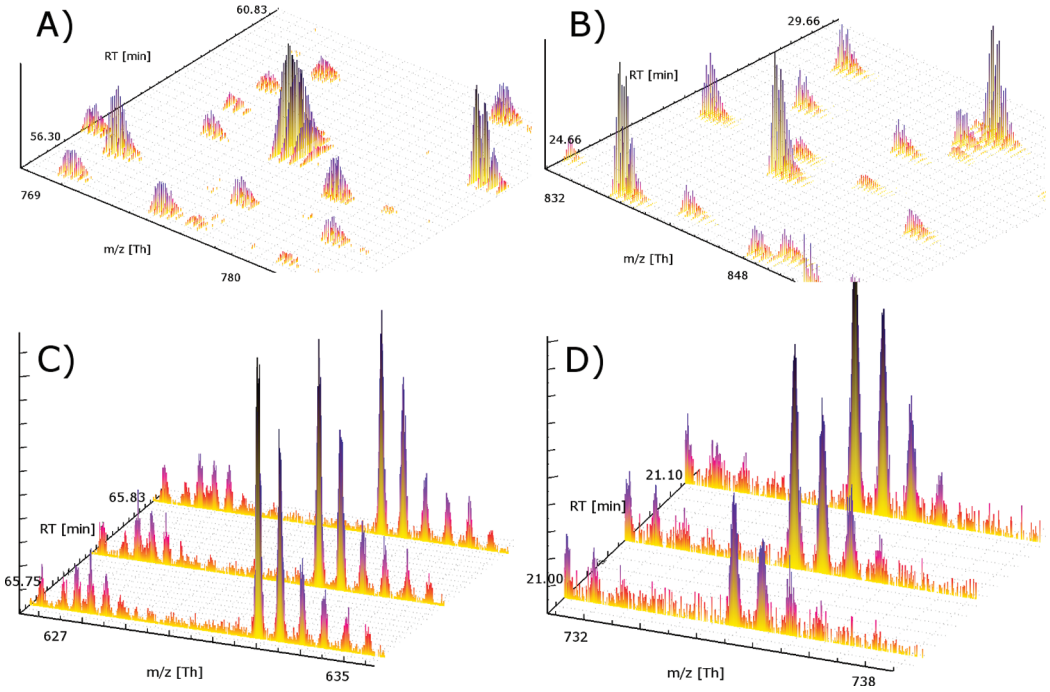


Figure 3. Comparison of real vs simulated data for FT and QTOF instruments. For clarity, data is shown on zoomed regions of an LC−MS map. (A) Real FT data, (B) simulated FT data, (C) real QTOF data, and (D) simulated QTOF data.

"debug" MS/MS spectra, which can used as a ground truth when assessing the disentangled ETISEQ spectra. Before submitting the data set via the ETISEQ webinterface (http://www.cancerresearch.unsw.edu.au/CRCWeb.nsf/page/Elution+time+ion+sequencing) using default parameters (except for "Exclude contaminant ions": set to *false*), we removed the debug spectra. Unfortunately, the data set returned by ETISEQ was not a valid mzXML file, which we fixed by applying regular expressions (see Section I, Supporting Information). We are currently in contact with the authors of ETISEQ to address further problems we encountered, and thus we cannot present results here. However, some of these flaws (e.g., wrong precursor information of deconvoluted MS/MS spectra) were only traceable due to our knowledge of ground truth and would have been very hard to find on real data. This shows that simulated data can indeed help to assess algorithm reliability.

Development and Quality Assessment of Labeled Quantification Algorithms

Developing quantitation algorithms for labeled or unlabeled mass spectrometry experiments is always a laborious task. Especially in labeled setups the algorithm optimization and comparison is always hindered by the unavailability of the complete set of labeled pairs or sets.

To prove the applicability of MSSimulator in benchmarking tools for SILAC quantitation we compared two known approaches for quantitation of stable-isotope labeling, XPRESS[28] and ASAPRatio.[29] We used the versions integrated into the Trans-Proteomic Pipeline (TPP)[30] v4.4.1 (VUVUZELA). The popular MaxQuant[31] unfortunately cannot be considered here, as it only supports the commercial RAW data format as input.

Both tools use a combination of MS/MS identifications and chromatographic peaks in the RAW data for quantitation. ASAPRatio, the more recently developed tool, has the more sophisticated data handling and error analysis and is expected to show better results than XPRESS.

To compare both quantitation approaches we generated a data set where the second channel was labeled with a modified version of lysine and arginine introducing a mass shift of ≈6.02 Da. The SILAC pairs were generated with the following ratios: 1:1, 1:2, 1:4 and 1:10. The data set contained 782 peptide features after digestion using the nave trypsin model and HPLC simulation on the default column. Following the simulation we generated exact identification results for all peptide features, removing the effect of inaccurate MS/MS identification in the analysis. These identification results were converted into the pepXML (http://tools.proteomecenter.org/wiki/index.php?title=Formats:pepXML) format using TOPP and analyzed by XPRESS and ASAPRatio. XPRESS as well as ASAPRatio produce annotated pepXML containing the computed peptide ratios. Figure 4 shows the resulting peptide ratios plotted against the simulated ratios. Both tools reconstruct most of the simulated ratios, however have (as expected) problems with overlapping signals. The results shown in Figure 4 reflect the expected superiority of ASAPRatio since it has a more robust error analysis then XPRESS.

The presented approach for the assessment of quantitation tools can be easily extended and automated for the evaluation of different tools under several conditions like changing noise levels, machine types, or changes in the resolution of the data and the effect of these changes can be quantified directly using the available ground truth (feature positions, simulated ratios, etc.).

Map Alignment Algorithm Stability

In this study, we aim to benchmark the ability of a map alignment strategy to correct for a retention time distortion between two simulated data sets, when the overlap of sample content is varied. We used the simulator to create feature maps of decreasing overlap in terms of protein content but constant

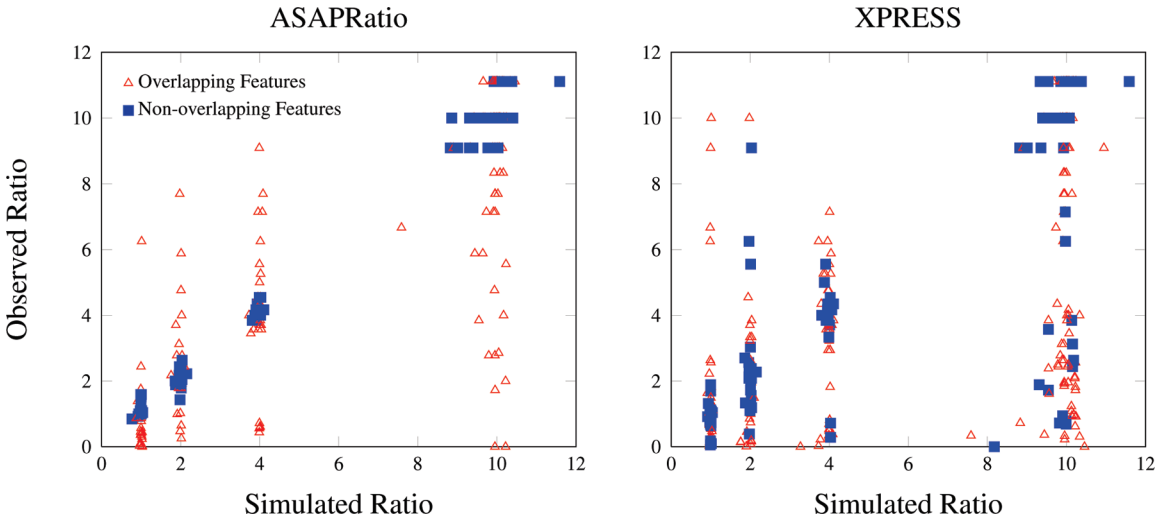


Figure 4. Ratios computed by ASAPRatio (left) and XPRESS (right) plotted against the ratios simulated by MSSimulator. Peptide features that overlap with at least one other feature (which is not the labeled partner) are marked with red triangles, nonoverlapping features are marked as blue squares.

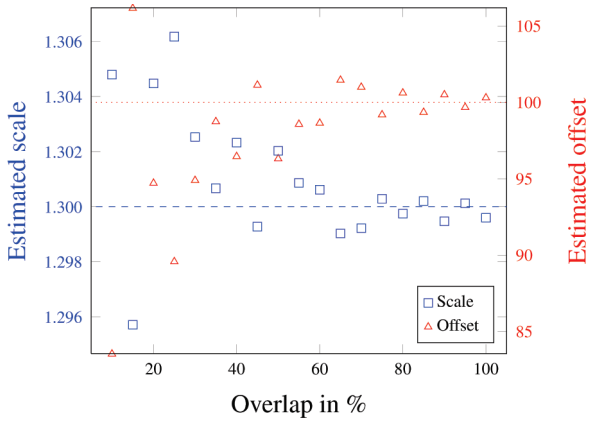


Figure 5. Quality of alignment when altering peptide overlap between the two data sets. The red triangles indicate the reconstructed offset in comparison to the simulated offset (red dotted line). The blue squares indicate the reconstructed scale in comparison to the simulated scale (blue dashed line).

number of features (∼4000) and applied the TOPP MapAligner tool to reconstruct the affine retention time shift, plus a local Gaussian distributed distortion. We chose offset = 100, scale = 1.3 and a local Gaussian distortion with sd = 3 for each feature. This scenario can give insight on how many corresponding features (i.e., alignment anchors) are sufficient to reconstruct the correct alignment. Another point of view is on how inefficient the feature identification is allowed to be on replicates, to reconstruct the RT shift. The results (see Figure 5) show that even a very small overlap does allow the reliable estimation of the true transformation.

Feature Detection in High-Resolution Data

Assessing the quality of feature detection algorithms is always a cumbersome task, since the exact location and charge of the real peptide signals are initially unknown. To overcome this problem feature detection algorithms are often developed and tested on manually annotated data sets.

As an addition to the established approach, we propose the use of simulated data sets as already done in previous works.[1,3] The exact knowledge of feature positions and properties eases the computation of essential quality values like false discovery rate (FDR) and true positive rate (TPR). Also the influence of various data specific properties like resolution, noise or chromatographic conditions can be easily quantified.

To cover this scenario we used simulated data to benchmark the performance of Hardklör[32] an established feature finding tool for high-resolution data and the FeatureFinder shipped with TOPP. The data was simulated using the settings for an FT instrument and input data described in Section V, Supporting Information. Hardklör was run with slightly modified parameters according to the "Sample Config Files" section on the Hardklör Web site (http://proteome.gs.washington.edu/software/hardklor/config.html). The configuration files for both tools and a description of the analysis steps can be found in Section VI, Supporting Information.

Both tools showed a good performance with a FDR of 0.07 and a TPR of 0.855 for Hardklör and 0.196 and 0.814 for the TOPP FeatureFinder respectively. To quantify the influence of poor chromatographic conditions on the performance of both tools, we repeated the analysis on a data set with an increased level of distortion for the simulated elution profiles. The performance of both tools dropped to FDR and TPR values of 0.115 and 0.798 for Hardklör and 0.203 and 0.642 for the TOPP FeatureFinder.

In this scenario, we showed that with only a small effort in data preparation (compared to manually annotating real data sets) one could quickly benchmark existing or self-developed software and assess the influence of data specific properties like chromatographic conditions on the performance of these tools.

CONCLUSION

MSSimulator is the most extensive collection of algorithms and models for MS simulation and allows for easy algorithm validation on a broad range of conditions, opening a wide range of benchmarking scenarios, which can easily be automated. The availability of a ground truth reduces the need for expensive manual validation on real data sets. Also, future labeling techniques can be added quickly by implementing our labeling interface.

We have shown that our simulated data is very similar to real data and allows easy validation of existing algorithms.

We invite the community to extend the solution presented here. Future extensions might include but are not limited to, automatic estimation of simulation parameters (e.g., resolution, sampling rate, noise level) from real data allowing to quickly generate benchmark data for analysis software, prediction of ionizability, incorporation of additional noise models, and more instrument specific properties (e.g., shadow peaks on Orbitrap instruments).

■ ASSOCIATED CONTENT

Supporting Information

Supplemental figures, tables, and data. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chris.bielow@fu-berlin.de; stephan.aiche@fu-berlin.de.

Author Contributions

†These authors are joint first authors

■ ACKNOWLEDGMENT

We thank Ole Schulz-Trieglaff for inspiration from the first implementation, the OpenMS development team for contributions to the library, and Silke Ruzek for her expertise on iTRAQ. Special thanks go to Alexandra Zerck for providing the essential parts of the precursor selection code. C.B. gratefully acknowledges funding by the European Commissions's seventh Framework Program (GA202222).

■ REFERENCES

(1) Morris, J. S.; Coombes, K. R.; Koomen, J.; Baggerly, K. A.; Kobayashi, R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* **2005**, *21*, 1764–75.

(2) Coombes, K. R.; Koomen, J. M.; Baggerly, K. A.; Morris, J. S.; Kobayashi, R. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Inform.* **2005**, *1*, 41–52.

(3) Schulz-Trieglaff, O.; Pfeifer, N.; Gröpl, C.; Kohlbacher, O.; Reinert, K. LC-MSsim—a simulation software for liquid chromatography mass spectrometry data. *BMC Bioinform.* **2008**, *9*, 423.

(4) Renard, B. Y.; Kirchner, M.; Steen, H.; Steen, J. A. J.; Hamprecht, F. A. NITPICK: peak identification for mass spectrometry data. *BMC Bioinform.* **2008**, *9*, 355.

(5) Yang, C.; He, Z.; Yu, W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinform.* **2009**, *10*, 4.

(6) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinform.* **2008**, *9*, 163.

(7) Kohlbacher, O.; Reinert, K.; Gröpl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **2007**, *23*, e191–7.

(8) Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4*, 1534–6.

(9) Siepen, J. A.; Keevil, E.-J.; Knight, D.; Hubbard, S. J. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J. Proteome Res.* **2007**, *6*, 399–408.

(10) Laughlin, G. M. M.; Nolan, J. A.; Lindahl, J. L.; Palmieri, R. H.; Anderson, K. W.; Morris, S. C.; Morrison, J. A.; Bronzert, T. J. Pharmaceutical Drug Separations by HPCE: Practical Guidelines. *J. Liq. Chromatogr. Relat. Technol.* **1992**, *15*, 961–1021.

(11) Pfeifer, N.; Leinenbach, A.; Huber, C. G.; Kohlbacher, O. Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *BMC Bioinform.* **2007**, *8*, 468.

(12) Lan, K.; Jorgenson, J. W. A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks. *J. Chromatogr. A* **2001**, *915*, 1–13.

(13) Kubinyi, H. Calculation of isotope distributions in mass spectrometry. A trivial solution for a non-trivial problem. *Anal. Chim. Acta* **1991**, *247*, 107–19.

(14) Matthiesen, R., Ed. *Mass Spectrometry Data Analysis in Proteomics*; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2007; p 336.

(15) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; Strupat, K.; Horning, S. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **2006**, *78*, 2113–20.

(16) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76*, 3908–22.

(17) Arnold, R. J.; Jayasankar, N.; Aggarwal, D.; Tang, H.; Radivojac, P. A machine learning approach to predicting peptide fragmentation spectra. *Pacific Symp. Biocomput.* **2006**, *230*, 219–30.

(18) Zhou, C.; Bowler, L. D.; Feng, J. A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinform.* **2008**, *9*, 325.

(19) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **2004**, *22*, 214–9.

(20) Frank, A. M. Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.* **2009**, *8*, 2226–40.

(21) Wong, J. W. H.; Schwahn, A. B.; Downard, K. M. ETISEQ—an algorithm for automated elution time ion sequencing of concurrently fragmented peptides for mass spectrometry-based proteomics. *BMC Bioinform.* **2009**, *10*, 244.

(22) Ross, P. L.;Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154–69.

(23) Ong, S.-E. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **2002**, *1*, 376–86.

(24) Mirgorodskaya, O. A.; Kozmin, Y. P.; Titov, M. I.; Körner, R.; Sönksen, C. P.; Roepstorff, P. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1226–32.

(25) Ramos-Fernández, A.; López-Ferrer, D.; Vázquez, J. Improved method for differential expression proteomics using trypsin-catalyzed 18O labeling with a correction for labeling efficiency. *Mol. Cell. Proteomics* **2007**, *6*, 1274–86.

(26) Martens, L. mzML - a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics* **2010**No. R110.000133.

(27) Klimek, J.; Eddes, J. S.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P. R.; Katz, J. E.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J. K.; Aebersold, R.; Martin, D. B. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J. Proteome Res.* **2008**, *7*, 96–103.

(28) Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **2001**, *19*, 946–51.

(29) Li, X.-J.; Zhang, H.; Ranish, J. A.; Aebersold, R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 6648–57.

(30) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.;

Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150–9.

(31) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–72.

(32) Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* **2007**, *79*, 5620–32.