Deuteration Distribution Estimation with Improved Sequence Coverage for HX/MS Experiments

Xinghua Lou[1], Marc Kirchner[1,3,5‡], Bernhard Y. Renard[1,3,‡], Ullrich Köthe[1],
Sebastian Boppel[1], Christian Graf[2], Chung-Tien Lee[2], Judith A. J. Steen[3,4],
Hanno Steen[3,5], Matthias P. Mayer[2], Fred A. Hamprecht[1,3,*]

‡Authors contributed equally.

[1] Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany

[2] Zentrum für Molekulare Biologie der Universität Heidelberg (ZMBH), DKFZ-ZMBH Alliance, Heidelberg, Germany

[3] Proteomics Center at Children's Hospital Boston, Boston, Massachusetts, USA

[4] Department of Neurobiology, Harvard Medical School, Boston, Massachusetts, USA

[5] Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA

April 2010

Abstract

Time-resolved hydrogen exchange (HX) followed by mass spectrometry (MS) is a key technology for studying protein structure, dynamics and interactions. HX experiments deliver a time-dependent distribution of deuteration levels of peptide sequences of the protein of interest. The robust and complete estimation of this distribution for as many peptide fragments as possible is instrumental to understanding dynamic protein-level HX behavior. Currently, this data interpretation step still is a bottleneck in the overall HX/MS workflow.

We propose *HeXicon*, a novel algorithmic workflow for automatic deuteration distribution estimation at increased sequence coverage. Based on an L_1-regularized feature extraction routine, HeXicon extracts the full deuteration distribution, which allows insight into possible bimodal exchange behavior of proteins, rather than just an average deuteration for each time point. Further, it is capable of addressing ill-posed estimation problems, yielding sparse and physically reasonable results. HeXicon makes use of existing peptide sequence information which is augmented by an inferred list of peptide candidates derived from a known protein sequence. In conjunction with a supervised classification procedure that balances sensitivity and specificity, HeXicon can deliver results with increased sequence coverage.

The entire HeXicon workflow has been implemented in C++ and includes a graphical user interface. It is available at http://hci.iwr.uni-heidelberg.de/software.php.

[*] to whom correspondence should be addressed

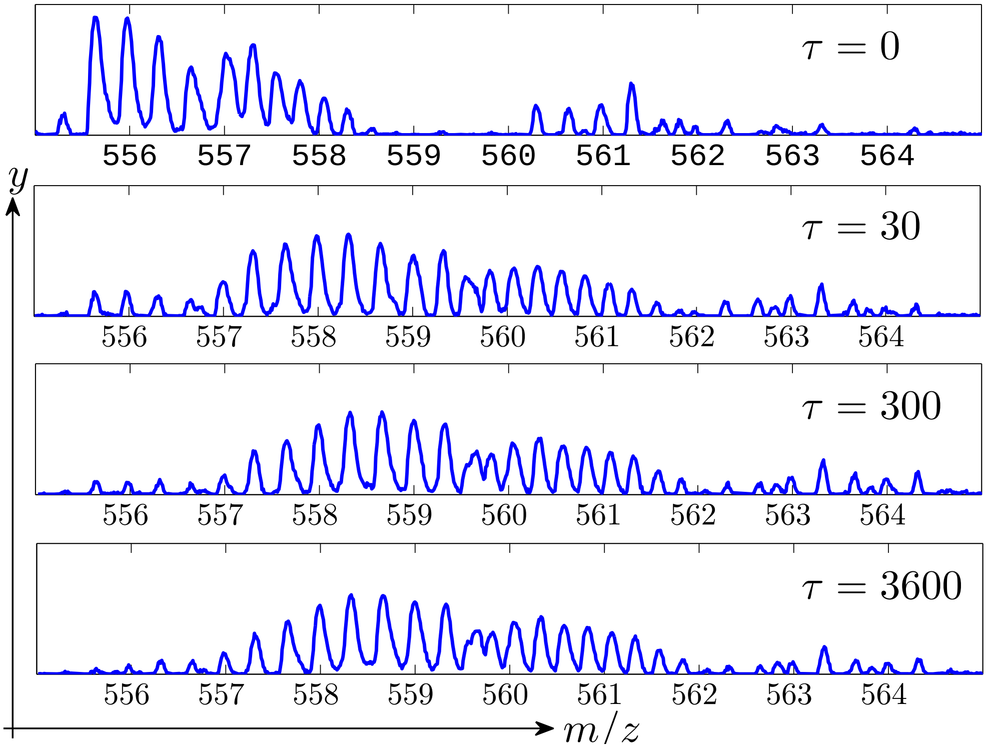


Figure 1: Examples of HX/MS spectrum data from an incubation time series of 0, 30, 300 and 3600 seconds: the isotope envelope shifts to higher m/z values because of deuterium incorporation. The deuteration content is encoded in a complex mixture of isotope distributions. Due to the noise and overlapping isotope clusters, the separation of individual peptides is non-trivial. The abundance of the spectrum is labeled as y.

1 INTRODUCTION

The determination of protein structure and dynamics is a key issue for the understanding of living systems [7]. By combining the information of the protein dynamics and other classical functional data, a more complete understanding of protein function can be obtained. In many cases, protein dynamics are directly related to specific protein functions such as conformational changes during enzyme activation and protein movements during binding [23]. Hydrogen exchange followed by mass spectrometry (HX/MS) has become a standard approach for interpreting HX experiments: the location and rate of deuteration are indicative of solvent accessibility and in particular hydrogen bonding and hence of conformation and dynamics [10]. They can be estimated by tracking the mass shift of peptide fragments in mass spectra over samples with different incubation times (Figure 1) [9]. In comparison to Nuclear Magnetic Resonance (NMR) spectroscopy, mass spectrometry requires lower protein concentrations and amounts, provides higher measurement speed and better scalability in terms of protein size, and detects coexisting conformations [12]. Whereas manifold improvements in experimental methodology and instrumentation have been implemented for HX/MS experiments, data processing still remains a major difficulty in the overall experimental workflow [9]. First of all, the precise deuteration distribution is represented by complex peak patterns that are difficult to separate and quantitate even in 2D LC/MS (Liquid Chromatography/Mass Spectrometry) representation. Secondly, the peptide sequences of interest have to be pre-determined via MS/MS search report or selected empirically, yielding suboptimal sequence coverage of the protein of interest. Finally, manual analysis is time-consuming, error-prone as well as inaccurate in case of overlapping isotope clusters (Figure 1).

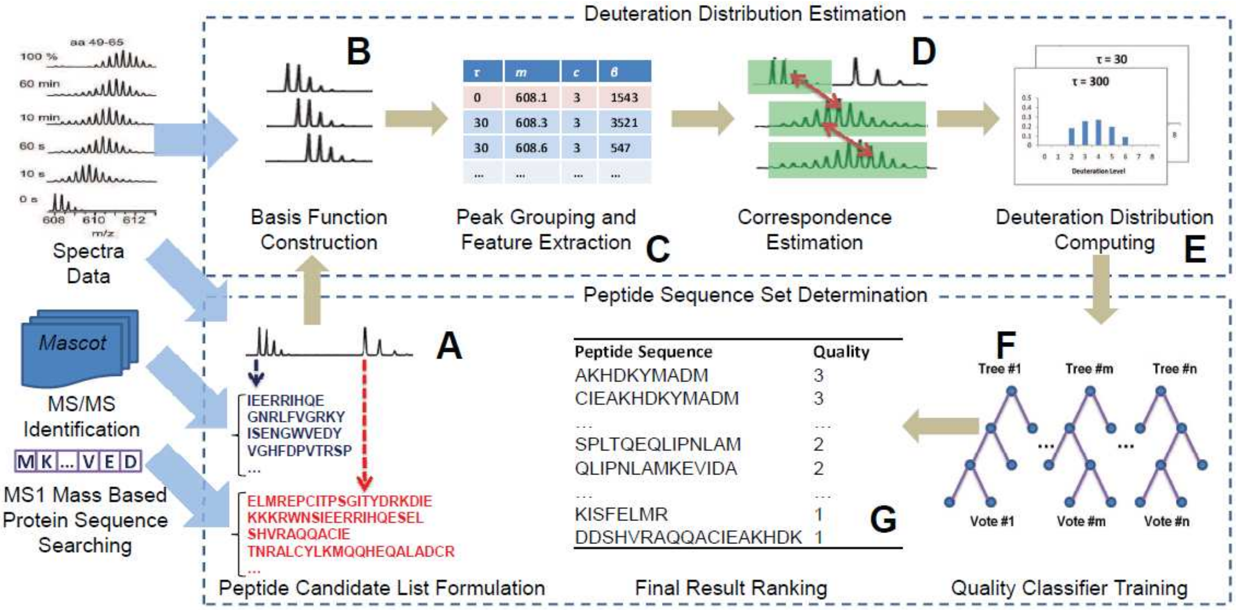
Several methods and tools have been developed to facilitate the manual analysis. Palmblad and colleagues [16] modeled the deuterium incorporation as a binomial distribution and used χ^2-statistics to extract the optimal parameter. Weis and Engen [24] designed HX-Express as a semi-automatic data processing tool which measures the deuteration by the width of the given isotope pattern. TOF2H [15] is an integrated software framework designed specifically for semi-automatic LC-MALDI (Matrix-Assisted Laser Desorption/Ionization) data analysis.

Note that while the approaches mentioned above facilitate the analysis of HX/MS data, they do not yield the complete deuteration distribution, but only the average deuteration. The true deuteration distribution offers a more detailed characterization and more insightful description of the exchange process. In particular, it is suitable for discovering bimodal exchange behaviors of large protein oligomers, which are not detectable by average deuteration levels.

The algorithms developed for extracting deuteration distribution information mainly fall into two categories. The first set of methods fit a hypothetical deuterated isotope pattern to the observed spectrum by least-squares regression [1, 13, 21]. They exhibit the advantage of speed but have difficulties in handling ill-posed problems, which, as shown in the following, are common in large-scale HX/MS data analysis. It is possible to make use of padding methods to regularize the ill-posed regression problem. Given the optimal degree of padding, this approach can address data truncation problems and avoid over-fitting to noise [6]. The second set of methods is based on maximum entropy deconvolution [25, 1]. Those methods can handle ill-posed problems and yield non-negative outputs; however, they are computationally much more expensive [25]. One common limitation of these two categories is that they are designed for well-tuned and small-scale problems, i.e. the peptide sequence of interest is pre-selected in a well-separated spectrum, thus making them less applicable in practice, especially for large-scale HX/MS data processing. These methods have been implemented by several software tools such as Deuterator [18, 17] and Hydra [21]. Both frameworks focus on incorporating existing algorithms and providing user-friendly GUI and powerful visualization.

We propose a novel algorithmic approach named HeXicon to the deuteration distribution estimation problem for large-scale HX/MS experiments. HeXicon exploits information in the retention time and m/z domains for optimized separation of large HX/MS data and applies NITPICK [19] for LC/MS feature extraction, resulting in a robust and regularized estimation of the deuteration distribution. It integrates protein sequence and protein identification information in an attempt to increase the sequence coverage.

Section 2 of the manuscript elaborates the methodological development of our approach. Sections 3 describes the experimental setup and reports the results, focusing on the novelty of delivering a robust estimate of the deuteration distribution and the comparison to manual analysis. Discussion and conclusion are offered in sections 4 and 5, respectively.



τ	m	c	β
0	608.1	3	1543
30	608.3	3	3521
30	608.6	3	547
...	...	...	...

Peptide Sequence	Quality
AKHDKYMADM	3
CIEAKHDKYMADM	3
...	...
SPLTQEQLIPNLAM	2
QLIPNLAMKEVIDA	2
...	...
KISFELMR	1
DDSHVRAQQACIEAKHDK	1

Figure 2: Workflow of HeXicon. **A:** The list of peptide identification from MS/MS searches is automatically extended by matching theoretical peptides to observed masses to find peptide sequence candidates for previously unidentified peptides; **B:** A basis function set is created by modeling all possible deuteration levels for each peptide sequence; **C:** The spectra and basis function sets are inserted into the LC/MS segmentation and NITPICK routine and groups of peaks with features are extracted; **D:** The correspondence of inter-experiment peak groups are identified via a weighted Euclidean distance measure; **E:** The deuteration distribution is derived; **F:** A random forest classifier discriminates high-quality results from low-quality results; **G:** The final results are ranked by their quality score.

2 METHODS

As illustrated in Figure 2, our approach consists of two major modules that jointly carry out our goals of robust deuteration distribution estimation and sequence coverage improvement. Given a hypothetical set of peptide sequences inferred in *Peptide Sequence Set Determination* (**A**), the *Deuteration Distribution Estimation* starts by constructing an over-complete set of basis functions (**B**) and then feeds them into the NITPICK algorithm to yield peak groups with features (**C**). Inter-experiment peak groups are then associated via correspondence estimation (**D**) and the deuteration distribution is derived for each association (**E**). The subsequent quality estimation of *Peptide Sequence Set Determination* retains the high-quality results and thus balances the sensitivity and specificity (**F, G**). Our approach makes extensive use of the NITPICK algorithm, a regularized, non-greedy, globally optimal linear mixture modeling algorithm for feature extraction from multicomponent mass spectra.

2.1 Deuteration Distribution Estimation

Definition Let p be a peptide sequence of interest. The deuteration level k is the number of deuterium exchanges at the back-bone hydrogens of p. The deuteration distribution $\rho(p, k, \tau)$ is the fraction of peptide with sequence p at deuteration level k for incubation time τ, where $k \in \{0, 1, \ldots, K(p)\}$ and $K(p)$ is the maximal possible deuteration level. The average deuteration $\eta(p, \tau)$ is the average deuteration level of all peptides with sequence p at incubation time τ.

2.1.1 NITPICK Algorithm

We formulate the deuteration distribution estimation as a regression problem. That is, the observed spectrum $\boldsymbol{s}$ is explained as a linear combination of constituent basis spectra which represent a particular peptide. Each feasible basis spectrum is specified by one column of the regression matrix $\boldsymbol{\Phi}$, and the regression coefficients $\boldsymbol{\beta}$ determine the abundance of those constituents in the mixture. If the matrix $\boldsymbol{\Phi}$ contains more basis functions than are actually present in any given mixture $\boldsymbol{s}$, the regression problem is ill-posed and has to be constrained. [22] showed that the introduction of a $L1$-constraint leads to a sparse solution vector $\boldsymbol{\beta}$ which assigns non-zero abundance only to those basis functions that are contained in the mixture with high probability. The resulting regression problem is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \big\{ ||\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{\beta}||_2^2 + \lambda \, ||\boldsymbol{\beta}||_1 \big\} \text{ subject to } \boldsymbol{\beta} \geq 0, \qquad (1)$$

which can be solved with the same computational efficiency as an ordinary least squares problem by the LARS algorithm [8]. The regularization parameter λ controls the model complexity based on the Bayesian Information Criterion (BIC) [20]. The NITPICK algorithm [19] determines its value automatically so that the number of degrees of freedom in the model is matched to the observed noise level of $\boldsymbol{s}$.

2.1.2 Basis Function Construction

Assuming that the peptide sequence set of interest P is known (see section 2.2), the solution to the regression problem must lie in a space spanned by all deuteration levels of all peptide sequences in the set (Figure 2 **B**). Thus, we build the basis function set $\boldsymbol{\Phi}$ by combining the theoretical isotope distribution for every deuteration level of each peptide sequence in P:

$$\boldsymbol{\Phi} = \bigcup_{\forall p \in P, \forall k \in [0, K(p)]} \phi(p, k) \qquad (2)$$

where $\phi(p, k)$ is the transformation function that computes the basis function for peptide sequence p at deuteration level k (i.e. its theoretical isotopic spectrum); $K(p)$ is the maximum number of exchangeable hydrogens [23]. To accommodate for the non-constant, m/z-dependent resolution [11, 5], we use a m/z-dependent peak shape function and learn its parametrization from the data (see **Supplementary Data**).

2.1.3 Quantitative LC/MS Feature Extraction

This feature extraction procedure provides two key steps for the workflow: firstly, it selects a subset of basis functions $\hat{\mathbf{\Phi}} \subseteq \mathbf{\Phi}$ that optimally explain the observed spectrum and thus determines the peptide sequences of interest; secondly, it extracts features of selected basis functions for the following deuteration distribution computation and correspondence estimation.

We first apply segmentation techniques to achieve optimized separation of the LC/MS data, which yields better signal-noise-ratio (SNR) and groups signals that belong to the same peptide. Manual analysis and some existing methods normally use a heuristic window-based approach for separating the LC/MS data. The integration of LC/MS peaks along the entire retention time window yields suboptimal SNR and fails in case of overlapping peak clusters [2]. Therefore, we integrate over retention time only within segments and are thus able to benefit from better SNR. The exact retention time position of the peptide is then determined via a sparse elution profile estimation on the LC/MS data segment [3]. Thereafter, to determine the ratio of different deuteration levels of the peptide sequence of interest, the abundance of their corresponding basis function $\phi(p, k)$ is estimated using the NITPICK algorithm (Figure 2 **C**). The regression problem (Equation 1) is normally ill-posed because the basis function construction yields an over-complete set of explanatory variables. Also, NITPICK provides sparse solutions which represent a subset of the over-complete basis function set that is indeed necessary to explain the observed spectrum.

Eventually, for each incubation time τ, the feature extraction procedure outputs a list of peak groups $\boldsymbol{\mathcal{G}}^{\tau}$, where a group $\boldsymbol{g}^{\tau}$ corresponds to a certain segment and contains peaks with features:

$$\boldsymbol{\mathcal{G}}^{\tau} = \{\boldsymbol{g}^{\tau}\} = \left\{(m, \beta, z, t)_{\boldsymbol{g}^{\tau}}\right\}$$

where m is the monoisotopic m/z position, β is the abundance of the corresponding basis function, z is the charge and t is the estimated retention time.

2.1.4 Correspondence Estimation

This step determines the correspondences between the peak groups over incubation time points and the peptide sequences of interest (Figure 2 **D**). Given a peptide sequence p of interest, its zero exchange peak group is first determined by matching a measured peak to its theoretical m/z value,

$$\hat{\boldsymbol{g}}^{0} = \underset{\boldsymbol{g}^{0} \in \boldsymbol{\mathcal{G}}^{0}}{\arg\min} |m_{\boldsymbol{g}^{0}} - f_{\text{theoretical}}(p, z, 0)\,|, \tag{3}$$

where $f_{\text{theoretical}}(p, z, k)$ computes the theoretical m/z of p at charge z and deuteration level k. The corresponding peak group at every other incubation time is determined by minimal weighted Euclidean distance

$$\hat{\boldsymbol{g}}^{\tau} = \underset{\boldsymbol{g}^{\tau} \in \boldsymbol{\mathcal{G}}^{\tau}}{\arg\min} \sqrt{(\boldsymbol{g}^{\tau} - \boldsymbol{g}^{0})^{T} \boldsymbol{S}^{-1} (\boldsymbol{g}^{\tau} - \boldsymbol{g}^{0})}, \tag{4}$$

where $\boldsymbol{S}$ is a diagonal matrix which normalizes and weights the contributions of different features to the distance measure. The matrix $\boldsymbol{S}$ is designed to express the characteristics of signals belonging to the same peptide sequence over incubation time. To speed up the computation, we also applied a filtering procedure to eliminate unlikely candidates by charge consistency and thresholding via retention time window and m/z accuracy cutoff (see **Supplementary Data**).

2.1.5 Deuteration Distribution Estimation

After determining the inter-experiment correspondence of peak groups with respect to a peptide sequence of interest, its deuteration distribution can be derived as (Figure 2 **E**)

$$\rho(p, k, \tau) = \frac{\beta_{\hat{\boldsymbol{g}}^{\tau}}(k)}{\sum \beta_{\hat{\boldsymbol{g}}^{\tau}}} \tag{5}$$

where the $\beta_{\hat{\boldsymbol{g}}^{\tau}}(k)$ is the abundance of the basis function corresponding to deuteration level k. The average deuteration is merely the average of the deuteration distribution over all deuteration levels.

2.2 Peptide Sequence Set Determination

To perform complete deuteration distribution estimation for the entire protein, optimized protein sequence coverage is desirable. We achieve this goal by extending the peptide sequence set via sequence search and later using a supervised classification approach to discard incorrect or ambiguous peptide sequences.

2.2.1 Unsupervised Peptide Sequence Inference

We use a two-step procedure to infer possible peptide sequences directly from the observed spectrum and from prior knowledge (i. e. the protein sequence and the MS/MS report). We first perform peak picking on the observed spectrum using the NITPICK algorithm. In a second step, the picked monoisotopic masses, for which no MS/MS identifications are available, are matched to theoretical peptide sequences extracted from the known protein sequence. Eventually, a list of candidate peptide sequences is generated, which consists of peptide sequences from two sources: peptides that are identified by MS/MS data and peptides that are extracted by searching the protein sequence for subsequences with a mass proximate to the picked peaks.

2.2.2 Supervised Quality Estimation

The unsupervised peptide sequence inference procedure exploits information without sufficient concern for multiple assignments of peptide sequences to the same peak or peptide sequences hallucinated from noise peaks. Despite the fact that this apparently improves the system's sensitivity, the payoff is a reduced specificity, i.e. false positives are mixed into the peptide sequence set. Therefore, HeXicon implements a quality estimation procedure to recover reasonable specificity while maintaining high sensitivity. We tackle this problem using a supervised classification approach: given training data $\{\boldsymbol{x}, q\}$ where $\boldsymbol{x} \in \mathcal{X}$ is the quality feature vector and $q \in \mathcal{Q}$ is the quality label, train a classifier $h : \mathcal{X} \to \mathcal{Q}$ that maps $\boldsymbol{x}$ to its estimated quality q. In particular, we use the Random Forest classifier [4], a supervised, decision-tree based ensemble learning method with high prediction accuracy and little sensitivity to the hyper-parameter settings (Figure 2 **F**).

A representative dataset was selected as training data and each reported peptide sequence was labeled with a quality score $q \in \{3, 2, 1\}$, in which 3 represents highly confident results, 2 indicates ambiguous results, i.e. unidentified peptide sequence resulting from multiple assignment to the same peak, and 1 contains all results containing no useful information. The quality features $\boldsymbol{x}$ are designed to characterize the quality of a peptide sequence from several different aspects. See the **Supplementary Data** for a full list of quality features. Retraining is necessary for different instruments.

3 RESULTS

HeXicon has been evaluated on two protein datasets of different complexity (Table 1): C terminus of Hsp70 Interacting Protein (CHIP) and High temperature protein G (HtpG). In each experiment, protein samples were first incubated in heavy water to induce a certain amount of exchange before being subjected to pepsin digestion. To identify peptic peptides from the investigated proteins we digested the undeuterated protein under the same conditions as later used for the exchange experiments. We then analyzed the peptic peptides by automated MS/MS using a 1 hour acetonitrile gradient either on a nanoLC-QSTAR MS system (CHIP, HtpG) and on a nanoLC-Orbitrap MS system (HtpG). Subsequently we determined, which of the identified peptides could be found consistently on the HPLC-QSTAR MS system using a 10 min acetonitrile gradient.

Both datasets have been processed manually, yielding average deuterations for selected peptide sequences that we use as ground truth. Segment retention time extensions are between 20s and 50s.

Measure	CHIP	HtpG
Protein length	303	636
Protein weight (kDa)	34.8	72.8
Data size (MB)	ca. 121	ca. 671
Incubation time (minutes)	0, 0.5, 5, 60	0, 5, 10, 30
Manually selected peptide sequences	21	39
Manual analysis time	ca. 2 days	ca. 1 week

Table 1: Summary of the CHIP and HtpG datasets.

Dataset	Measure	Manual Analysis	HeXicon
CHIP	Number of extracted peptide sequences	21	31
	Sequence coverage	84.2%	90.4%
	Analysis time	2 days	1 hour
HtpG	Number of extracted peptide sequences	39	90
	Sequence coverage	78.5%	85.5%
	Analysis time	1 week	3 hours

Table 2: Comparison to manual analysis: sequence coverage and analysis time.

3.1 Deuteration Distribution Estimation

For the CHIP spectra in Figure 3 (first column), HeXicon provides a sparse and condensed estimation of the deuteration distribution which exhibits smoothness along the deuteration levels, as shown in Figure 3 (second column). For comparison, we created a well-posed regression problem by constructing basis functions for the corresponding peptide CIEAKHDKYMADM and applied the non-negative least-squares regression based method described in [6]. We optimized the degree of padding by manually estimating the maximal deuteration level, yielding a solution very similar to the HeXicon's. Without optimizing the degree of padding, i.e. padding to the theoretically maximal possible deuteration level, Chik's approach selects several spurious basis functions due to overfitting (fourth column). Figure 4 (top) shows a mixture of signals from two peptide sequences: AAERERELE and IAKKKRWNSIEER. HeXicon yields condensed deuteration distributions for both peptide sequences, as shown in Figure 4 (bottom, first column). After padding optimization, Chik's approach gives a similar distribution for AAERERELE but the estimate for IAKKKRWNSIEER is questionable, see Figure 4 (bottom, second column).

3.2 Sequence Coverage Enhancement

Combining MS/MS identifications and inferred peptide sequences, HeXicon yields an apparent improvement on the number of extracted peptide sequences with concomitant increases in sequence coverage when compared to the manual analysis (Table 3.2). For the manual analysis we only used those peptides identified that we could find consistently in the 10 min gradient runs on the QSTAR system.

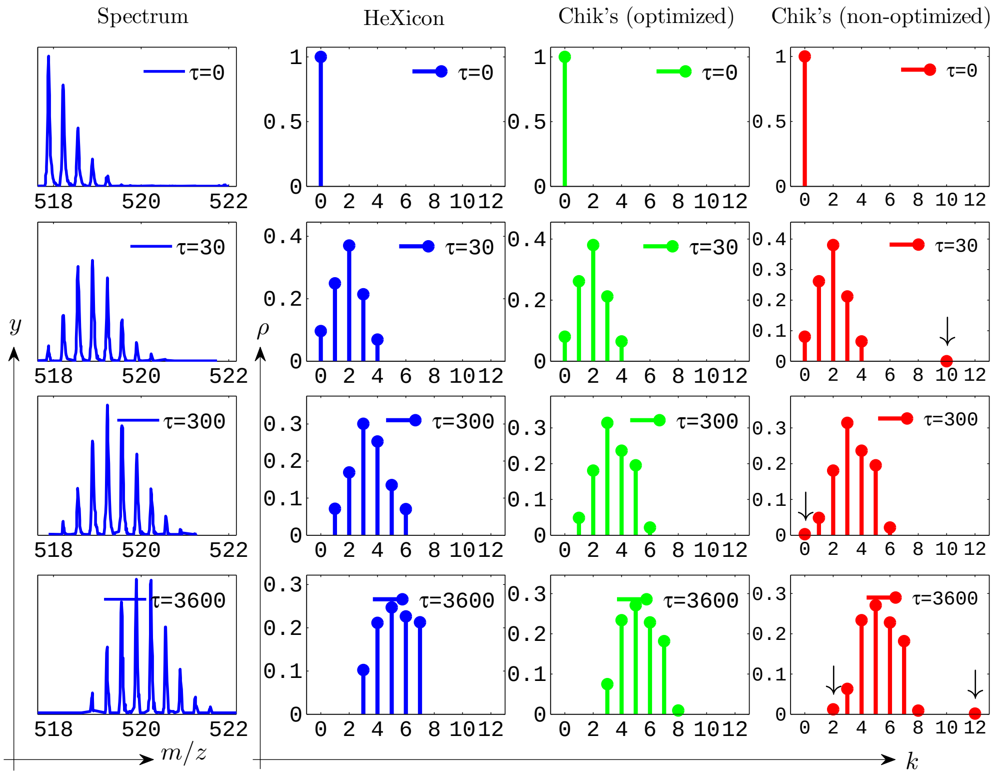


Figure 3: Comparison of deuteration distribution estimation of CIEAKHDKYMADM from a time series of 0, 30, 300 and 3600 seconds (first column). HeXicon yields condensed solution and avoids overfitting (second column). With manually optimized degree of padding, Chik's approach results in similar estimates (third column). Without padding optimization, Chik's approach selects several spurious peaks (fourth column, marked by "↓") due to overfitting.

3.3 Exchange Rate Inference

The deuteration distribution estimated by HeXicon can easily be transformed into an average deuteration estimate $\eta^{\mathrm{H}}(p,\tau)$ by computing the empirical mean. We validated HeXicon by comparing its average deuteration estimate to the manually obtained average deuteration estimate $\eta^{\mathrm{M}}(p,\tau)$. We applied two metrics to measure the accuracy: (i) the average m/z difference Δ_m is computed by $\Delta_m = \sum_\tau \left|\eta^{\mathrm{M}}(p,\tau) - \eta^{\mathrm{H}}(p,\tau)\right| / N_\tau$, where N_τ is the total number of incubation time points; (ii) the relative exchange rate difference Δ_κ is computed by $\Delta_\kappa = \left|\kappa^{\mathrm{M}} - \kappa^{\mathrm{H}}\right| / \max\left(\kappa^{\mathrm{M}}, \kappa^{\mathrm{H}}\right)$, where κ is the exchange rate inferred by fitting the average deuteration to the HX kinetic model function [14]. Since the fitting is non-linear and non-convex and since its first-order and second-order derivatives could be derived analytically, we applied a generalized Newton method to approximate the optimal solution (see **Supplementary Data**).

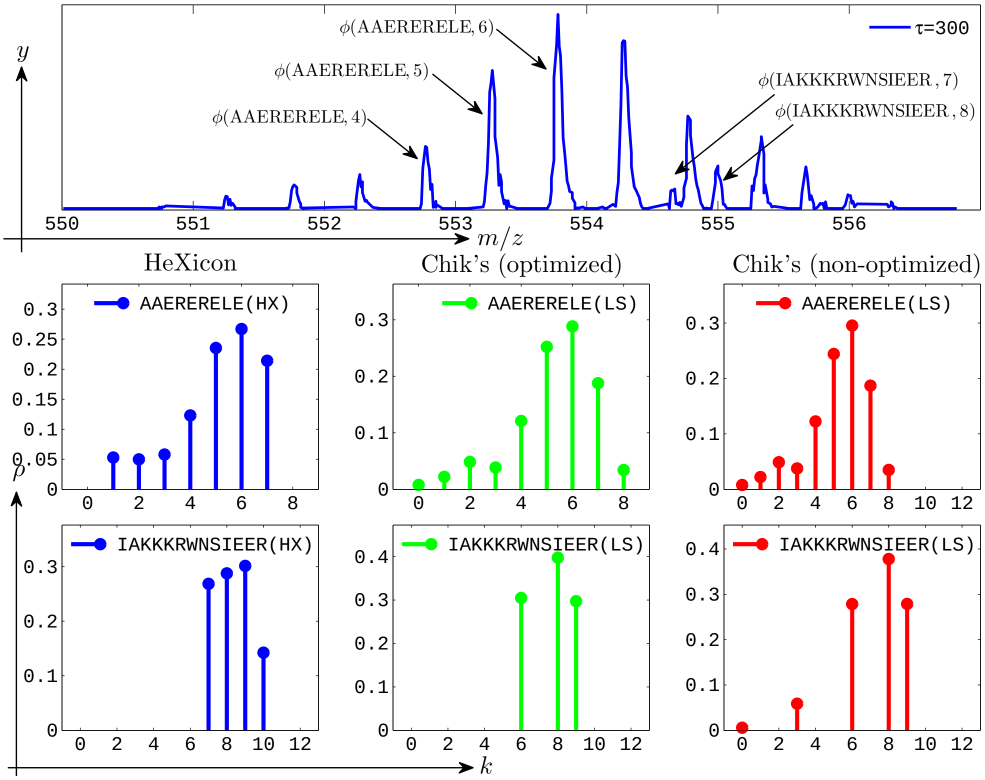


Figure 4: Comparison of deuteration distribution estimation for overlapping patterns. Overlapping patterns consist of AAERERELE and IAKKKRWNSIEER (top). HeXicon yields condensed and smooth solutions for both peptide sequences (bottom, first column). Even with padding optimization, Chik's approach overfits the spectrum and yields an unrealistic deuteration distribution for IAKKKRWNSIEER (bottom, second column). Here $\phi(p, k)$ indicates the maximum peak position of the basis function of peptide p at deuteration level k.

For the CHIP dataset and 20 of 21 manually selected peptide sequences, the estimates by HeXicon coincide well with the manual analysis (see examples in Figure 5, top-left and top-right), yielding an average m/z difference of 0.0688 ± 0.0307 Da (mean ± standard deviation) and a relative exchange rate difference of 0.0994 ± 0.0847. For the HtpG dataset, HeXicon correctly estimates the average deuteration for 32 of 39 manually selected peptide sequences and yields an average m/z difference of 0.0578 ± 0.0339 Da and a relative exchange rate difference of 0.1205 ± 0.0958 (e.g. Figure 5, bottom-left). For the remaining seven manually selected peptides, the estimates are inaccurate (e.g. Figure 5, bottom-right). The complete list of peptide sequences and their average deuteration is given in the **Supplementary Data**.

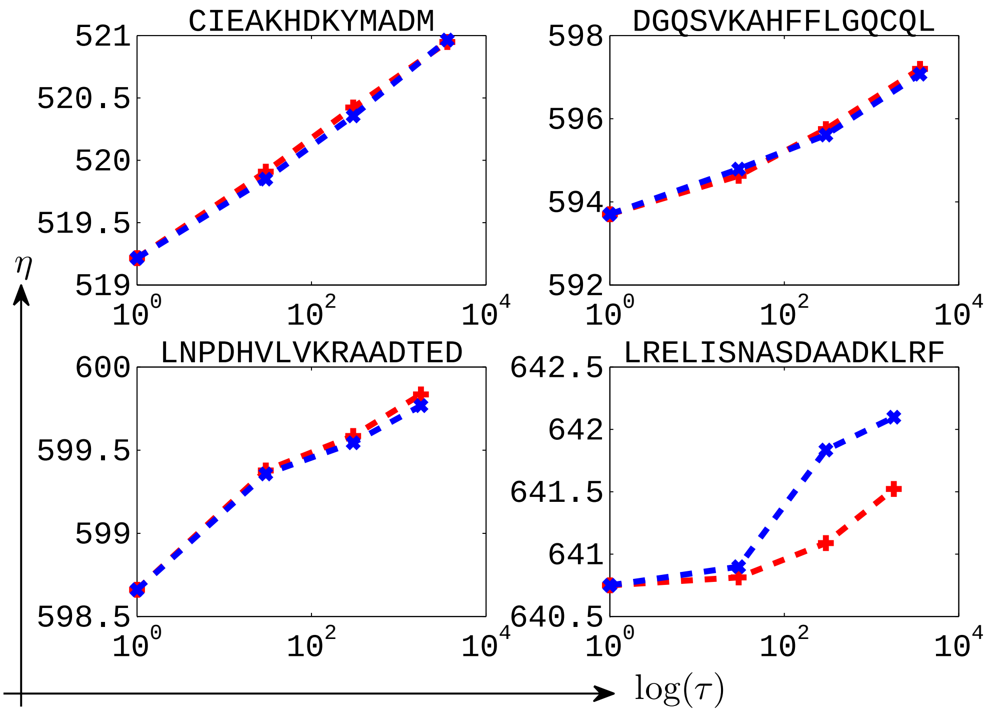


Figure 5: Comparison of the exchange rate inference between manual analysis (red) and HeXicon (blue) for selected examples. While the estimate by HeXicon coincides well with the manual analysis for the peptides displayed on the top-left, top-right and bottom-left, the estimate for LRELISNASDAADKLRF (bottom-right) is incorrect due to under-segmentation of overlapping peptides in the LC/MS spectrum.

3.4 Quality Filtering Accuracy

The quality estimation step aims at identifying high-quality results and discarding the remaining results. We measure the cross validation performance of this step using common criteria from information theory: recall, precision and F-score. The results given in Table 3 indicate that the quality estimation step is accurate and generalizes well across data sets, providing an F-score over 90%.

3.5 Runtimes and Implementation

HeXicon has been implemented in C++ and the compiled software is available at http://hci.iwr.uni-heidelberg.de/software.php As indicated in Table 3.2, HeXicon strongly reduces the analysis time compared to manual analysis. Since HeXicon is fully automated, it does not require any real-time user-interaction. Experiments were carried out without replicates. In order to perform replicate analysis, HeXicon results need to be obtained separately for

Measure	Class 1	Class 2	Class 3
Recall	98.8	91.6	92.2
Precision	98.7	92.9	89.2
F-score	98.7	92.2	90.7

Table 3: Cross validation performance: recall, precision and F-score (in %) are given for high quality (Class 3), medium quality (Class 2) and low quality (Class 1) results.

each replicate and subsequently aggregated. The software package requires the spectrum data as mzXML files and other information (i.e. the protein sequence and the MS/MS search result) as plain text files. CSV files are the output.

4 DISCUSSION

As shown in section 3.1, to avoid overfitting Chik's approach requires padding optimization by user-input or pre-processing. The reason is that the least-squares regression attempts to use each predictor without any restriction and thus overfits the data and causes several spurious basis functions to be selected, as shown in Figure 3 (fourth column, marked by "↓"). The proposed approach benefits from the sparsity of the $L1$-regularization and discards those spurious deuteration levels automatically, and thus requires no additional processing such as thresholding or any further user interaction. This overfitting problem becomes worse when overlapping patterns occur. As shown in Figure 4 (bottom, right column), Chik's approach (with padding optimization) gives a reasonable distribution for AAERERELE, but yields an unrealistic estimate for IAKKKRWNSIEER, i.e. the large gaps between neighboring deuteration levels. HeXicon, on the other hand, keeps the intrinsic smoothness and sparsity of the deuteration levels. Although the estimate for the low-intensity IAKKKRWNSIEER is subject to low SNR, it is still represented by a compact deuteration distribution at the most relevant positions and appears to be physically reasonable. While maximum entropy deconvolution based methods [25] might theoretically be appealing, they are not applicable to the problem since they require a pre-defined noise level [1] which is usually not available to the users and may vary among different m/z regions or experiments. Further, these approaches are prone to overfitting and are computationally expensive [13].

The improved sequence coverage provided by HeXicon is particularly helpful to gain a more complete and detailed understanding of a dataset. Due to under-segmentation of crowded regions in the LC/MS data, HeXicon did not recover all manually selected peptide sequences from the HtpG dataset, but it still managed to yield a higher sequence coverage because other peptide sequences were selected to compensate for the missing ones. Further, as shown in Table 3.2, HeXicon finds more than twice the number of peptide sequences selected by human experts, which allows exchange behavior prediction in finer regions. For instance, the estimation of exchange rate at positions 279-284 can be inferred from both HLQRVGHFD-PVTRSPLTQEQLIPNL (position 259-284) and HLQRVGHFDPVTRSPLTQE (position 259-278). Since we only considered those HeXicon results with the highest quality score for the computation of the sequence coverage, this number can be regarded as a conservative estimate. Additional lower quality results provided by HeXicon can guide users towards further targeted experiments. For instance, ambiguous results, when multiple peptide sequences could be assigned to the same spectrum, might motivate additional MS/MS run on specific peptide sequences of interest, and thereby allow further improvement on the sequence coverage.

5 CONCLUSION

In this article, we introduced HeXicon, a novel algorithmic workflow for the robust estimation of deuteration distributions with increased sequence coverage for HX/MS experiments. Comparisons to previous methods showed that the $L1$-regularization adopted in our method provides a sparse estimation of deuteration distributions and avoids over-fitting. The overall sequence coverage is increased by inferring peptide sequences from prior knowledge, and the tradeoff between sensitivity and specificity is balanced using a supervised classification procedure.

In comparison to manual analysis, we showed that HeXicon succeeds in accurately extracting the deuteration content while improving sequence coverage and reducing analysis time.

Acknowledgement

The authors would like to thank Thorben Kröger, Anna Kreshuk and Michael Hanselmann (Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany) for comments, suggestions, and fruitful discussions. The authors also gratefully acknowledge financial support by the Heinz Goetze Memorial Fellowship (X.L.), the Helmholtz Alliance on Systems Biology (X.L., F.A.H.), the Karl-Steinbuch-Fellowship (B.Y.R.), the DFG (SFB 638: M.P.M., C.G., C.T.L.), the DFG (HA4364/2-1:B.Y.R., F.A.H.), and the Alexander von Humboldt-Foundation (3.1-DEU/1134241: M.K.).

References

[1] R. R. Abzalimov and I. A. Kaltashov. Extraction of local hydrogen exchange data from HDX CAD MS measurements by deconvolution of isotopic distributions of fragment ions. *J Am Soc Mass Spectrom*, 17(11):1543–1551, 2006.

[2] A. H. P. America and J. H. G. Cordewener. Comparative LC-MS: A landscape of peaks and valleys. *Proteomics*, 8(4):731–749, 2008.

[3] S. Boppel, B. Y. Renard, M. Kirchner, H. Steen, J. A. J. Steen, U. Koethe, and F. A. Hamprecht. Sparse profile reconstruction for LC/MS feature extraction. In *American Society for Mass Spectrometry.* Annual Conference, 2008.

[4] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[5] I. V. Chernushevich, A. V. Loboda, and B. A. Thomson. An introduction to quadrupole-time-of-flight mass spectrometry. *J Mass Spectrom*, 36(8):849–865, 2001.

[6] J. K. Chik, J. L. V. Graaf, and D. C. Schriemer. Quantitating the statistical distribution of deuterium incorporation to extend the utility of H/D exchange MS data. *Anal Chem*, 78(1):207–214, 2006.

[7] C. M. Dobson. Protein folding and misfolding. *Nature*, 426(6968):884–890, 2003.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann Stat*, 32(2):407–451, 2004.

[9] J. R. Engen. Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. *Anal Chem*, 81(19):7870–7875, 2009.

[10] S. W. Englander. Hydrogen exchange and mass spectrometry: a historical perspective. *J Am Soc Mass Spectrom*, 17(11):1481–1489, 2006.

[11] M. Guilhaus. Special feature: Tutorial. Principles and instrumentation in time-of-flight mass spectrometry. Physical and instrumental concepts. *J Mass Spectrom*, 30(11):1519–1532, 1995.

[12] A. N. Hoofnagle, K. A. Resing, and N. G. Ahn. Protein analysis by hydrogen exchange mass spectrometry. *Annu Rev Biophys Biomol Struct*, 32(1):1–25, 2003.

[13] M. Hotchko, G. S. Anand, E. A. Komives, and L. F. Ten Eyck. Automated extraction of backbone deuteration levels from amide H/2H mass spectrometry experiments. *Protein Science*, 15(3):583–601, 2006.

[14] L. Konermann, X. Tong, and Y. Pan. Protein structure and dynamics studied by mass spectrometry: H/D exchange, hydroxyl radical labeling, and related approaches. *J Mass Spectrom*, 43(8):1021–1036, Aug 2008.

[15] P. Nikamanon, E. Pun, W. Chou, M. D. Koter, and P. D. Gershon. "TOF2H": A precision toolbox for rapid, high density/high coverage hydrogen-deuterium exchange mass spectrometry via an LC-MALDI approach, covering the data pipeline from spectral acquisition to HDX rate analysis. *BMC Bioinformatics*, 9:387, 2008.

[16] M. Palmblad, J. Buijs, and P. Håkansson. Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions. *J Am Soc Mass Spectrom*, 12(11):1153–1162, 2001.

[17] B. D. Pascal, M. J. Chalmers, S. A. Busby, and P. R. Griffin. Hd desktop: an integrated platform for the analysis and visualization of h/d exchange data. *J Am Soc Mass Spectrom*, 20(4):601–610, Apr 2009.

[18] B. D. Pascal, M. J. Chalmers, S. A. Busby, C. C. Mader, M. R. Southern, N. F. Tsinoremas, and P. R. Griffin. The Deuterator: software for the determination of backbone amide deuterium levels from H/D exchange MS data. *BMC Bioinformatics*, 8(1):156, 2007.

[19] B. Y. Renard, M. Kirchner, H. Steen, J. A. J. Steen, and F. A. Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9:355, 2008.

[20] G. Schwarz. Estimating the dimension of a model. *Ann Stat*, 6(2):461–464, 1978.

[21] G. W. Slysz, C. A. H. Baker, B. M. Bozsa, A. Dang, A. J. Percy, M. Bennett, and D. C. Schriemer. Hydra: software for tailored processing of H/D exchange data from MS or tandem MS analyses. *BMC Bioinformatics*, 10(1):162, 2009.

[22] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J Roy Statist Soc Ser B*, 58(1):267–288, 1996.

[23] T. E. Wales and J. R. Engen. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom Rev*, 25(1):158–170, 2006.

[24] D. D. Weis, J. R. Engen, and I. J. Kass. Semi-automated data processing of hydrogen exchange mass spectra using HX-Express. *J Am Soc Mass Spectrom*, 17(12):1700–1703, 2006.

[25] Z. Zhang, S. Guan, and A. G. Marshall. Enhancement of the effective resolution of mass spectra of high-mass biomolecules by maximum entropy-based deconvolution to eliminate the isotopic natural abundance distribution. *J Am Soc Mass Spectrom*, 8(6):659–670, 1997.