Journal of proteome research

Article

pubs.acs.org/jpr

The C-Score: A Bayesian Framework to Sharply Improve Proteoform Scoring in High-Throughput Top Down Proteomics

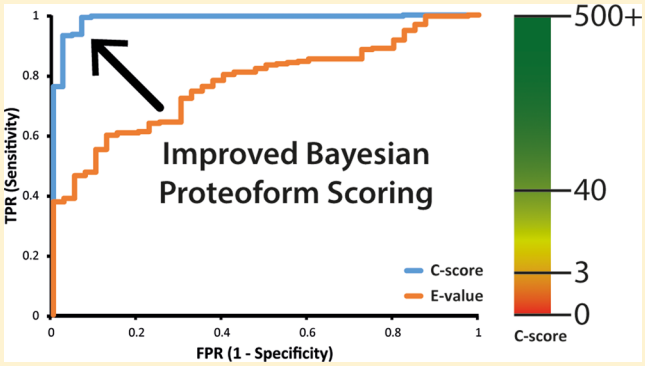
Richard D. LeDuc,*[,†] Ryan T. Fellers,[‡] Bryan P. Early,[‡] Joseph B. Greer,[‡] Paul M. Thomas,[‡] and Neil L. Kelleher*[,‡]

[†]National Center for Genome Analysis Support, Indiana University, 2709 E. 10th Street, Bloomington, Indiana 47408, United States

[‡]Departments of Chemistry and Molecular Biosciences, and the Proteomics Center of Excellence, Northwestern University, 2145 N. Sheridan Road, Evanston, Illinois 60208, United States

Supporting Information

ABSTRACT: The automated processing of data generated by top down proteomics would benefit from improved scoring for protein identification and characterization of highly related protein forms (proteoforms). Here we propose the "C-score" (short for Characterization Score), a Bayesian approach to the proteoform identification and characterization problem, implemented within a framework to allow the infusion of expert knowledge into generative models that take advantage of known properties of proteins and top down analytical systems (e.g., fragmentation propensities, "off-by-1 Da" discontinuous errors, and intelligent weighting for site-specific modifications). The performance of the scoring system based on the initial generative models was compared to the current probability-based scoring system used within both ProSightPC and ProSightPTM on a manually curated set of 295 human proteoforms. The current implementation of the C-score framework generated a marked improvement over the existing scoring system as measured by the area under the curve on the resulting ROC chart (AUC of 0.99 versus 0.78).



KEYWORDS: *top down proteomics, proteoform characterization, Bayesian scoring*

INTRODUCTION

Top down mass spectrometry describes an analytical process for the identification and characterization of whole proteins.[1,2] The canonical "top down" experiment consists of a precursor scan to obtain the intact mass of the proteoform(s)[3] under study and a tandem mass spectrum (MS/MS) obtained using ion fragmentation techniques such as ECD,[4] ETD,[5] HCD,[6] CID,[7] or UVPD.[8] The defining characteristic of a top down experiment is that the precursor ion is an intact proteoform,[3] not the typical small peptides (less than 3 kDa) produced from intentional enzymatic digestion prior to mass spectrometry (MS). Thus, the mass of the precursor ion should represent a native proteoform present in the sample, with its fragment ion masses providing extensive characterization and verification of the primary structure. As larger proteins are targeted, experiments tend toward acquisition logic involving spectral averaging for both the precursor and fragment ions to improve the data quality. These combined data are then analyzed to infer the neutral masses of all intact and fragment ion species observed. For proteins analyzed by electrospray ionization, this Analysis to Infer Mass (AIM) uses either isotopic spacings for direct charge state assignment[9] and/or deconvolution of protein charge states.[10,11]

Historically, top down mass spectrometry has targeted the in-depth characterization of a small number of proteoforms;[12,13] however, the past 5 years has seen a gradual transition to "top down proteomics" where thousands of proteoforms from dozens of LC−MS/MS runs are analyzed.[14,15] While the number of proteins able to be identified has risen into the thousands, the extent of characterization of each individual proteoform varies and currently there is no scoring framework that captures this aspect of top down proteomics. In certain cases, a protein (arising from a specific gene) may be identified confidently with no inference problem whatsoever, yet may be only partially characterized as shown in Figure 1. In Figure 1B, two "equivalent" proteoforms, one with a post-translational modification (PTM) in the first position and the other with the PTM in the second, will each have the same number of matching fragment ions. Scoring systems based only on matching ions will report equal scores for these two proteoforms. This example illustrates how a proteoform can be clearly identified (i.e., the evidence supporting the identification of either of the two positional isomers is strong), yet not fully characterized. This problem of PTM site

Received: December 20, 2013
Published: June 12, 2014

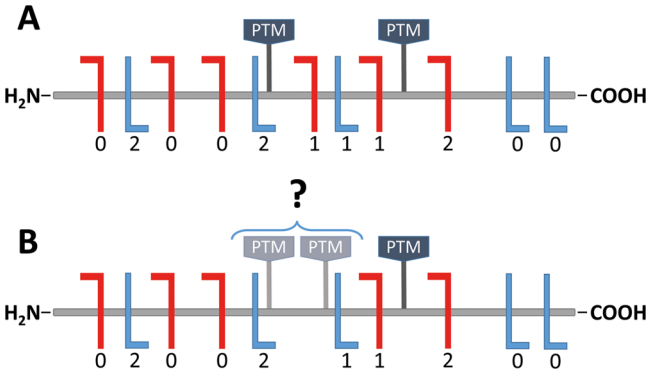


Figure 1. Incomplete fragmentation may or may not lead to partial characterization of a protein molecule. In panel (A), the matched fragment ions uniquely determine the location of the post-translational modifications (PTMs). In panel (B), incomplete fragmentation in the middle of the protein backbone does not yield a definitive PTM localization. In this particular example, the PTM could be located at either of two amino acids, resulting in an identified, but partially characterized proteoform. The C-score framework was developed to handle such cases, routinely encountered in top down proteomics. N-terminal fragment ions are colored red, while C-terminal fragment ions are shown in blue. The numbers below the fragment ions indicate how many PTMs are reported on by each ion.

localization is encountered in bottom up, and has been dealt with in various ways.[16−18] In targeted top down MS generating just a few spectra, manual reanalysis of data, or curation of the primary literature, can be used to select one protein form over another highly related one. In such cases, expert decisions are made to distinguish proteoforms with similar primary structures taking into account cleavage propensity of pairs of amino acids, complementary ion pairs, sequence tags, and mass errors of precursor and fragment ions.

With the shift to high-throughput, fully automated data collection, we now seek a framework for scoring of protein identification and proteoform characterization that builds in domain knowledge to achieve the quality of manual analysis without requiring it. All of the mathematical symbols used below are aggregated into Table 1. Fundamentally, the problem of identifying which proteoform is most consistent with an experiment combines both aspects of protein identification (which gene) and characterization (which proteoform). In practice, the problem is one of testing a series of hypotheses about which proteoform was present in the mass spectrometer, and then picking the "best" hypothesis from the list. Each candidate proteoform (not simply its protein sequence) constitutes a hypothesis, but since candidate proteoforms does not represent all possible hypotheses to be tested, a scoring system must allow for the possibility that the best scoring proteoform is near the correct answer, but not exactly correct. Further, it is frequently the case that the observed data cannot conclusively differentiate between two or more related proteoforms (as in Figure 1B), or the case where multiple proteoforms were actually fragmented together. Typically, the way a proteoform score is used experimentally is not too dissimilar the use of scores in annotating novel nucleotide sequences with BLAST. For a given search, a minimum cutoff threshold is picked by the operator (a priori), and for each query the answer accepted as correct must be both the highest scoring result and greater than the cutoff threshold.

The problem of inferring which proteoform, from the articulated "prior" list of proteoforms in a database, was observed within the mass spectrometer is well-suited to a Bayesian approach (Figure 2). In this case, Bayes law can be rearticulated as follows:

$$\Pr(\text{Proteoform}_i|\text{Data}_{MS/MS}) = \frac{\Pr(\text{Proteoform}_i)\ \Pr(\text{Data}_{MS/MS}|\text{Proteoform}_i)}{\Pr(\text{Data}_{MS/MS})} \quad (1)$$

where (1) $\Pr(\text{Proteoform}_i|\text{Data}_{MS/MS})$ is read as the probability of the ith Proteoform given the MS/MS data, and is known as the *posterior probability* of proteoform i given the observed data; (2) $\Pr(\text{Proteoform}_i)$ is known as the *prior probability* of proteoform i; (3) $\Pr(\text{Data}_{MS/MS}|\text{Proteoform}_i)$ is read as the probability of the data given proteoform i, and is known as the *likelihood* of the data given proteoform i; (4) $\Pr(\text{Data}_{MS/MS})$ is known as the *probability of the data*. By convention, this can be taken as the sum of all prior probabilities multiplied by their

Table 1. Mathematical Symbols Used in this Article

symbol	meaning
$\Pr(\text{Proteoform}_i\|\text{Data}_{MS/MS})$	Probability of the ith proteoform given the MS/MS data, and is known as the *posterior probability* of proteoform i
$\Pr(\text{Proteoform}_i)$	The *prior probability* of proteoform i
$\Pr(\text{Data}_{MS/MS}\|\text{Proteoform}_i)$	The probability of the data given proteoform i, and is known as the *likelihood* of the data given proteoform i
$\Pr(\text{Data}_{MS/MS})$	The *probability of the data*. By convention, this can be taken as the sum of all prior probabilities multiplied by their likelihoods across all hypotheses.
M_O	Observed precursor mass
m_i	Mass of the ith of n observed fragment ions
$\{m_i\}_{i=1}^{n}$	The set of all n observed fragment ions
ϕ_q	The qth of k candidate proteoforms in the database
$\Pr(\phi_q)$	The prior probability of ϕ_q.
$\Pr(M_O,\{m_i\}\|\phi_q)$	The likelihood of proteoform ϕ_q given the observed precursor mass and fragment masses.
k	The number of proteoforms potentially explaining the observed masses.
$\Pi_{i=1}^{n}\Pr(m_i\|\phi_q)$	The product of the individual likelihoods for each observed fragment ion, given the qth candidate proteoform.
M_j	The theoretical intact mass of candidate proteoform j
δ_m	The difference between the observed and a candidate mass. Either at the MS1 or MS2 level.
ω_i	A weight representing the propensities for the ith pair of amino acids to cleave during fragmentation. It is taken as proportional to the product of the cleavage frequencies for both flanking amino acids.
ω_{noise}	The weight assigned to observed MS2 masses that do not match a theoretical MS2 mass for a given candidate proteoform.
t_j	The total area under the MS2 generative model. An intermediate value used to find probabilities.

dx.doi.org/10.1021/pr401277r | J. Proteome Res. 2014, 13, 3231−3240

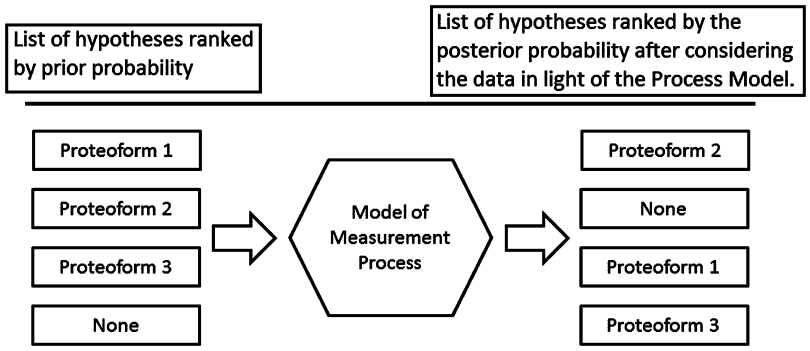


Figure 2. Problem of assigning which proteoform was present in the mass spectrometer during automated data collection can be envisioned as sorting a list of candidate proteoforms based on the observed MS data and a mathematical model of the process by which the observations were collected. Bayes law provides a useful foundation for building these models. In practice, it is not required that the list of candidate proteoforms be explicitly articulated prior to the analysis. It is sufficient that all candidate forms can be calculated for an explicitly stated set. For example, listing base protein sequences and the PTMs to be considered on these sequences defines a list of proteoforms, even if the list is never explicitly written.

likelihoods across all hypotheses.[19] Notice that this term scales the posterior probability to be the fraction of the total of the numerators over all proteoforms interrogated. Thus, a multiple testing correction across the database search is integral in this approach. This differs from controlling for the overall false discovery rate of an experiment which can, for example, be handled posthoc with a search against a scrambled sequence database in a manner analogous to that used in earlier work.[15]

To be precise, we will define the following variables to restate Bayes law from eq 1. Let M_O = observed precursor mass, m_i = mass of the ith of n observed fragment ions, so $\{m_i\}_{i=1}^{n}$ is the set of all n observed fragment ions, and ϕ_q = the qth of k candidate proteoforms in the database. The posterior probability of hypothesis ϕ_q, as per eq 1, can be restated as

$$\Pr(\phi_q|M_O, \{m_i\}) = \frac{\Pr(\phi_q)\ \Pr(M_O, \{m_i\}|\phi_q)}{\Pr(M_O, \{m_i\})}$$

where $\Pr(\phi_q)$ is the "prior" term and $\Pr(M_O,\{m_i\}|\phi_q)$ is the "likelihood" term.

Since the probability of the data is, by convention,[19] known, two values are needed to calculate the posterior probability for each hypothesis: the prior probability and the likelihood for the data. In many applications, the prior probabilities can be taken to be "all hypotheses are equally probable" (i.e., uniform prior probabilities). If there are "k" competing hypotheses, that is, "k" proteoforms in the candidate list, then each proteoform has a prior probability of $1/k$, but this does not need to be the case. It is possible that one would want to assign proteoforms of proteins that contain experimentally demonstrated PTMs (or transcripts informed by RNA-Seq) higher prior probabilities than proteoforms that contain chemically possible, but otherwise rarely observed modifications. For example, if there are two proteoforms that differ only in the location of a PTM such as in Figure 1B, and one of the PTMs was known to occur, while the other had never been reported, and the set of observed fragment ions failed to differentiate the two forms, then the known form should be considered the most probable form to have been observed. This degree of differentiation can be achieved with such a scoring system by first setting and then improving the prior probabilities with continued experimentation. Prior probabilities are, by definition, based on information known prior to data collection and are in practice always somewhat arbitrary in their determination. In sharp contrast, the likelihood of the data given the proteoform is calculated under an explicit mathematical model of the processes used to generate top down MS data; calculation of this likelihood requires generative models.

Generative models "generate" the probability of the observed data, given the proteoform in question. Therefore, they take as input the proteoform, and as much knowledge of the measurement process as can be encoded in the model-creation process. For example, a generative model could include the propensity of individual pairs of amino acids to dissociate during fragmentation (e.g., X-P cleavage in ECD and ETD is not possible,[4] yet DP bonds are preferentially cleaved in threshold dissociation[20]), or the model could include a function for the difference between the observed and theoretical intact mass. The more knowledge of the measurement process the generative model includes, the better, but since the task is to rank order the list of candidate proteoforms (Figure 2), some details can be safely excluded from the generative model if they do not shift the proteoform rankings.

This Bayesian approach offers many advantages. First, it logically follows what many are already doing in the field. For example, during manual data interrogation of an error-tolerant, top down search result, many researchers prefer the observed intact mass to match the theoretical, but allow for the possibility that the masses may not match because the best proteoform in the database is not the correct one. Any scoring system that does not include the closeness of the observed MS1 to the theoretical MS1 is ignoring a valuable observation. Next, this Bayesian framework explicitly states the process in the form of two generative models; one for the precursor mass, and the other for the fragment ion masses. Since the models are clearly articulated, they can be defended, rejected, or modified based on community discourse. In practice, this means that the process model can be modified and updated to reflect new understanding of the measurement process, or to reflect changes to the process used. For example, an automated data analysis system can use the same scoring mechanism, but employ different parameter sets that reflect experimental differences. Experiments that use CID or HCD will employ different parameter sets in the generative models used for scoring than the same top down experiment employing ECD/ETD or UVPD[8] for protein ion fragmentation.

It should be noted that the calculation of scores in this system differs from Bayesian approaches commonly seen in biological sequence analysis.[21] In those applications, the generative models usually have unknown parameters that

dx.doi.org/10.1021/pr401277r | J. Proteome Res. 2014, 13, 3231−3240

need to be estimated from the sequence data, which is usually taken as being perfectly known. The application here is more like that of Edwards where the process is considered known.[22] As will be seen, our generative models have no unknown parameters; instead all values are taken from our knowledge of mass spectrometry, or from prior studies that focused on determining the needed value. Thus, instead of inferring values from the data collected for the study in question, we have a framework for applying knowledge gained in analytic chemistry to the problem of protein inference.

Currently, ProSightPC and ProSight PTM report an expectation, or E-value, for each proteoform. This score is calculated by multiplying the probability of getting at least the observed number of matching fragment ion due to chance by the number of proteoforms interrogated. For this calculation, the probability of getting at least a given number of matching fragment ions is determined using Poisson-based model.[23] Here we describe the implementation of this Bayesian approach to scoring, and compare one set of generative models to the ProSight PTM expectation value used within ProSightPTM and ProSightPC.[23−28] We show that our implementation, with its current generative models, provides an improvement over the existing scoring system as measured by area under the curve (AUC) on the resulting ROC chart[29] (AUC of 0.99 versus 0.78 for a complex human example and AUC 0.85 versus 0.80 in *Pseudomonas*).

METHODS

Initial Assumptions for Probability Calculations

The observed data in a canonical MS/MS experiment includes (a) the neutral precursor mass, which gives the total molecular weight of the proteoform under study, and (b) a set of neutral fragment ion masses, that is, masses of the products resulting from fragmentation of the proteoform. Also required is a database of possible proteoforms, each of which serves as a "hypothesis" that could potentially explain the observed MS/MS data. The database of possible proteoforms was generated by combinatorial expansion of all potential proteoforms using the known modification information in the UniProt Knowledgebase.[25−28]

Derivation of the Likelihood

Since the observed data M_O and each fragment ion, m_i, are independently conditioned on φ_q, we could take

$$\Pr(M_O, \{m_i\}|\varphi_q) = \Pr(M_O|\varphi_q) \prod_{i=1}^{n} \Pr(m_i|\varphi_q)$$

Unfortunately, this approach suffers from two limitations. First the magnitude of $\Pi_{i=1}^{n} \Pr(m_i|\phi_q)$ scales with n. Larger lists of MS2 fragment ions lower the calculated value of the MS2 likelihood, relative to lists with few fragment ions, which makes it impossible to directly compare scores between separate queries with differing numbers of MS2 fragment ions. This can be mitigated by weighting by the geometric mean of the number of MS2 fragment ions; $1/n$. Second, this approach weighs the MS1 data as simply a single additional matching fragment ion. To avoid this, we introduce two scaling functions, f and g, to map the ranges of the individual generative model to a normalized range;

$$\Pr(M_O, \{m_i\}|\varphi_q) = f(\Pr(M_O|\varphi_q))\, g((\prod_{i=1}^{n} \Pr(m_i|\varphi_q))^{1/n})$$

For the precursor generative model, the function f is simply the identity function. For the fragment ion generative model, we define g by its logarithm base, which is simply a linear function on the logarithm base 10 of the fragment probability; g sets the logarithm base 10 of the minimum possible fragment probability to the logarithm base 10 of the minimum possible precursor probability and likewise for the maxima. If $\min_1$ is the minimum precursor probability, $\max_1$ is the maximum precursor probability, $\min_2$ is the minimum fragment probability, and $\max_2$ is the maximum fragment probability, then

$$\log_{10} g((\prod_{i=1}^{n} \Pr(m_i|\phi_j))^{1/n}) = \frac{(\log_{10} \max_1 - \log_{10} \min_1)}{(\log_{10} \max_2 - \log_{10} \min_2)} \times (\log_{10}(\prod_{i=1}^{n} \Pr(m_i|\phi_j))^{1/n} - \log_{10} \min_2) + \log_{10} \min_1$$

Derivation of the Posterior Probability

Under an assumption of uniform prior, and given the likelihood functions from above, we have

$$\Pr(\varphi_q|M_O, \{m_i\}) = \frac{\Pr(\varphi_q)\, \Pr(M_O, \{m_i\}|\varphi_q)}{\sum_j \Pr(\varphi_j)\, \Pr(M_O, \{m_i\}|\varphi_j)}$$

$$\Pr(\varphi_q|M_O, \{m_i\}) = \frac{f(\Pr(M_O|\varphi_q))\, g((\prod_i \Pr(m_i|\varphi_q))^{1/n})}{\sum_j f(\Pr(M_O|\varphi_j))\, g(\prod_i \Pr(m_i|\varphi_j))} \quad (2)$$

Equation 2 therefore reduces the posterior probability of a hypothesis to the data likelihood computation, with a normalization factor equal to the sum of the likelihoods under all possible hypotheses. All that is needed now to calculate posterior probabilities are the generative models. Since we have assumed that the given database of proteoforms is an exhaustive set of hypotheses, these generative models must allow for the possibility of observing related proteoforms that are not present in the database.

Generative Models

The C-score system requires two generative models; one for the precursor mass M_O and the other for the set of observed fragment ion masses, $\{m_i\}$. These models fully prescribe how to compute the likelihood terms on the right-hand side of eq 2. Note that the particular form of eq 2 implies that all the likelihood (probability) terms need only be computed up to a constant factor. This constant factor cancels between the numerator and denominator of the RHS of eq 2.

The probability $\Pr(M_O|\phi_j)$ of an arbitrary protein sequence ϕ_j (of theoretical mass M_j) producing the observed precursor mass M_O can be modeled as a function of δ_m, the difference between the M_O and M_j. We create the probability distribution $\Pr(M_O|\phi_j)$ such that masses within δ_m of M_j have the highest probability, and this probability reduces as a truncated Gaussian function with $\mu = 1$, $\sigma = 30$ Da, and a minimum value of 1×10^{-300}. Notice that we only need to specify $\Pr(M_O|\phi_j)$ to a constant factor (i.e., we only need to specify a non-negative function $f(M_O, \phi_j)$ proportional to the probability), and the normalization factor $(1/\int_0^{\max(M_O)} f(M_O,\phi_j)\mathrm{d}M_O)$ that converts it to a probability density function is assumed implicitly.

dx.doi.org/10.1021/pr401277r | J. Proteome Res. 2014, 13, 3231−3240

To specify the probability distribution $\Pr(m_i|\phi_j)$, we need an MS2 generative model based on our knowledge of the measurement process during tandem mass spectrometry. We note that each fragmentation event involves cleavage of the intact proteoform at *one* bond on the protein backbone, resulting in exactly *two* fragments (although both may not be observed in the spectrometer). The fragmentation propensity depends on the pair of adjacent amino acids flanking the cleavage site. The generative model we choose is based on this observation, and uses the following basic ideas:

(1) Each theoretically possible fragment ion mass defines a region of width $2\delta_m$ $(m - \delta_m, m + \delta_m)$ called a *permissible region*. An observed mass m_i within a permissible region has a probability proportional to the cleavage propensity of gas phase protein ions at that permissible region.

(2) An observed fragment ion mass m_i outside of any permissible region has a small, constant probability.

These ideas are captured in the probability distribution function of Figure 3. In principal, for any given ϕ_j, the probability distribution $\Pr(m_i|\phi_j)$ is constructed by assigning a height to every point in the range $(0, M_j)$, where M_j is the theoretical precursor mass. In practice, MS2 mass lists can contain unexpected ions with mass values greater than M_j, and so an arbitrarily large mass of 4 million Daltons is taken in place of M_j. (The value of 4 million Daltons was selected as it is safely above the mass of the largest protein known, titin at 3.9 MDa, and allows for a modified form of titin to be present. In practice, this value is effectively infinity.) The assigned heights across the entire range are divided by the total area under the curve, thus defining a probability density function. The heights are assigned as follows:

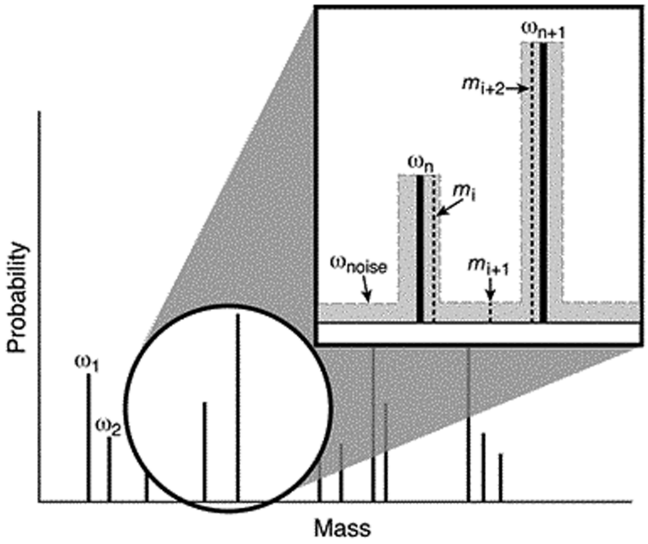


Figure 3. MS2 generative model for fragment ions observed in tandem mass spectrometry (MS/MS). For instruments capable of MS/MS with accurate mass, thin (e.g., low part-per-million wide) permissible regions occur with a search-defined width around each theoretical fragment ion mass. The heights of these regions are scaled by the propensity of the cleavage events required to form the theoretical ion. Different fragmentation methods require different weights, for example ECD has different fragmentation propensities than CID. A very low probability, *the weight for chemical or electronic noise*, is assigned to any observed mass that does not match one of the theoretical masses.

Step 1: Determine all theoretical fragment ions that ϕ_j can give rise to, noting their mass and the N- and C-terminal flanking amino acids at each cleavage site.

Step 2: For each theoretical fragment ion, calculate a weight, ω, proportional to the product of the cleavage frequencies for both flanking amino acids as previously determined for the appropriate fragmentation method.[30] Thus, if α and β are the two flanking amino acids, and f_α and f_β are their respective cleavage frequencies, we set $\omega = f_\alpha f_\beta$. The permissible region associated with a theoretical fragment ion mass, m, is assigned a consistent height equal to the weight ω computed for the corresponding theoretical fragment ion.

Step 3: Any point in the range $(0, M_j)$ but outside of all permissible regions is assigned a height of ω_{noise}, which is a constant. This term represents spectral noise.

We noted above that the probability density function is obtained by dividing the above-mentioned "height" function by the sum of the area under the curve, which is different for different ϕ_j. This total area, denoted by t_j, is computed as follows: the kth permissible region has a height equal to ω_k and a width equal to $2\delta_m$. Regions outside of permissible regions have a height of ω_{noise} and a total width of $P - 2(\text{len} - 1)\delta_m$, where len is the length of the protein sequence. Thus, $t_j = \omega_{noise}(P - 2(\text{len} - 1)\delta_m) + \Sigma_k 2\delta_m\omega_k$.

The resulting probability density function is

$$\Pr(m_i|\varphi_j) = \begin{cases} \dfrac{\omega_k}{t_j} & \text{for } m_i \text{ in a permissible region with weight } \omega_k, \\ \dfrac{\omega_{noise}}{t_j} & \text{for } m_i < P, \text{ but not in a permissible region,} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The expressions on the right-hand side specify a probability *density* while the expression on the left-hand side is a probability *mass*, which, strictly speaking, should be equated to the probability density over a very small mass interval, δ. However, such a correction can be safely ignored here because the probability mass $\Pr(m_i|\phi_q)$ is used only in the context of eq 2, with equal powers of δ in numerator and denominator.

Using eqs 2 and 3, it is now possible to calculate a posterior probability $\Pr(\phi_q|M_O,\{m_i\})$ for every sequence ϕ_q in the database. This posterior probability is proportional to the likelihood $\Pr(M_O,\{m_i\}|\phi_q)$ since we assume uniform priors (eq 2). Therefore, our search for the maximum a posteriori hypothesis ϕ_q is equivalent to a maximum likelihood estimation (MLE) search, that is, we report the ϕ_q that maximizes $\Pr(M_O, \{m_i\}|\phi_q)$.

The Characterization Score, or C-Score

We report the best hypothesis ϕ_q along with its "*Phred-like*" *characterization score*,[31] which can be written as $C = -10\log_{10}(1 - \Pr(\phi_q|M_O,\{m_i\}))$. This final C-score transformation scales the posterior probability of ϕ_q to the familiar range used in many other bioinformatic applications. C-scores span the standard Phred-like score range of 0 to >500. Practical ranges of the C-score are evaluated with specific examples and reported in the main text below. Therefore, a C-score of 40 is sufficient to judge a proteoform as extensively or fully characterized, while proteoforms with C-scores between 3 and 40 are identified, but only partially characterized. A C-score below 3

dx.doi.org/10.1021/pr401277r | J. Proteome Res. 2014, 13, 3231−3240

indicates insufficient evidence for either identification or characterization. Note also that since the C-score represents a nonlinear transformation of the posterior probability, which is itself normalized by $Pr(Data_{MS/MS})$, there is a functional relationship between the highest score in a search, and the second highest score (Supporting Information Figure 1).

Approach to Comparing Scores

The typical usage of a score such as the C-score is in high throughput data processing. An operator picks a threshold level of the score and then asserts that any query scoring above the threshold identifies the target data. This is done routinely in annotating DNA sequences with tools such as BLAST. To test the utility of a new scoring model, a set of data where the correct answer can be considered known is needed; this forms the "ground truth" of the analysis. When the new scoring system gives the known true answer both as the highest score, and the scoring above the operator-defined threshold, the system is said to have delivered a "true positive". Likewise, when the best proteoform scoring above threshold is not the known correct answer, it is scored as a false positive, and so on for both true and false negatives. *Sensitivity* is the proportion of positives which are actually classified as such for a given threshold. Likewise, *specificity* is the proportion of negatives that are correctly identified as such.

For any given arbitrary threshold of a score, a specificity and sensitivity can be calculated from a known data set. By iterating over a number of thresholds, and plotting the resulting sensitivity against its corresponding specificity (or by convention, against 1-specificity) a "receiver operating characteristic" curve is generated. This curve is known to be indicative of the utility of the classifier. By repeating this iterative analysis of with multiple scoring systems, the relative utility of the systems can be directly compared. So, to compare the new C-score with the existing ProSight E-value, test input data sets with known correct answers are needed. These data sets are searched against the appropriate proteoform database and, for each proteoform returned, the E-value and C-score is calculated. Next, an arbitrary list of threshold values is generated for each score, and for each threshold value in this list, a list of identified proteoforms is generated. The identified proteoform is taken as the form that is both above the threshold score, and to be the highest scoring proteoform. It is therefore possible for a search not to return a proteoform, if no proteoform scores above the cutoff threshold. When searching a test data set with known correct answers, it is possible to classify the search results as either a true positive, true negative, false positive, or false negative. From these classifications, specificity and sensitivity is calculated for each threshold level. These specificity and sensitivity couplets then become points on the ROC chart curve for a given score.

Test Data Sets

Data Collection. The human data files used in this analysis were acquired in the course other studies published previously.[32,33] Briefly, mitochondrial membrane proteins were isolated from HeLa S3 cells and separated using a GELFrEE 8100 Fractionation system (Expedeon) as described previously.[32,33] For the analysis at hand, 295 top down experiments from a nanoLC-Velos Orbitrap Elite MS analysis of a GELFrEE fraction containing ~15−20 kDa proteins were selected for intensive manual interrogation to provide a set of highly curated "known positives" to test parameter sets within the generative models used in development of the C-score framework. *Pseudomonas aeruginosa* were grown on rich media to mid-log phase, were isolated by centrifugation, lysed, separated with GELFrEE, and analyzed with mass spectrometry using the methods referenced above. For the secondary data set, 136 top down experiments were curated to select true-positives for analysis.

Manual Data Analysis. A descriptive analysis of intact and tandem mass spectral data was performed using RawMeat 2.1 (http://vastsci.com/rawmeat/, VAST Scientific). Using these data, related MS2 scans were merged and individual ProSight Upload Format (PUF) files[28] were created for each data set. A combination of QualBrowser and ProSightPC 3.0 (both ThermoFisher, San Jose, CA) was used, and all potential precursors predicted to have been in the MS1 Isolation Window were added as potential proteoforms for the experiments. Both MS1 and MS2 data were deisotoped using the Xtract algorithm embedded within ProSightPC 3.0, generating 623 single experiment PUF files. Each experiment was manually searched against "homo_sapiens_2012_02_top_-down_simple" (containing 460 000 forms) and "homo_sapiens_2012_02_top_down_complex" (containing just over 10 million forms) which are ProSight Warehouses (.pwf files) for human proteins built against UniProt release 2012_02, available for download (ftp://prosightftp:gsX1gON@prosightpc.northwestern.edu/2012_02/Eukaryotes/Homo%20sapiens/). Experiments with single "correct" proteoforms were selected and verified by at least two group members trained in top down proteomics data analysis, to be considered as true answers, and while subjective each met the following two criteria: the most abundant fragment ions must be accounted for, and the intact mass difference between theoretical and experimental must be small (<10 ppm) or must be explainable (±1 Da errors from deisotoping, a previously unknown or unannotated post-translational modification, oxidation, etc.). This will typically involve the assignment of >50% of fragment ions appearing at a signal-to-noise of 10:1 or higher. However, we note that both experimental conditions (e.g., overfragmentation, generating internal ions) and the selection of data processing parameters will affect the quality of data and thus the fraction of ions matched for each experiment.

If evidence was not sufficient to uniquely identify one proteoform for an experiment (arising either from poor data quality or from multiple proteins being fragmented simultaneously), the experiment was excluded from further study here. Using these heuristics, 295 PUF files where only one proteoform was present were selected for use here. These files are described in Supporting Information Table 1 and provided in Supporting Information Data Set 1.

Implementation

To facilitate a comparison of C-scores and E-values, a custom C# .NET 4.5 console application was developed. This application takes, as input, a collection of PUF files, the list of manually validated correct answers, and the C-score version number (with a specific parameter set) to use. The C-score version allows various iterations of generative models to be tested against the E-value and each other. The output from the application is an Excel spreadsheet containing a correct score and actual score for both the E-value and the C-score. During execution, this application runs ProSightPC[27] Absolute Mass and BioMarker searches on each PUF file to compile actual search scores. The correct scores are then calculated using the

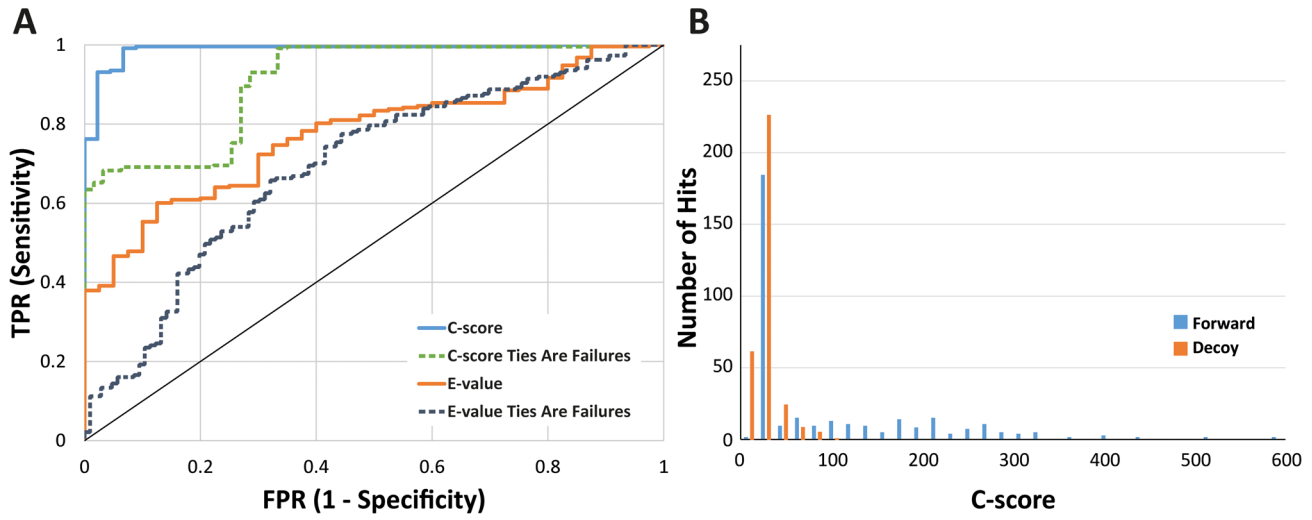


Figure 4. (A) Receiver operating characteristic curves comparing C-scores and E-values on the 295 experiments in the manually curated test data set. The area under the curve for the Blue C-score is 0.99 compared to 0.78 for the Orange E-value. Notice the large difference in sensitivity of the E-value in the low FPR region, between when ties for the best score are considered as acceptable for identification or not. Although present in C-scores, the problem is much less pronounced, and at a much higher sensitivity. (B) Histogram of the 295 C-scores obtained from searching the human database (forward) compared to the C-scores obtained from searching the same experimental data against a scrambled decoy database. Of the decoy hits, only 7% had a C-score above 40, likewise 42% of the forward C-scores are above this value.

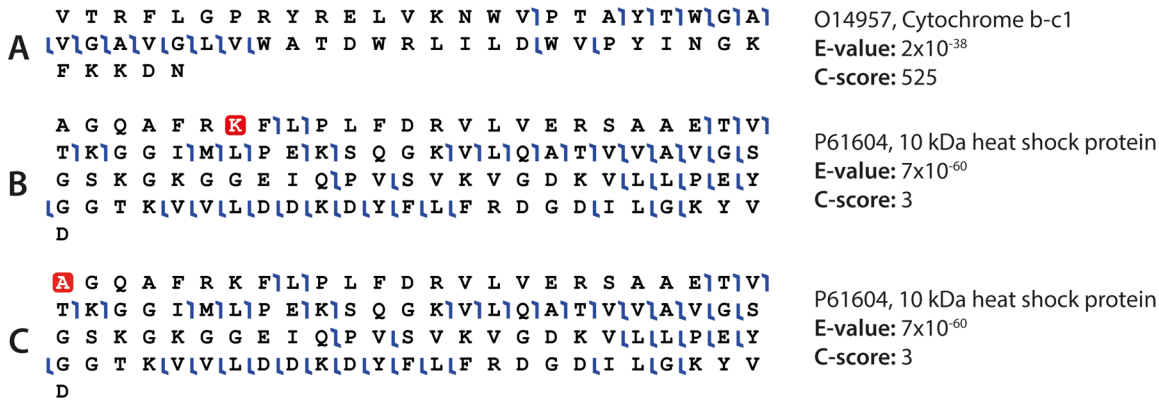


Figure 5. (A) A result with a very high C-score of 525, indicating a fully characterized proteoform of protein cytochrome *b*-*c*1, O14957. (B, C) Two proteoforms with equivalent E-values and C-scores for 10 kDa heat shock protein, P61604. These data show the complementarity between the E-value (scores protein identification well) and the C-score (reflects confidence in characterization of related proteoforms on a Phred-like score from 0 to >500).

list of correct answers (Supporting Information Table 1). The C-score will be available for testing within the Proteoform Characterization Tool,[34] hosted by the Consortium for Top Down Proteomics (http://www.topdownproteomics.org).

Comparison

To compare the E-value to the C-score, we chose as our first input a set of 295 PUF files (described above) and a C-score version of 1.0. The list of correct proteoforms for each of the 295 targets was provided as a CSV file. The custom console application was then used to generate an Excel spreadsheet that was further analyzed. Two ROC curves (one for each score) were plotted on the same chart by checking, for each target, whether the top scoring proteoform was the correct proteoform (Figure 4). Subsequently, additional ROC curves were generated, where a target was counted as incorrect if the correct proteoform merely tied for the best score, but other proteoforms also shared the same score value. We also calculated the area under each curve as a quantitative measure of difference. Two additional ROC curves where generated as before but the correct answer would only be considered a true positive if it uniquely had the best score. Thus, for these curves, if the correct proteoform failed to out-score all other proteoforms in the database it was not considered a true positive. This process was repeated for 136 proteoforms from *Pseudomonas aeruginosa* (data are provided in Supporting Information Data Set 2).

RESULTS AND DISCUSSION

The C-score was substantially better than the E-value at identifying and characterizing the correct proteoform from the data set of 295 human test cases, as well as the 136 *Pseudomonas* test cases. Figure 4 shows a receiver operating characteristic (ROC) curve for both scores on the human data, and the area under the curve for the C-score was 0.99 compared to only 0.78 for the E-value. For the *Pseudomonas* test cases, the C-score also outperformed the E-value with AUCs of 0.85 to 0.80, respectively.

dx.doi.org/10.1021/pr401277r | *J. Proteome Res.* 2014, 13, 3231−3240

Clearly, the C-score (with the v1.0 parameter set) dominates the E-value for all levels of specificity and sensitivity. Since the E-value is simply a nonlinear transformation of the number of matching fragment ions, it seems reasonable to assert that this improvement comes from the new score's ability to include additional factors such as known fragmentation propensities and MS1 mass differences both of which are known to be relevant in characterizing proteoforms. These and other such factors are not considered by the current E-value, nor is the score easily extended to consider such factors.

When the data are sufficient to completely characterize a proteoform, the C-score is often well above 40, indicating hyperconfidence in the characterization power of the underlying data (Figure 5A). However, a major limitation of the E-value occurs when many proteoforms share the same (seemingly confident) score. This can be seen in Figure 4 as the vertical distance between the two E-value lines in the specificity range of 0.8−1.0 (FPR 0.0−0.2). Figure 5B and C shows two example cases where there is equal evidence for two distinct proteoforms, one with an N-terminal acetylation and the other with the acetylation localized to K7. In this example, these two proteoforms share a confident E-value of 7×10^{-60}; however, their C-scores are tied, yet at a much less confident value of 3. Using the data from these two manually annotated sample sets, and with the understanding of the C-score model, we assert three operating ranges for the C-score: C-score > 40 → proteoform is both identified and fully characterized; 3 ≤ C-score ≤ 40 → proteoform is identified, but only partially characterized; C-score < 3 → proteoform is neither identified nor characterized.

To achieve separation between the two proteoforms in Figure 5B and C, future iterations of the C-score will allow for differential probabilities for N-terminal acetylation and internal acetylation. It has been reported that over 80% of all human proteins are constitutively N-terminally acetylated,[35] so one may posit that, in the absence of evidence to the contrary, N-terminal acetylation is more likely than internal acetylation. The C-score model can incorporate this bias and many similar issues that provide biochemical and biological context. To achieve even more specificity and sensitivity in the C-score framework, one may use a proteomic database, such as GPMdb, to use peptide-level information linked to the specific gene product to inform these differential probabilities in top down proteomics experiments.[36] In the case of UniProt entry P61604, GPMdb reports that N-terminal acetylation was observed in 69/69 studies, making the proteoform reported in Figure 5B a much more likely than the proteoform reported in Figure 5C.

The characterization of high-throughput LC−MS/MS experiments, where hundreds of MS/MS targets are automatically generated for each LC run, forms the foundation of high throughput top down proteomics. Our probabilistic formulation leverages the knowledge of fragmentation propensities and may be further extended to incorporate other types of prior information to improve the scoring (vide supra). To characterize unknown PTMs (e.g., a previously unknown methylation), we can extend our C-score model to give a higher probability to common PTM mass shifts (i.e., 42.0105 Da for acetylation, 79.966 Da for phosphorylation, etc.), thereby boosting the data likelihood. This will be particularly useful in complex eukaryotic proteomes where the inclusion of all theoretically possible proteoforms in the database (even at search run time) is computationally prohibitive.

This framework provides a great deal of flexibility for having an appropriate scoring systems for a given experimental procedure. Our current assumption of uniform priors on candidate proteins makes our procedure a maximum likelihood (ML) inference, but the framework can be readily extended to allow for nonuniform priors without additional overhead. Nonuniform priors may be useful when researchers have a priori belief that some protein forms (e.g., particular combinations of PTMs or splice variant) are more likely to be present in the sample (e.g., from RNA-seq data). Further, the score remains robust when used with instruments with decreasing mass accuracy. Supporting Information Figure 2 shows the ROC chart resulting from rerunning the analysis on the human test data set with increasing mass tolerances. As tolerance increases, the discriminative power of the score decreases as more fragment ions will match by chance.

■ CONCLUSION

In sum, we have shown a promising scoring system for protein identification and proteoform scoring to better capture the information content in high-resolution top down proteomics. The C-score model lays forth a path to increase the sophistication of protein identification and characterization platforms to extract maximum value from MS-based proteomics in automated fashion. The conceptual and demonstrable advances outlined above provide a deterministic process to advance the utility of proteomics for nonexperts. We strive to enable both proteomics experts as well as "non-bioinformaticists/non-proteomicists" to advance protein-based science based on proper description of the fully articulated primary structure for whole proteoforms. When coupled with high value experimentation, we posit that the faithful and more efficient conversion of data streams into knowledge will significantly advance major breakthroughs in mechanistic biology and on the front lines of disease research including improved discovery and validation of protein-based biomarkers.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary Table S1, Supplementary Figures 1 and 2, zipped XML files containing precursor and fragment ions for all 295 human top down experiments used in this study (Supplementary Data Set 1), and zipped XML files containing precursor and fragment ions for all 136 *Pseudomonas aeruginosa* top down experiments used in this study (Supplementary Data Set 2). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*(N.L.K.) Tel: 847-467-4362. Fax: 847-467-3276. E-mail: n-kelleher@northwestern.edu.
*(R.D.L.) Tel: 812-856-0752. E-mail: rleduc@iu.edu.

Notes

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, the National Center for Genome Analysis Support, Northwestern University, or Indiana University.

The authors declare the following competing financial interest(s): Several of the authors are also developers on ProSightPC, a piece of commercial software mentioned in the manuscript.

■ ACKNOWLEDGMENTS

The authors thank Adam Catherman and Ioanna Ntai for collecting the data used in this paper, and Saurabh Sinha for many helpful conversations with the early development of Bayesian approach. This research is based upon work supported by the National Science Foundation under Grant No. ABI-1062432 to Indiana University (R.D.L.). The project described was supported by Award No. DA018310 from the National Institute on Drug Abuse (N.L.K.), and Award No. GM067193 from the National Institute of General Medical Sciences (N.L.K.).

■ ABBREVIATIONS

CID, collisionally induced dissociation; HCD, higher energy collisional dissociation; ECD, electron capture dissociation; ETD, electron transfer dissociation; AUC, area under the curve; ROC, receiver-operating characteristic; RHS, right-hand side (of the equation); Pr, probability; PUF, ProSight Upload Format; CSV, comma separated values; FPR, false positive rate; TPR, true positive rate; ML, maximum likelihood; MS, mass spectrometry; MS1, intact/precursor scan; MS2 (or MS/MS), tandem mass spectrometry scan, fragmentation; PTM, post-translational modification; UVPD, ultraviolet photodissociation

■ REFERENCES

(1) Kelleher, N. L. Top-Down Proteomics. *Anal. Chem.* **2004**, *76* (11), 197A−203A.
(2) Reid, G. E.; McLuckey, S. A. 'Top Down' Protein Characterization Via Tandem Mass Spectrometry. *J. Mass Spectrom.* **2002**, *37* (7), 663−675.
(3) Smith, L. M.; Kelleher, N. L. Consortium for Top Down Proteomics, Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10* (3), 186−187.
(4) Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W. Electron Capture Dissociation for Structural Characterization of Multiply Charged Protein Cations. *Anal. Chem.* **2000**, *72* (3), 563−573.
(5) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (26), 9528−9533.
(6) Wu, J.; Warren, P.; Shakey, Q.; Sousa, E.; Hill, A.; Ryan, T. E.; He, T. Integrating Titania Enrichment, Itraq Labeling, and Orbitrap CID-HCD for Global Identification and Quantitative Analysis of Phosphopeptides. *Proteomics* **2010**, *10* (11), 2224−2234.
(7) Kelleher, N. L.; Senko, M. W.; Little, D. P.; O'Connor, P. B.; McLafferty, F. W. Complete Large-Molecule High-Resolution Mass Spectra from 50-Femtomole Microvolume Injection. *J. Am. Soc. Mass Spectrom.* **1995**, *6* (3), 220−221.
(8) Shaw, J. B.; Li, W.; Holden, D. D.; Zhang, Y.; Griep-Raming, J.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S. Complete Protein Characterization Using Top-Down Mass Spectrometry and Ultraviolet Photodissociation. *J. Am. Chem. Soc.* **2013**, *135* (34), 12646−12651.
(9) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (4), 320−332.
(10) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass-Spectrometry of Large Biomolecules. *Science* **1989**, *246* (4926), 64−71.
(11) Mann, M.; Meng, C. K.; Fenn, J. B. Interpreting Mass-Spectra of Multiply Charged Ions. *Anal. Chem.* **1989**, *61* (15), 1702−1708.
(12) Whitelegge, J. P.; Zabrouskov, V.; Halgand, F.; Souda, P.; Bassiliana, S.; Yan, W.; Wolinsky, L.; Loo, J. A.; Wong, D. T. W.; Faull, K. F. Protein-Sequence Polymorphisms and Post-Translational Modifications in Proteins from Human Saliva Using Top-Down Fourier-Transform Ion Cyclotron Resonance Mass Spectrometry. *Int. J. Mass Spectrom.* **2007**, *268* (2−3), 190−197.
(13) Pan, J. X.; Han, J.; Borchers, C. H.; Konermann, L. Hydrogen/Deuterium Exchange Mass Spectrometry with Top-Down Electron Capture Dissociation for Characterizing Structural Transitions of a 17 kDa Protein. *J. Am. Chem. Soc.* **2009**, *131* (35), 12801−12808.
(14) Ansong, C.; Wu, S.; Meng, D.; Liu, X. W.; Brewer, H. M.; Kaiser, B. L. D.; Nakayasu, E. S.; Cort, J. R.; Pevzner, P.; Smith, R. D.; Heffron, F.; Adkins, J. N.; Pasa-Tolic, L. Top-Down Proteomics Reveals a Unique Protein S-Thiolation Switch in *Salmonella typhimurium* in Response to Infection-Like Conditions. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (25), 10153−10158.
(15) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. Mapping Intact Protein Isoforms in Discovery Mode Using Top-Down Proteomics. *Nature* **2011**, *480* (7376), 254−258.
(16) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285−1292.
(17) Taus, T.; Kocher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and Confident Phosphorylation Site Localization Using PhosphoRS. *J. Proteome Res.* **2011**, *10* (12), 5354−5362.
(18) Bailey, C. M.; Sweet, S. M. M.; Cunningham, D. L.; Zeller, M.; Heath, J. K.; Cooper, H. J. SloMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra. *J. Proteome Res.* **2009**, *8* (4), 1965−1971.
(19) MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
(20) Breci, L. A.; Tabb, D. L.; Yates, J. R., 3rd; Wysocki, V. H. Cleavage N-Terminal to Proline: Analysis of a Database of Peptide Tandem Mass Spectra. *Anal. Chem.* **2003**, *75* (9), 1963−1971.
(21) Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. J. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, UK, 1998.
(22) Edwards, A. W. F. *Likelihood*; Cambridge University Press: Cambridge, UK, 1972.
(23) Meng, F.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L. Informatics and Multiplexing of Intact Protein Identification in Bacteria and the Archaea. *Nat. Biotechnol.* **2001**, *19* (10), 952−957.
(24) Meng, F.; Cargile, B. J.; Patrie, S. M.; Johnson, J. R.; McLoughlin, S. M.; Kelleher, N. L. Processing Complex Mixtures of Intact Proteins for Direct Analysis by Mass Spectrometry. *Anal. Chem.* **2002**, *74* (13), 2923−2929.
(25) Taylor, G. K.; Kim, Y. B.; Forbes, A. J.; Meng, F.; McCarthy, R.; Kelleher, N. L. Web and Database Software for Identification of Intact Proteins Using "Top Down" Mass Spectrometry. *Anal. Chem.* **2003**, *75* (16), 4081−4086.
(26) LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L. ProSight PTM: An Integrated Environment for Protein Identification and Characterization by Top-Down Mass Spectrometry. *Nucleic Acids Res.* **2004**, *32* (Web Serverissue), W340−345.
(27) LeDuc, R. D.; Kelleher, N. L. Using ProSight PTM and Related Tools for Targeted Protein Identification and Characterization with

High Mass Accuracy Tandem Ms Data. *Curr. Protoc. Bioinf.* **2007**, DOI: 10.1002/0471250953.bi1306s19.
(28) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: Improved Protein Identification and Characterization for Top Down Mass Spectrometry. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W701−706.
(29) Sonego, P.; Kocsor, A.; Pongor, S. Roc Analysis: Applications to the Classification of Biological Sequences and 3d Structures. *Briefings Bioinf.* **2008**, *9* (3), 198−209.
(30) Kruger, N. A.; Zubarev, R. A.; Carpenter, B. K.; Kelleher, N. L.; Horn, D. M.; McLafferty, F. W. Electron Capture Versus Energetic Dissociation of Protein Ions. *Int. J. Mass Spectrom.* **1999**, *183*, 1−5.
(31) Ewing, B.; Green, P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.* **1998**, *8* (3), 186−194.
(32) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. Large-Scale Top Down Proteomics of the Human Proteome: Membrane Proteins, Mitochondria, and Senescence. *Mol. Cell. Proteomics* **2013**, *12*, 3465−3473.
(33) Catherman, A. D.; Li, M.; Tran, J. C.; Durbin, K. R.; Compton, P. D.; Early, B. P.; Thomas, P. M.; Kelleher, N. L. Top Down Proteomics of Human Membrane Proteins from Enriched Mitochondrial Fractions. *Anal. Chem.* **2013**, *85* (3), 1880−1888.
(34) Dang, X.; Scotcher, J.; Wu, S.; Chu, R. K.; Tolic, N.; Ntai, I.; Thomas, P. M.; Fellers, R. T.; Early, B. P.; Zheng, Y.; Durbin, K. R.; Leduc, R. D.; Wolff, J. J.; Thompson, C. J.; Pan, J.; Han, J.; Shaw, J. B.; Salisbury, J. P.; Easterling, M.; Borchers, C. H.; Brodbelt, J. S.; Agar, J. N.; Pasa-Tolic, L.; Kelleher, N. L.; Young, N. L. The First Pilot Project of the Consortium for Top-Down Proteomics: A Status Report. *Proteomics* **2014**, *14* (10), 1130−1140.
(35) Arnesen, T.; Van Damme, P.; Polevoda, B.; Helsens, K.; Evjenth, R.; Colaert, N.; Varhaug, J. E.; Vandekerckhove, J.; Lillehaug, J. R.; Sherman, F.; Gevaert, K. Proteomics Analyses Reveal the Evolutionary Conservation and Divergence of N-Terminal Acetyltransferases from Yeast and Humans. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (20), 8157−8162.
(36) Craig, R.; Cortens, J. P.; Beavis, R. C. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *J. Proteome Res.* **2004**, *3* (6), 1234−1242.