J. Am. Soc. Mass Spectrom. (2011) 22:856–866
DOI: 10.1007/s13361-010-0055-y

RESEARCH ARTICLE

Automated Lipid A Structure Assignment from Hierarchical Tandem Mass Spectrometry Data

Ying S. Ting,[1] Scott A. Shaffer,[1,5] Jace W. Jones,[1,2] Wailap V. Ng,[3] Robert K. Ernst,[4] David R. Goodlett[1]

[1]Department of Medicinal Chemistry, University of Washington, Box 357610, Seattle, WA 98195-7610, USA
[2]Jones Environmental, Inc., Fullerton, CA, USA
[3]Department of Biotechnology and Laboratory Science in Medicine, Institute of Biotechnology in Medicine, Institute of Biomedical Informatics, and Center for Systems and Synthetic Biology, National Yang Ming University, Taipei, Taiwan
[4]Department of Microbial Pathogenesis, University of Maryland, Baltimore, MD, USA
[5]University of Massachusetts, Medical School, Worcester, MA, USA

Abstract

Infusion-based electrospray ionization (ESI) coupled to multiple-stage tandem mass spectrometry (MS^n) is a standard methodology for investigating lipid A structural diversity (Shaffer et al. J. Am. Soc. Mass. Spectrom. 18(6), 1080–1092, 2007). Annotation of these MS^n spectra, however, has remained a manual, expert-driven process. In order to keep up with the data acquisition rates of modern instruments, we devised a computational method to annotate lipid A MS^n spectra rapidly and automatically, which we refer to as hierarchical tandem mass spectrometry (HiTMS) algorithm. As a first-pass tool, HiTMS aids expert interpretation of lipid A MS^n data by providing the analyst with a set of candidate structures that may then be confirmed or rejected. HiTMS deciphers the signature ions (e.g., A-, Y-, and Z-type ions) and neutral losses of MS^n spectra using a species-specific library based on general prior structural knowledge of the given lipid A species under investigation. Candidates are selected by calculating the correlation between theoretical and acquired MS^n spectra. At a false discovery rate of less than 0.01, HiTMS correctly assigned 85% of the structures in a library of 133 manually annotated *Francisella tularensis* subspecies *novicida* lipid A structures. Additionally, HiTMS correctly assigned 85% of the structures in a smaller library of lipid A species from *Yersinia pestis* demonstrating that it may be used across species.

Key words: HiTMS, Hierarchical, Multi-stage, Mass spectrometry, High-throughput, Lipid A, Lipidomic, Algorithm, Automated, Structural, Cross-correlation

Introduction

Lipid A, the endotoxic portion of lipopolysaccharides (LPS) responsible for gram-negative bacterial virulence, is imbedded in the outer leaflet of the gram-negative bacterial outer membrane. As an essential component of gram-negative bacterial membranes, lipid A exhibits species-specific structural diversity [1]. The general structure consists of a backbone of two glucosamine residues present as a β(1-6)-linked dimer [2]. This backbone can be diversified in response to specific environmental signals or between bacterial species. Specifically, changes in the fatty acid content varying both in the length and number of fatty acid side chains (e.g., tetra-acylated or hexa-acylated) and

Electronic supplementary material The online version of this article (doi:10.1007/s13361-010-0055-y) contains supplementary material, which is available to authorized users.

Correspondence to: David R. Goodlett; *e-mail:* goodlett@uw.edu

Received: 7 October 2010
Revised: 10 December 2010
Accepted: 10 December 2010
Published online: 5 March 2011

phosphorylation patterns (Supplemental Figure S.1) [3]. Additional glycan modifications of the phosphate residues by monosaccharides, such as aminoarabinose or galactosamine are also observed [2, 4]. Environmentally induced changes to lipid A diversity is not only observed among species, but also within species (Supplemental Figure S.1). Variability in lipid A structures is an adaptive mechanism that increases bacterial survival, often increasing resistance to host killing mechanism or in the avoidance of the host innate immune system [5, 6]. The diversity of such species and environmental-driven structural modification makes complete lipid A structural analysis challenging.

To date, the single most important structural tool for characterization of lipid A is tandem mass spectrometry (MS/MS). Due to the existence of structural isomers, commonly referred to as isobars present in individual lipid A preparations, the use of a hierarchical, multi-stage tandem mass spectrometry (MS^n) approach coupled with accurate mass assignment proved to be an essential, albeit labor intensive, tool in our prior work to understand lipid A structural diversity [7]. This hierarchical MS^n strategy acquires tandem mass spectra on each precursor ion and all of their derived fragment ions until ion signal for a given precursor-fragment ion lineage drops below the threshold (Figure 1). In this method a series of precursor ion genealogies are revealed that allow structures to be derived from the recorded parent-fragment ion lineages. However, the acquisition process is automated and leads to a large number of tandem MS files; each precursor having multiple fragment ion lineages consisting of MS^2 to MS^4 data. In our prior work on *Francisella tularensis* subspecies *novicida* (Fn) grown under two different environmental conditions, this hierarchical MS^n strategy revealed 30 unique lipid A species from only a half dozen unique precursor ions [8, 9]. While we demonstrated that high-throughput (HTP) hierarchical MS^n profiling of lipid A mixtures can accelerate our discovery of structural diversity, a new bottleneck was created by the large amounts of data that had to be manually interpreted [6, 7, 10, 11]. Thus, a major challenge for further implementation of our hierarchical MS^n lipidomic strategy lay in the computational and bioinformatic demands of handling the large amounts of data [12]. Therefore, to achieve a viable HTP lipid A structural analysis scheme an automated approach for structure assignment was needed.

A thorough review of the literature revealed a number of software tools available for lipid structure assignment; however, none of these tools are compatible with lipid MS^n data analysis. Some of these tools are limited to the type of instrument used to obtain the data, such as Fatty Acid Analysis Tool (FAAT), Lipid Profiler and Lipid Inspector [13–16]. Some are limited to the lipid classes covered by their fragment ion databases, such as LipidQA [17]. Another software tool lipID supported more comprehensive lipid classes, but unfortunately lacked the ability to analyze MS^n data [18]. In 2009, AMDMS-SL was developed to automate the analyses of multidimensional mass spectrometry (MDMS)-based shotgun lipidomics [19, 20]. AMDMS-SL was able to identify individual lipid molecular species within a predetermined individual lipid class [21]. However, the general structure of lipid A is very different from the majority of lipid classes, e.g., glycerolipids, glycerophospholipids, and sphingolipids, which were well characterized by AMDMS-SL built-in database. In addition, the initial analysis of structures by tandem mass spectrometry alone is often not sufficient to decipher the structures of lipid A, which we have found often require exhaustive MS^n analysis to successfully characterize the isobaric components [7, 10, 11, 22, 23].

Thus, while there are a number of computational tools available for the analysis of ESI generated data of lipids in general, none deal with the structural diversity of lipid A specifically and none have the ability to analyze HTP hierarchical MS^n mass spectra. As mentioned previously, recent work from our laboratory demonstrated the use an infusion-based HTP ESI-MS^n strategy to characterize lipid A structural diversity in Fn [7]. In this study, 30 lipid A structures were determined by manual spectral interpretation for two growth conditions (25 °C and 37 °C). The extensive tandem mass spectral library generated in this prior study was used as the initial training dataset for the development of an automated structure assignment tool that we term hierarchical tandem mass spectrometry (HiTMS) algorithm. Data analysis by HiTMS is based on predicted ion dissociations and complementary neutral losses, including fatty acids, phosphate, and monosaccharide substituents that are used to construct a species-specific library. Here, we describe HiTMS as well as use of a cross-correlation scoring routine to assign individual lipid A structures objectively from complex mixtures of Fn and *Yersinia pestis* (Yp).

Materials and Methods

Preparation of Bacterial Lipid A

Francisella tularensis subspecies *novicida* (Fn) strain U112 was grown with aeration in tryptic soy broth (Gibco BRL, Grand Island, NY, USA) supplemented with 0.1% cysteine at 25 °C and harvested in the stationary phase. Yp strain KIM6+ was grown in Luria broth (pH 7.4) at 37 °C with aeration and harvested in the late exponential phase referred to as Yp wild type (Yp_WT) [24]. Lipid A C-1 and C-4' phosphatase, LpxE and LpxF, respectively, have been expression cloned in Fn [25]. The individual plasmids with the structural genes of LpxE or LpxF and an ampicillin resistance gene were incorporated into KIM6+ cell via electroporation [11]. The phosphatase expressing strains were then grown in Luria broth containing 100 μg/mL ampicillin (pH 7.4) at 37 °C with aeration and harvested in the late exponential phase, and designated as Yp_LpxE and Yp_LpxF. Fn and Yp LPS were extracted using the hot phenol/water extraction method as previously described [26]. Lipid A

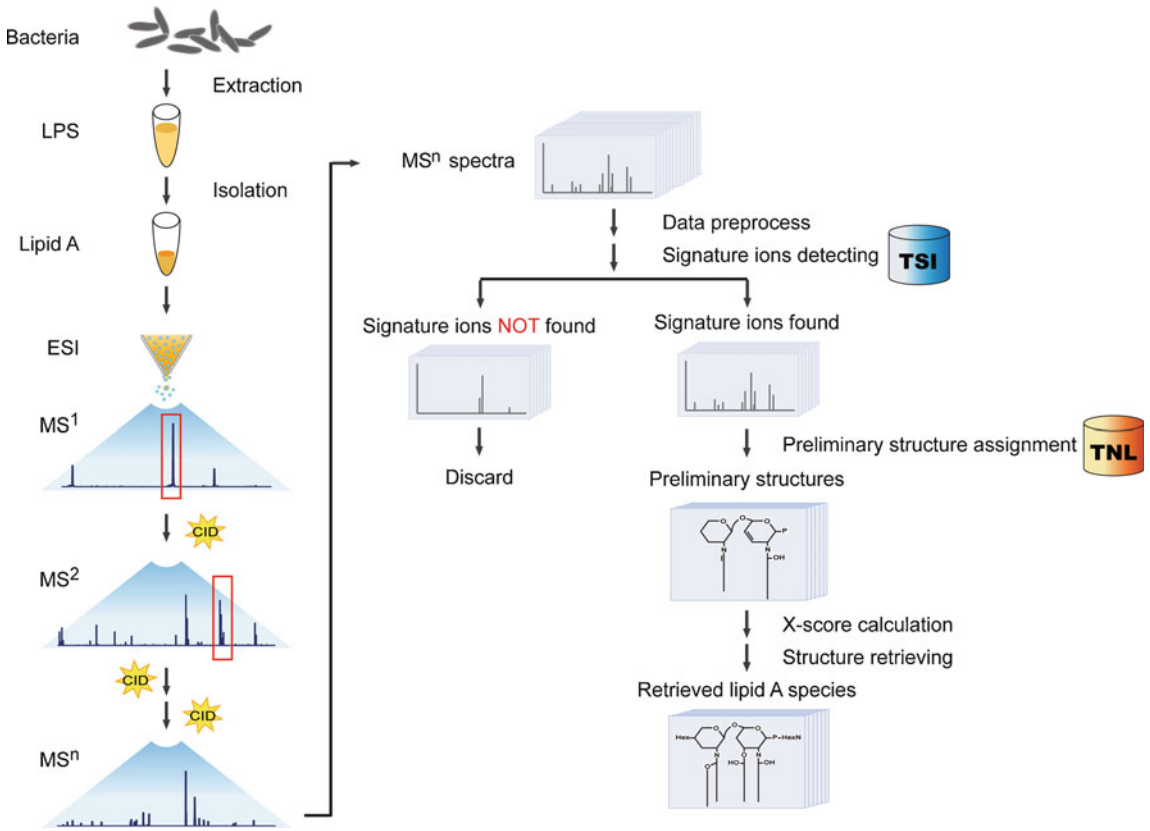

Figure 1. Overview of HiTMS interpretation of lipid A hierarchical ESI-MS^n data. Bacterial lipid A is isolated from LPS extraction and analyzed by ESI tandem mass spectrometry with hierarchical MS^n strategy that acquires tandem mass spectra on each precursor ion and all of the derived fragment ions. The collection of MS^n spectra is searched against the theoretical signature ion (TSI) database for observed signature ions. The neutral losses of signature ions in each spectrum are then searched against the theoretical neutral losses (TNL) database to identify dissociation formulae. Lipid A preliminary structures for each MS^n spectral set are then proposed. Every assignment of preliminary structures is given a X-score based on the correlation between theoretical and acquired spectra. All candidate structures that pass the X-score cutoff are considered as accurate assignments

was isolated after LPS was treated with RNase A, DNase I, and proteinase K by the method of Caroff et al. [27].

Mass Spectrometric Analysis

The isolated Fn lipid A was analyzed by electrospray ionization (ESI) in the negative ion mode on a hybrid linear ion trap Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (LTQ-FT) (Thermo Scientific, San Jose, CA, USA). Lipid A was prepared at ~0.5 mg/mL in methanol/chloroform (2:1) and infused at 1.0 µL/min into a heated capillary inlet maintained at 400–450 °C. MS^n spectra were acquired according to a "target" MS scheme predetermined from previous studies [7]. Briefly, in this scheme 15 deprotonated molecular ions were selected individually for MS^2 for the initial loss of 12:0 (number of carbons:number of double bonds), 14:0, 16:0, 18:0, and 20:0 fatty acids, each of which were determined previously to be esterified through the 2' position fatty acid at the 3-hydroxy position of the lipid A deprotonated anions. Each of the subsequent ions was selected for MS^3 for the combined loss of galactosamine and the 3-position 3-hydroxy fatty acid (12:0, 14:0, 16:0, 18:0, and 20:0). Ions representing these combined losses were in turn selected for MS^4 and monitored for the observation of Y_1/Z_1 ion pairs using 1 min scan averaging. Each MS^2 and MS^3 "channel" was selected regardless of product ion spectra observed [28], and only the MS^4 were used for the determination of the individual lipid A structures. Ion population in the LTQ was set at 10,000 and collision energies employed for MS^n ranged from 25%–35%. For Yp Lipid A, data were acquired on an LTQ-FT Ultra (Thermo Scientific) as described elsewhere [10, 11].

Theoretical Databases Construction

Theoretical database constructor program was written with Perl ver. 5.8.8 (http://www.perl.org) built for x86_64-Linux platform. A species-specific theoretical database was constructed based on the manual interpretation of lipid A fragmentation rules in tandem mass spectra, which included phosphate patterns as well as fatty acid and monosaccharide

substituents. Direct bond cleavages of lipid A structures were considered as the general template for fragmentation and structural inference.

Each species-specific theoretical database contains two sub-databases for: (1) theoretical signature ions (TSI), and (2) theoretical neutral losses (TNL) (Figure 2). Observed signature ions are unique ions that help hypothesize the molecule's structure. The observed lipid A signature ions were usually determined from the conserved characteristic of lipid A diglucosamine and named according to the nomenclature described by Domon and Costello [29]. Based on the observed fragmentation templates of lipid A, signature ions were calculated and put into the theoretical signature ion (TSI) database. To increase the structural diversity of lipid A represented in the TSI database, a user-defined carbon range of fatty acids was applied (i.e., 12:0 to 20:0 fatty acids). By systematically altering the fatty acid side chain lengths and positions, all possible signature ions were computed and incorporated into the TSI database. To facilitate the structure assignment, neutral losses of signature ions were also calculated and put in the theoretical neutral loss (TNL) database. Additionally, common observed neutral losses that come from direct bond cleavages of lipid A other than cleavages of signature ions were also included in the TNL database. Similarly, to increase the structural diversity covered by TNL databases, fatty acid compositions of TNL were systematically altered within the user-defined carbon range.

DeltaMass is a user assigned HiTMS parameter that defines the mass tolerance used to represent the acceptable mass difference between theoretical and observed ions. DeltaMass was applied in all searches against the TSI and TNL databases using the values consistent with the mass accuracy of the acquired data.

Data Preprocessing

Raw data files were converted into mzXML data format by ReAdW, available in Xcalibur software (Thermo Scientific). The peak information from either individual or averaged mass spectra were then extracted using MassSpecWavelet, a wavelet transform based peak detection software provided by the Bioconductor project (http://www.bioconductor.org/) [30]. Resulting peak information of each MS^n tandem mass spectrum was recorded in a peak list file (referred as MS^n spectra hereafter).

Hierarchical Tandem Mass Spectrometry (HiTMS) Algorithm

HiTMS was implemented in Perl ver. 5.8.8 (http://www.perl.org) and run on a 64-bit GNU/Linux platform. Acquired MS^n spectra were searched against TSI database to find possible signature ions and spectra without any identifiable signature ions were discarded (Figure 1). Any identified signature ions suggest formulae corresponding to the reducing and/or non-reducing portions of lipid A. By subtracting the mass of signature ions from their precursors, the neutral losses of signature ions are subsequently calculated and searched against the TNL database. The combination of signature ions and matched neutral losses provides a preliminary candidate structure (Figure 3). The calculated neutral losses of all the ions in each spectrum were also searched against the TNL database to provide needed information for spectrum annotation. To each lipid-spectrum match (LSM) an X-score is applied to evaluate the closeness of fit between every MS^n spectrum and its preliminary candidate structures (see Section 2.6). After preliminary structures were assigned, neutral loss of every MS^n spectrum's precursor ion

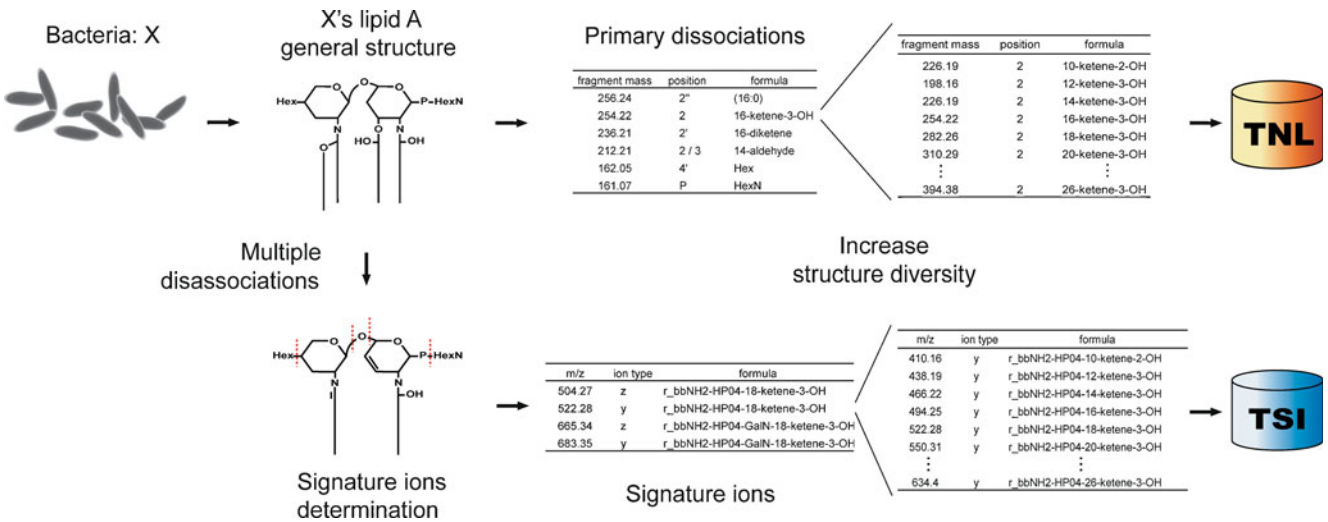

Figure 2. Flowchart of the construction of theoretical databases. Lipid A general structure based on previous studies are used as a template for theoretical signature ion (TSI) and theoretical neutral loss (TNL) database construction. For the TNL database, fragment ion masses of primary dissociations are incorporated. The theoretical formula and masses of dissociations are calculated within a range of fatty acid lengths. Based on this logic, theoretical *m/z* values of signature ions are calculated and stored in the TSI database

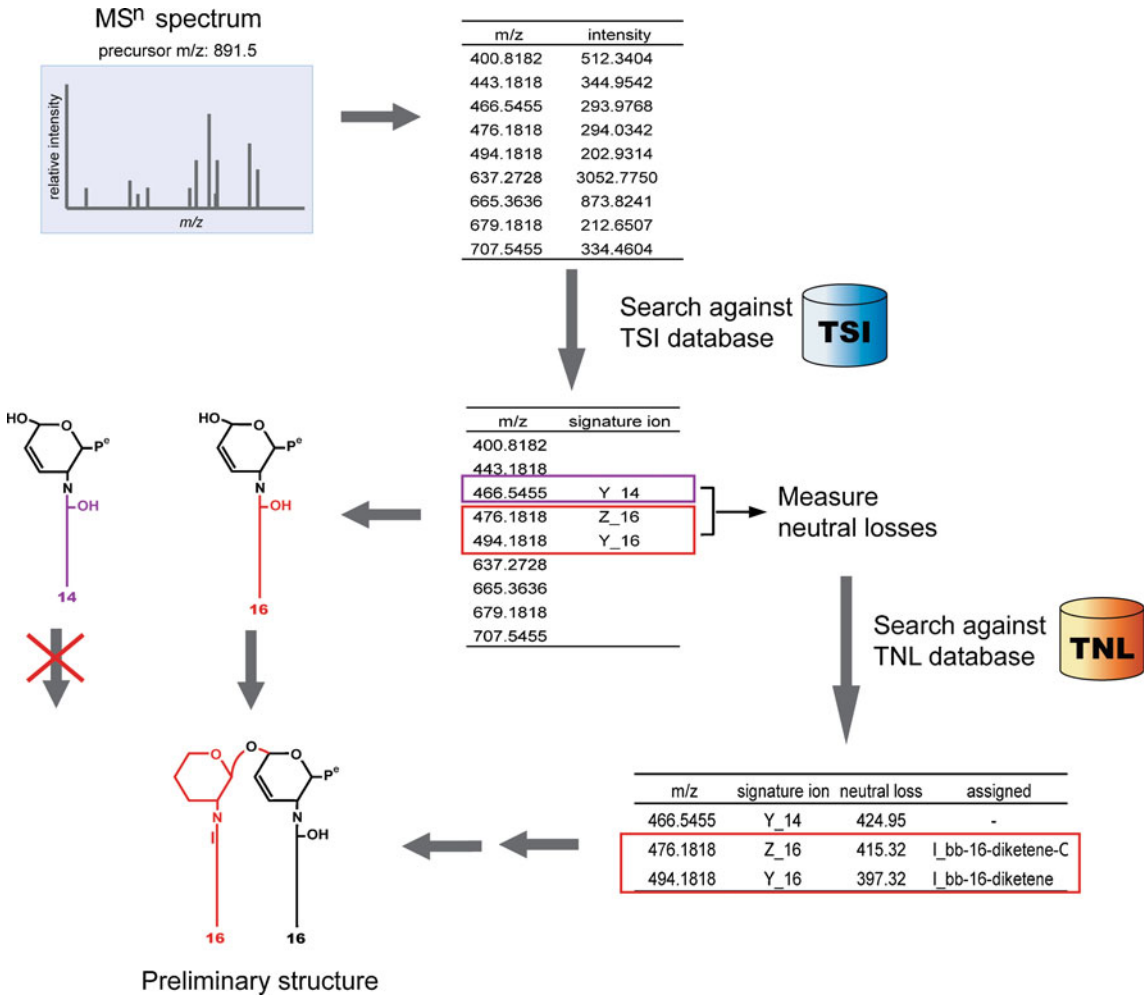

Figure 3. Flowchart of preliminary candidate structural hypotheses. A peak list is extracted from MS^n spectra by MassSpecWavelet peak detection algorithm. The extracted *m/z* peaks are searched against the theoretical signature ion (TSI) database for signature ions, which suggest the formulae of the reducing end of lipid A. Neutral losses of all possible ion indicators are measured and searched against the theoretical neutral losses (TNL) database. Identified neutral losses provide information on the remaining structures. Preliminary structures are then proposed by combining the complementary portions

was calculated in the corresponding MS^{n-1} spectrum and searched against TNL database again to identify the possible dissociation patterns. HiTMS continues the above procedures in an iterative manner until the MS^1 level is reached. The final structures are deduced by integrating the information gained from the different levels of MS^n data.

Cross Correlation (X-score)

The X-score uses a closeness of fit measurements between an acquired and theoretical tandem mass spectrum similar to SEQUEST xcorr [31, 32]. For every LSM, hypothetical lipid A structure is fragmented in silico based primarily on aforementioned direct bond cleavages, including glycosidic bond cleavages (i.e., A/X, B/Y, C/Z type ions), losses of O- and N-linked acyl chains, losses of phosphate, losses of monosaccharide, and perturbations representing combined losses. Fragmentations are later combined into a reconstructed mass spectrum representing the theoretical dissociation of the candidate structure. The peak intensity of each reconstructed mass spectrum is assigned a Boolean value where 1 represents the existence of a fragmentation of such *m/z* value. The X-score between the acquired mass spectrum and the reconstructed mass spectrum of hypothetical structure is measured as follows:

$$X - score = x_0 \cdot y' \quad where \;\; y' = y_0 - \left[\sum_{\tau=-75, \tau \neq 0}^{\tau=+75} y_\tau \right] / 150$$

Each X-score calculation is a scalar dot product between reconstructed mass spectrum x and the preprocessed acquired mass spectrum y′ where τ is the correction factor, as described in previous publications [31, 32]. DeltaMass is used as the bin size to convert mass spectra into vectors. X-score is used by HiTMS to measure the closeness of fit of every LSM.

On-The-Fly Decoy Generation

In the world of proteomics, a decoy database is often employed to help evaluate the significance of peptide spectra matches. A decoy database comprises protein sequences that have been shuffled or reversed, generated from the given target database beforehand or on-the-fly [33–35]. HiTMS uses this target-decoy strategy, generating decoys by shuffling the candidate lipid A structure on-the-fly while analyzing each MS^n spectrum. To avoid destroying the lipid A biochemistry, shuffling only occurs on the position and length of fatty acid side chains. This approach ensures that every decoy lipid A exhibits precisely the same molecular composition and mass as the target (i.e., candidate) lipid A structures. X-score of both candidate and decoy LSM are then calculated to help evaluate the significance.

Results and Discussion

Manual Structural Analysis of Fn Lipid A

In prior work, we demonstrated the power of an infusion-based high-throughput (HTP) hierarchical ESI-MS^n strategy (Figure 1) to generate lipid A tandem mass spectra [7]. In this strategy, each precursor ion and all subsequent fragment ions for each generation up to the fourth level were fragmented in a hierarchical MS^n fashion. For example, a precursor ion might generate a set of fragment ions A_1 ... A_n, where n is the number of fragments produced by precursor ion A. Next, each of these first generation fragment ions was fragmented to produce a series of second generation fragment ions $A_{1,1}$ ... $A_{1,m}$ where m is the number of second generation fragment ions produced by fragment ion A_1 such that the process of continued up to the level of the fourth generation where ion intensity generally dropped below the detectable threshold. Thereby, each precursor ion had a hierarchical set of tandem mass spectra (similar to a surname genealogical tree that branches out from the progenitor parental line) associated with it (Figure 3) that could be used to aid structure assignment. While lipid A structures were assigned manually in our prior efforts [7, 10, 11], the objective of the current work was to develop HiTMS, an automated structure assignment algorithm, and demonstrate its effectiveness on two structural different types of lipid A. To this end we first tested HiTMS's accuracy for high throughput structure assignments on a library of 133 unique Fn lipid A structures involving 58 variations of fatty acid combinations. This library consisted of a set of 30 previously published structures [7] and an additional set of 103 unpublished structures, all of which were manually assigned; see Supplemental Table S.1 for new structures. The 133 unique lipid A structures in the library were derived from 49,943 tandem mass spectra (i.e., MS^1 up to MS^4). While HiTMS is capable of analyzing individually all 49,943 tandem mass spectra, our manually derived library of 133 structures came from 284 unique mass spectra produced by averaging 1 min intervals of infusion data. Here, we focused HiTMS analysis on only these 284 averaged spectra for which structures had been manually confirmed. In order to determine the accuracy of structure assignments by HiTMS the 284 averaged mass spectra including 7 MS^1, 16 MS^2, 55 MS^3, and 206 MS^4 spectra, were analyzed by HiTMS in an automated manner.

Species-Specific Construction of TSI and TNL Databases

Analogous to the use of species-specific genomic databases in proteomics, HiTMS uses species-specific theoretical databases to identify the origin of fragment ions. These species-specific theoretical databases require some basic knowledge of the lipid A structural configuration under investigation, which may be inferred initially from the precursor ion spectrum. For example, in the case of lipid A isolated from Fn, the theoretical signature ions were derived primarily from Y- and Z-type ions as per Domon and Costello [29]. These theoretical signature ions were deposited in a species-specific Fn-TSI database including the combination of Y_1/Z_1ions with a C-1 phosphate, one of 12:0 to 20:0 fatty acids (where chain length range is user-defined) on 2-position, and a potential galactosamine (GalN) substitution on C-1 phosphate. Additionally, the fact that Fn lipid A only has a C-1 phosphate that gives characteristic Y_1/Z_1ions was included as signature ions while cross-ring cleavages were not included. Finally, all theoretical neutral losses that could be generated from direct bond cleavage of Fn lipid A were deposited in Fn-TNL databases, including (n:0) -3-OH ketene, (n:1) ketene, (n:0) acid, GalN, glucosamine (GlcN), and a phosphate group; n is the range of carbon atoms in the fatty acids from 12 to 20. In addition, the neutral losses resulting from glycosidic cleavages were also included in Fn-TNL.

In the case of lipid A isolated from Yp, theoretical signature ions were derived primarily from A-type ions. Unlike Fn, Yp lipid A is diphosphorylated and required unique TSI and TNL databases. The phosphorylation patterns of lipid A in Yp are C-1 and C-4'bisphosphate, C-1 pyrophosphate, and C-4' pyrophosphate [11]. Yp lipid A has also been observed to be heavily modified with up to two aminoarabinose (Ara4N) moieties [10]. At the mammalian host temperature of 37 °C, the major Yp lipid A structures consists of a β-1,6-linked diglucosamine backbone with two phosphate groups and four primary 14:0 fatty acids [36–38]. Such tetra-acylated lipid A with four identical fatty acids distributed evenly on the reducing and non-reducing end is likely to result in symmetric lipid A structures. The symmetric pattern of bisphosphorylated lipid A produces B/Y- and C/Z-type ions that fail to distinguish reducing end from non-reducing end. Thus, A-type ions are crucial in Yp structure assignment because they fragment across the reducing glycan and result in distinguishable reducing and non-reducing fragments providing unique signature ions [10, 11].

We also examined two Yp phosphatase mutant strains, LpxE and LpxF, which overexpress phosphatases resulting

in the dephosphorylation of the 1- and 4'-positions, respectively, yielding asymmetric lipid A. To account for this asymmetry, both A-type ions as well as Y- and Z-type ions were also included in Yp-TSI database to detect a C-1 phosphate modification. In total, the Yp-TSI database consisted of $^{0,2}A_2$, $^{0,4}A_2$, Y_1, and Z_1 ions with combinations of 0, 1, or 2 phosphates; 0, 1, 2, or 3 primary acyl chains; and 0 or 1 secondary acyl chains, while the range of fatty acid carbons was from 12 to 16. All theoretical neutral losses that can be generated from direct bond cleavage of Yp lipid A were deposited in Yp-TNL databases, including (n:0)-3-OH ketene, (n:1) ketene, (n:0) acid, Ara4N, and up to two phosphate groups; where *n* is the fatty acid chain length. Possible neutral losses from any signature ions were also calculated and included in Yp-TNL databases. Yp-TSI and Yp-TNL databases were used in structural analysis of Yp_WT, Yp_LpxE, and Yp_LpxF datasets.

HiTMS Analysis of Fn Lipid A Mass Spectra

HiTMS analysis begins with the examination of a set of tandem mass spectra from the highest order of MS^n mass spectra available, which by default is the least complex, and works backward toward the precursor ion scan. In the case of the Fn dataset, HiTMS began with MS^4 tandem mass spectra, which were examined for the presence of species-specific, theoretical signature ions deposited in the Fn-TSI database. In our dataset of 284 averaged tandem mass spectra, HiTMS found that 147 (71%) of the 206 MS^4 tandem mass spectra contained signature ions present in the Fn-TSI database. This meant that these 147 MS^4 tandem mass spectra came from lipid A species that were assignable based on the known Fn lipid A biochemistry from which the Fn-TSI database was defined. As confirmation, the assigned tandem mass spectra from the signature ion matches were further evaluated for the presence of expected neutral losses present in the Fn-TNL database. This secondary Fn-TNL database search provided preliminary structural hypotheses for each of the 147 mass spectra that were subsequently evaluated by a cross-correlation (X-score) analysis.

In order to evaluate the significance of a match, HiTMS employed a target-decoy strategy similar to that frequently used in proteomics to evaluate the false discovery rate of peptide tandem mass spectral matches. For each of the 147 assigned MS^4 tandem mass spectra, HiTMS generated six on-the-fly decoys by shuffling the positions and lengths of the fatty acid side chains based on the species-specific candidate structure. Decoys had the exact same chemical composition as the candidate structure and maintained lipid A-like fragmentation rules. An X-score, which is a correlation score used to evaluate the closeness of fit between an acquired and a theoretical MS^n spectrum was then calculated for every LSM (as defined in Section 2) including candidates and decoys.

The X-score analysis calculated the similarity between tandem mass spectra generated from the derived hypothetical structures used to create the TSI and TNL databases and those from the observed data. To generate the X-scores, peak lists of spectra were converted into vectors binned by DeltaMass mass tolerance (default 0.8 Da). This X-score process produced a set of values similar to those generated by SEUQEST, a popular tool for matching peptide tandem mass spectra to the amino acid sequences of proteins in a database [32]. As shown in Figure 4, the X-score distribution from candidate matches was much higher than X-score from decoy matches. Thus, an X-score value of 3.0, at which the two distributions intersected, was selected as the default X-score cutoff that successfully rejected more than 99% of decoys and resulted in an FDR <0.01.

HiTMS analysis of the 284 averaged mass spectra matched 120 lipid A structures (Supplemental Table S.2). Comparison to our database of 133 manually assigned lipid A structures revealed that 109 unique lipid A structures were correctly retrieved by HiTMS and 11 putative new structures were hypothesized. Only 24 of the original structures were undetected, which could be due to a number of reasons; e.g., some of the manually assigned structures may be incorrect or the automated threshold used by HiTMS may have missed some ions selected during the manual analysis. The performance of HiTMS was assessed using F-measure, which is a tool for calculating precision and recall [39]. For the Fn dataset, a balanced F-measure of 0.86 with a precision of 0.91 and a recall of 0.82 was observed, which suggests that HiTMS was able to assign lipid A structures with both a high precision and high recall. Interestingly, out of the 206 MS^4 total tandem mass spectra 147 were uniquely assignable and 57 annotated with more than one lipid A structure. The latter results suggest the existence of isomers even after the sample was examined down to the MS^4 level. The isomeric structures detected in these 57 MS^4 spectra are most likely due to various combinations of the fatty acid side chains [7].

HiTMS Versus Manual Annotations

Of the 133 manually assigned lipid A tandem mass spectra, 24 (18%) were not annotated by HiTMS (Supplemental

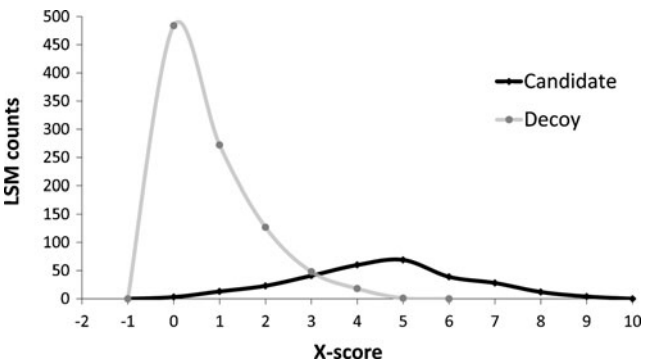

Figure 4. Example of X-score distribution from the Fn lipid A data. For every MS^n spectrum set, HiTMS generated six decoys on-the-fly based on the candidate structures and calculated the X-score of each lipid-spectrum match (LSM). X-score distribution from candidates was much higher than X-score from decoys that had both median and mean around 0

Table S.3). However, while HiTMS produced structural hypotheses for 11 of these 24, their X-scores were below the acceptable cutoff value of 3.0. These low X-score structure predictions could be due to the lack of detectable signature ions or other supporting ions, such as evidence of direct bond cleavage of fatty acids in these 11 tandem mass spectra. It should be noted that as the hierarchical MS^n strategy approaches an MS^4 acquisition that data quality naturally declines along with declining ion signal strength. This lower data quality is one likely reason for failing to detect informative signature ions. Thus, as spectral quality decreases, lower X-score values are expected. In addition, reconstruction of theoretical spectra is based on the most ideal fragmentations that include every possible direct bond cleavage and their combinations, not the most likely to cleave chemical bonds. While lipid A structures in Fn are very complex, it may be possible that in some cases the initial spectra might have been misinterpreted manually. For the moment though this manually curated library of structures is the standard from which we judge success for HiTMS. Finally, while we have not done so here, it should be noted that HiTMS is capable of examining, individually, each of the 49,943 acquired mass spectra that were averaged to produce 248 spectra for manual interpretation. This strategy will likely reveal even more complexity, but given the difficulties in confirming accuracy of so many putative structures, we have limited our analysis here to the manually curated data set.

HiTMS Spectral Annotation

As an example of how HiTMS annotates mass spectra in general and the ability of HiTMS to aid manual structure assignment, Figure 5 shows an annotated tandem mass spectrum and the subsequent lipid A structural hypothesis that, in this case, was not detected by manual interrogation, but can be easily confirmed with the annotated spectrum. Note that one reason this structure may have been missed is due to the low intensity signature ions that HiTMS detected, which could be easily neglected during manual assignment due to the fact that there were many more dominant ions to account for during interpretation. Further, as mentioned above, distinguishing isomers is a difficult task that HiTMS analysis of MS^n data was designed to handle. To do so, HiTMS uses a sophisticated peak detection algorithm, MassSpecWavelet [30], to improve the detection of low intensity peaks while allowing the analyst to adjust the threshold accordingly. In these cases, HiTMS not only identified many potential new structures, but also annotated them, facilitating manual review of the structure assignments.

Yp Lipid A Structural Analysis

Lipid A structural diversity is reflected, in part, in the various combinations of fatty acids (numbers and types), as well as in the phosphorylation patterns, which appear to be species-specific. Like fatty acid composition, phosphorylation pattern has been shown to have strong influence on bacterial pathogenicity [40]. Thus, to insure that HiTMS could make accurate structure assignments where such modifications were common, we also analyzed a previously published Yp lipid A data sets [10, 11]. This data was generated for lipid A isolated from Yp after growth in rich media at 37 °C as part of a study to determine the phosphorylation pattern in Yp. Specifically, Yp lipid A is structurally unique from Fn and has been shown to have diverse phosphorylation patterns, including bisphosphorylation and pyrophosphorylation. Unlike Fn, Yp lipid A contains two phosphates and is usually detected in bisphosphorylated forms [10]. In addition, there are additional Ara4N modifications and differences in fatty acid side chains compared with Fn. Thus, this data set on a symmetrical form of lipid A provided a number of unique opportunities to test HiTMS.

These Yp datasets contained MS^1 to MS^3 tandem mass spectra, but no MS^4 tandem mass spectra. Regardless of this difference and the fact that Yp datasets were much smaller than the Fn data set, HiTMS was able to correctly assign two diphosphorylated lipid A structures from Yp_WT dataset (i.e., the major structures) as well as five lipid A structures from genetically modified Yp, which resulted from in vivo removal of C-1 phosphate (Yp_LpxE dataset) and C-4' phosphate (Yp_LpxF dataset) [10, 11].

Specifically, the Yp data set included spectra from MS^1 to MS^3consisting of 31,521 tandem mass spectra in Yp_WT, 2807 in Yp_LpxE, and 3187 in Yp_LpxF that were averaged (as per Fn) for HiTMS analysis. Analysis of 148 averaged MS^3 spectra from Yp_WT, 35 from Yp_LpxE, and 40 from Yp_LpxF, respectively, showed that HiTMS correctly assigned these lipid A structures as well as the structures known to be made by genetically modified LpxE and LpxF phosphatase mutants. As with the Fn data, an X-score cutoff of 3.0 was applied to filter the search results. An example of HiTMS annotation of Yp_LpxE data is depicted in Figure 6. The detected $^{0,4}A_2$ ion suggests the fatty acid on the 2-position is (14:0)-3-OH, but the fragmentation of unsaturated 14:1 fatty acid could be on 2'- or 3'-position. This process produced two candidate structures, which are presented for expert review to confirm/refute the proposed structures. Based on the X-score and the number of matched ions, the 14:1 acyl chain structure is more likely to be on the 2'-position, however, the possibility of it being on the 3'-position could not be ruled out. In addition, the mass difference from MS^1 to MS^2 was annotated with a fatty acid composition of a 12:0 plus a (14:0)-3-OH ketene or a 14:0 fatty acid plus a (12:0)-3-OH ketene. In this case, a simple mass difference alone is not enough to determine which composition is more likely correct, but we have shown that HiTMS can annotate correctly even the correlated MS^2 spectra provided by A-ions.

Conclusions

The HTP hierarchical MS^n data acquisition strategy we developed previously [7] produced thousands of tandem mass spectra in a few days' time that were reduced to 1-min averages

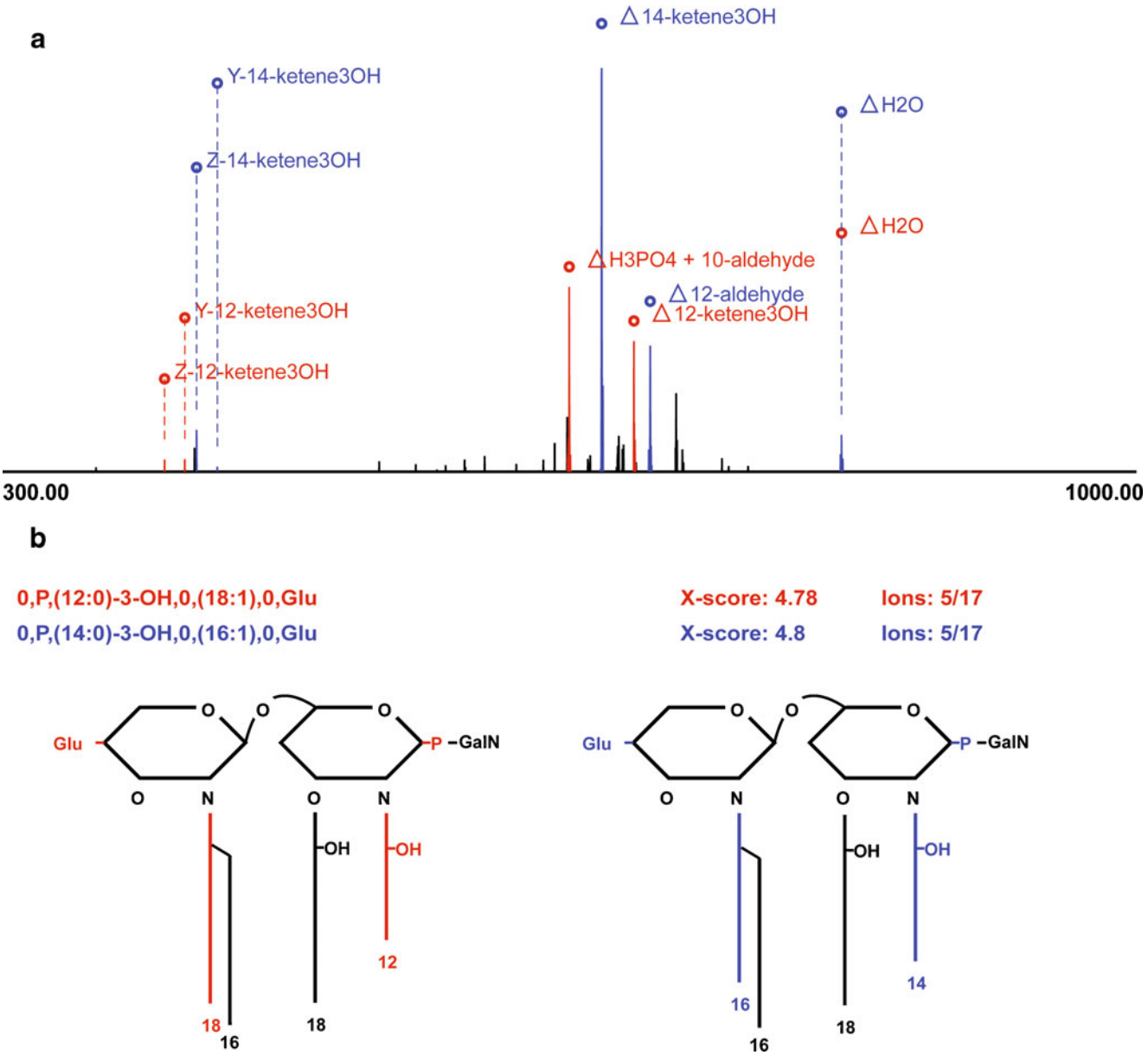

Figure 5. Example of a putative lipid A structure derived from a MS^4 spectrum. **(a)** Two lipid A isomers were found by HiTMS in MS^4 of *m/z* 1025.7 (from MS^3 of 1486.9, from MS^2 of 1743.1), including one known lipid A structure (A3_3, blue) and one potential new structure in red. HiTMS labeled signature ions and neutral losses. For example, label "Y-14-ketene3OH" represented a Y-ion with a (14:0)-3-OH fatty acid attached on 2-position, which is equal to a Y-ion backbone plus a (14:0) ketene. Label "Δ" represented a neutral loss of the following molecule. Detected versus theoretical ion ratios are shown next to the X-score of corresponding assignments. **(b)** Retrieved lipid A structures. Red and blue colors indicated the corresponding difference of lipid A. Black indicated the structural information obtained from the neutral losses during MS^1 to MS^2 and to MS^3

for manual interpretation. Even then, Fn lipid A interpretation required several months of expert analysis time to produce 133 unique lipid A structures. Thirty of these 133 were previously reported, and here we report 103 additional structures (Supplemental Table S.1) derived from manual interpretation of the same original dataset. In order to automate lipid A structure assignment so that annotation rates would be more in line with the data acquisition rates, we developed HiTMS. This automated algorithm relies on species-specific theoretical libraries of signature ions and neutral losses to produce structure assignments. HiTMS correctly identified 85% of the 133 structures in our Fn lipid A library at <0.01 FDR and produced 11 hypothesized structures not revealed by manual analysis. These additional structures are likely due to the sensitivity of an automated routine to assign confidence to low intensity signals that manual analysis overlooked. Finally, HiTMS was also shown to work on the Yp lipid A data acquired up to the MS^3 level. Thus, HiTMS has been shown to be a reliable tool for systematic, automated interpretation of hierarchical MS^n lipid A tandem mass spectral data sets and

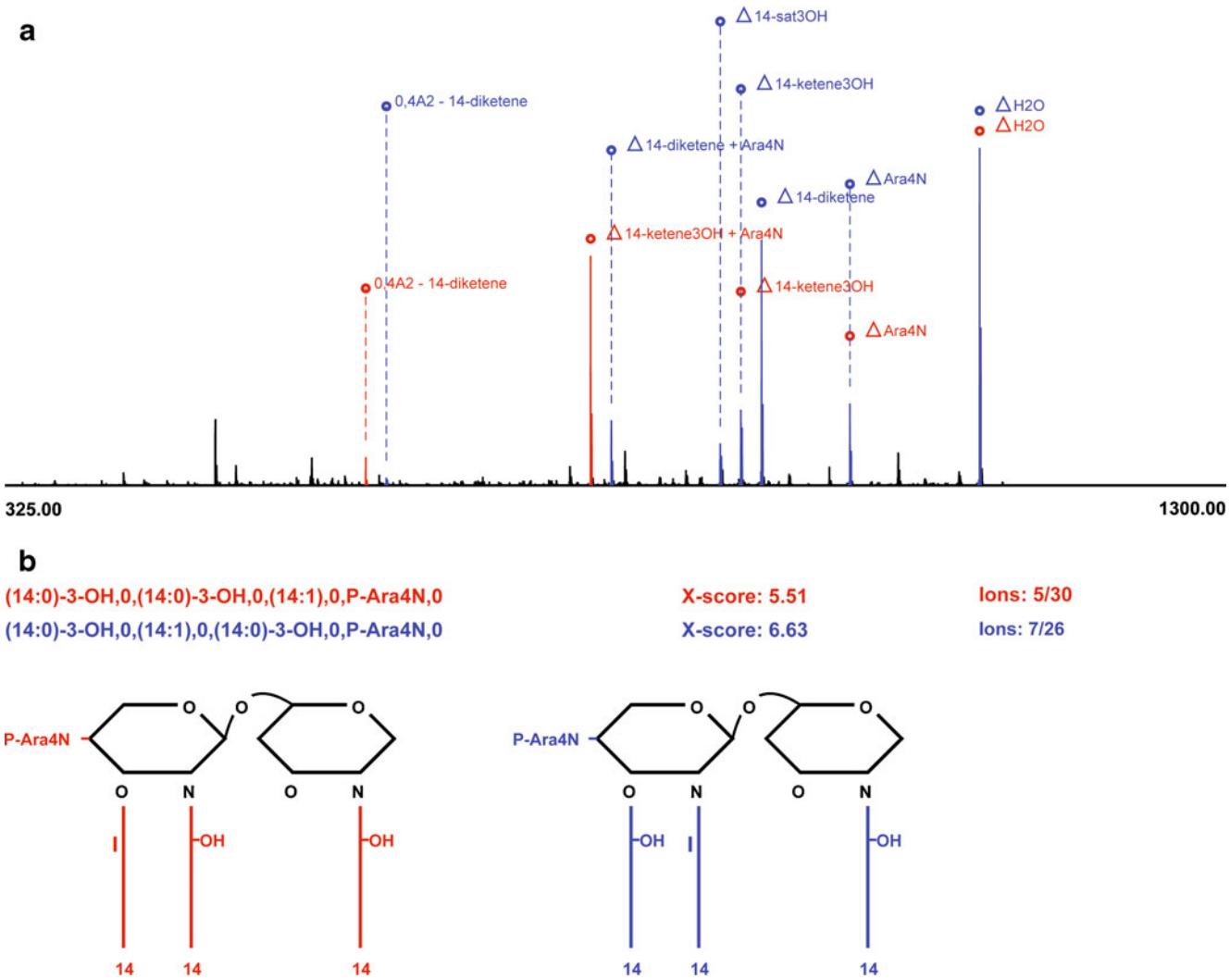

Figure 6. Example of ambiguous lipid A structural annotations derived from a Yp MS^3 spectrum. **(a)** Negative ion mode ESI-LTQ CID MS^3 spectrum of the ion at *m/z* 1192.7 (from MS^2 of 1637.1), from Yp_LpxE lipid A. HiTMS proposed two isobaric structures labeled in red and blue. The mass difference of MS^1 and MS^2 were associated with the combination of a 12:0 fatty acid and a 14:0 ketene or a 12:0 ketene and a 14:0 fatty acid. Label "0,4A2-14-diketene" represented a $^{0,4}A_2$-ion with a (14:1) fatty acid attached on either 2' or 3'-position, which is equal to a $^{0,4}A_2$-ion backbone plus a C14 diketene. Label "Δ" represented a neutral loss of the following molecule. **(b)** Preliminary lipid A structures proposed by HiTMS. HiTMS did not propose the final structures because of the ambiguous annotation of the neutral losses during MS^1 to MS^2

capable of proposing structures not considered by manual analysis. The success of HiTMS to assign the structures of two chemically distinct species of lipid A from Fn and Yp species suggests lipid A from other species could also be interpreted via HiTMS, as well as other classes of more generic glycolipids. This expectation is based on the use of an unbiased, scrambled fatty acid composition permutation strategy that is similar to the amino acid sequence scrambling strategy used to assign FDR values in proteomic experiments. This type of analysis provides the user with fast structure assignments based on objectively assigned X-scores at a user defined FDR or a X-score threshold that is applicable for both symmetric and asymmetric lipid A species. Unlike other lipid identification related software that primarily examine lipid classification and lipidomic profiles, HiTMS is capable of analyzing thousands of tandem mass spectra generated by a HTP hierarchical MS^n data acquisition strategy. Finally, while we have used HiTMS on lipid A from well studied species, it also has the potential to be used to examine lipid A data from novel bacteria by empirically optimizing the libraries in an iterative fashion at a given X-score cutoff.

Acknowledgments

The authors thank the NIAID Regional Center of Excellence for Biodefense and Emerging Infectious Diseases Mass Spectrometry Core (U54 AI057141-03), the National Center for Research Resources (1S10RR023044-01), and NIEHS sponsored Center for Ecogenetics and Environmental Health (P30ES07033) for funding.

References

1. Miller, S.I., Ernst, R.K., Bader, M.W.: LPS, TLR4 and infectious disease diversity. *Nat. Rev. Microbiol.* **3**(1), 36–46 (2005)
2. Raetz, C.R., Whitfield, C.: Lipopolysaccharide endotoxins. *Annu. Rev. Biochem.* **71**, 635–700 (2002)
3. Ernst, R.K., Yi, E.C., Guo, L., Lim, K.B., Burns, J.L., Hackett, M., Miller, S.I.: Specific lipopolysaccharide found in cystic fibrosis airway Pseudomonas aeruginosa. *Science* **286**(5444), 1561–1565 (1999)
4. Ernst, R.K., Hajjar, A.M., Tsai, J.H., Moskowitz, S.M., Wilson, C.B., Miller, S.I.: *Pseudomonas aeruginosa* lipid A diversity and its recognition by Toll-like receptor 4. *J. Endotoxin. Res.* **9**(6), 395–400 (2003)
5. Hornef, M.W., Wick, M.J., Rhen, M., Normark, S.: Bacterial strategies for overcoming host innate and adaptive immune responses. *Nat. Immunol.* **3**(11), 1033–1040 (2002)
6. Coats, S.R., Jones, J.W., Do, C.T., Braham, P.H., Bainbridge, B.W., To, T.T., Goodlett, D.R., Ernst, R.K., Darveau, R.P.: Human Toll-like receptor 4 responses to *P. gingivalis* are regulated by lipid A 1- and 4'-phosphatase activities. *Cell Microbiol.* **11**(11), 1587–1599 (2009)
7. Shaffer, S.A., Harvey, M.D., Goodlett, D.R., Ernst, R.K.: Structural heterogeneity and environmentally regulated remodeling of *Francisella tularensis* subspecies *novicida* lipid A characterized by tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **18**(6), 1080–1092 (2007)
8. Schneiter, R., Brugger, B., Sandhoff, R., Zellnig, G., Leber, A., Lampl, M., Athenstaedt, K., Hrastnik, C., Eder, S., Daum, G., Paltauf, F., Wieland, F.T., Kohlwein, S.D.: Electrospray ionization tandem mass spectrometry (ESI-MS/MS) analysis of the lipid molecular species composition of yeast subcellular membranes reveals acyl chain-based sorting/remodeling of distinct molecular species en route to the plasma membrane. *J. Cell. Biol.* **146**(4), 741–754 (1999)
9. Han, X., Gross, R.W.: Shotgun lipidomics: electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. *Mass Spectrom. Rev.* **24**(3), 367–412 (2005)
10. Jones, J.W., Shaffer, S.A., Ernst, R.K., Goodlett, D.R., Turecek, F.: Determination of pyrophosphorylated forms of lipid A in Gram-negative bacteria using a multivaried mass spectrometric approach. *Proc. Natl. Acad. Sci. U.S. A.* **105**(35), 12742–12747 (2008)
11. Jones, J.W., Cohen, I.E., Turecek, F., Goodlett, D.R., Ernst, R.K.: Comprehensive structure characterization of lipid A extracted from *Yersinia pestis* for determination of its phosphorylation configuration. *J. Am. Soc. Mass Spectrom.* **21**(5), 785–799 (2010)
12. Wenk, M.R.: The emerging field of lipidomics. *Nat. Rev. Drug. Discov.* **4**(7), 594–610 (2005)
13. Schwudke, D., Oegema, J., Burton, L., Entchev, E., Hannich, J.T., Ejsing, C.S., Kurzchalia, T., Shevchenko, A.: Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. *Anal. Chem.* **78**(2), 585–595 (2006)
14. Leavell, M.D., Leary, J.A.: Fatty acid analysis tool (FAAT): An FT-ICR MS lipid analysis algorithm. *Anal. Chem.* **78**(15), 5497–5503 (2006)
15. Ekroos, K., Chernushevich, I.V., Simons, K., Shevchenko, A.: Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer. *Anal. Chem.* **74**(5), 941–949 (2002)
16. Ejsing, C.S., Duchoslav, E., Sampaio, J., Simons, K., Bonner, R., Thiele, C., Ekroos, K., Shevchenko, A.: Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. *Anal. Chem.* **78**(17), 6202–6214 (2006)
17. Song, H., Hsu, F.F., Ladenson, J., Turk, J.: Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. *J. Am. Soc. Mass Spectrom.* **18**(10), 1848–1858 (2007)
18. Hubner, G., Crone, C., Lindner, B.: lipID–a software tool for automated assignment of lipids in mass spectra. *J. Mass Spectrom.* **44**(12), 1676–1683 (2009)
19. Yang, K., Cheng, H., Gross, R.W., Han, X.: Automated lipid identification and quantification by multidimensional mass spectrometry-based shotgun lipidomics. *Anal. Chem.* **81**(11), 4356–4368 (2009)
20. Han, X., Gross, R.W.: Shotgun lipidomics: multidimensional MS analysis of cellular lipidomes. *Expert Rev. Proteom.* **2**(2), 253–264 (2005)
21. Han, X., Yang, J., Cheng, H., Ye, H., Gross, R.W.: Toward fingerprinting cellular lipidomes directly from biological samples by two-dimensional electrospray ionization mass spectrometry. *Anal. Biochem.* **330**(2), 317–331 (2004)
22. Mikhail, I., Yildirim, H.H., Lindahl, E.C., Schweda, E.K.: Structural characterization of lipid A from nontypeable and type f *Haemophilus influenzae*: variability of fatty acid substitution. *Anal. Biochem.* **340**(2), 303–316 (2005)
23. Schilling, B., McLendon, M.K., Phillips, N.J., Apicella, M.A., Gibson, B.W.: Characterization of lipid A acylation patterns in Francisella tularensis, *Francisella novicida*, and *Francisella philomiragia* using multiple-stage mass spectrometry and matrix-assisted laser desorption/ionization on an intermediate vacuum source linear ion trap. *Anal. Chem.* **79**(3), 1034–1042 (2007)
24. Une, T., Brubaker, R.R.: In vivo comparison of avirulent Vwa- and Pgm- or Pstr phenotypes of *yersiniae*. *Infect. Immun.* **43**(3), 895–900 (1984)
25. Wang, X., McGrath, S.C., Cotter, R.J., Raetz, C.R.: Expression cloning and periplasmic orientation of the *Francisella novicida* lipid A 4'-phosphatase LpxF. *J. Biol. Chem.* **281**(14), 9321–9330 (2006)
26. Westphal, O., Jann, K.: Bacterial Lipopolysaccharides: extraction with phenol-water and further applications of the procedure. Methods Carbohydr. Chem. (5), 83–91 (1965)
27. Caroff, M., Tacken, A., Szabo, L.: Detergent-accelerated hydrolysis of bacterial endotoxins and determination of the anomeric configuration of the glycosyl phosphate present in the "isolated lipid A" fragment of the *Bordetella pertussis* endotoxin. *Carbohydr. Res.* **175**(2), 273–282 (1988)
28. Panchaud, A., Scherl, A., Shaffer, S.A., von Haller, P.D., Kulasekara, H.D., Miller, S.I., Goodlett, D.R.: Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* **81**(15), 6481–6488 (2009)
29. Domon, B., Costello, C.E.: A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjug. J.* **5**, 397–409 (1988)
30. Du, P., Kibbe, W.A., Lin, S.M.: Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **22**(17), 2059–2065 (2006)
31. Eng, J.K., McCormack, A.L., Yates, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**(11), 976–989 (1994)
32. Eng, J.K., Fischer, B., Grossmann, J., Maccoss, M.J.: A fast SEQUEST cross correlation algorithm. *J. Proteome Res.* **7**(10), 4598–4602 (2008)
33. Moore, R.E., Young, M.K., Lee, T.D.: Qscore: An algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13**(4), 378–386 (2002)
34. Klammer, A.A., MacCoss, M.J.: Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **5**(3), 695–700 (2006)
35. Park, C.Y., Klammer, A.A., Kall, L., MacCoss, M.J., Noble, W.S.: Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **7**(7), 3022–3027 (2008)
36. Rebeil, R., Ernst, R.K., Gowen, B.B., Miller, S.I., Hinnebusch, B.J.: Variation in lipid A structure in the pathogenic *yersiniae*. *Mol. Microbiol.* **52**(5), 1363–1373 (2004)
37. Kawahara, K., Tsukano, H., Watanabe, H., Lindner, B., Matsuura, M.: Modification of the structure and activity of lipid A in *Yersinia pestis* lipopolysaccharide by growth temperature. *Infect. Immun.* **70**(8), 4092–4098 (2002)
38. Knirel, Y.A., Lindner, B., Vinogradov, E.V., Kocharova, N.A., Senchenkova, S.N., Shaikhutdinova, R.Z., Dentovskaya, S.V., Fursova, N.K., Bakhteeva, I.V., Titareva, G.M., Balakhonov, S.V., Holst, O., Gremyakova, T.A., Pier, G.B., Anisimov, A.P.: Temperature-dependent variations and intraspecies diversity of the structure of the lipopolysaccharide of *Yersinia pestis*. *Biochemistry* **44**(5), 1731–1743 (2005)
39. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. 249–254 (1999)
40. Park, B.S., Song, D.H., Kim, H.M., Choi, B.S., Lee, H., Lee, J.O.: The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. *Nature* **458**(7242), 1191–1195 (2009)