Quantifying Homologous Proteins and Proteoforms*

Dmitry Malioutov‡, Tianchi Chen§, Edoardo Airoldi¶, Jacob Jaffe¶, Bogdan Budnik**, and Nikolai Slavov§‡‡

Many proteoforms—arising from alternative splicing, post-translational modifications (PTM), or paralogous genes—have distinct biological functions, such as histone PTM proteoforms. However, their quantification by existing bottom-up mass-spectrometry (MS) methods is undermined by peptide-specific biases. To avoid these biases, we developed and implemented a first-principles model (HI\textsuperscript{quant}) for quantifying proteoform stoichiometries. We characterized when MS data allow inferring proteoform stoichiometries by HI\textsuperscript{quant} and derived an algorithm for optimal inference. We applied this algorithm to infer proteoform stoichiometries in two experimental systems that supported rigorous benchmarking: alkylated proteoforms spiked-in at known ratios and endogenous PTM proteoforms quantified relative to internal standards. When compared with the benchmarks, the proteoform stoichiometries inferred by HI\textsuperscript{quant} without using external standards had relative error of 5–15% for simple proteoforms and 20–30% for complex proteoforms. A HI\textsuperscript{quant} server is implemented at: https://web.northeastern.edu/slavov/2014_HIquant/

Alternative mRNAs splicing and post-translational modifications (PTM)\textsuperscript{1} produce multiple protein isoforms per gene, termed proteoforms by Smith et al. (1). Furthermore, protein isoforms can be produced by distinct but highly homologous open reading frames, i.e. paralogous genes. Despite having similar sequence, proteoforms and protein isoforms often have distinct, even opposite biological functions (2, 3). For examples: (1) some Bcl-x isoforms promote apoptosis whereas other Bcl-x isoforms inhibit apoptosis (4); (2) the methylation of histone 3 can cause either transcriptional activation (lysine 4) or repression (lysine 9) depending on the modified lysine (5); (3) pyruvate kinase isoforms have different metabolic regulation, activities, and roles in aerobic glycolysis (6–8); and (4) minor amino acid differences between actin and tropomyosin isoforms alter the tropomyosin position on actin (9).

Understanding such systems demands accurate methods to distinguish and quantify proteoform abundances (3, 10). However, the high sequence homology among proteoforms makes their quantification by bottom-up LC-MS/MS very challenging because protease digestion (which is an essential part of high-throughput bottom-up proteomics) is likely to produce mostly peptides that are shared by multiple proteoforms (11–13). One approach to overcoming the challenge of shared peptides and distinguishing among proteoforms is to analyze the intact proteoforms without protease digestion, which is known as top-down proteomics (10, 14–16). Combining advances in chromatographic separation of proteins with top-down proteomics has enabled quantifying functionally distinct isoforms of skeletal and cardiac muscle proteins (9, 17), combinatorial histone modifications (18, 19) and discovery analysis of thousands of human proteoforms (10, 20). Distinguishing proteoforms has been powered by efficient fragmentation methods, such as electron capture dissociation (ECD) (21) and electron transfer dissociation (ETD) (22). Combining these methods with infrared photoactivation (23) and with clever algorithms has made it possible to localize PTMs and identify the sequences and modified sites even for peptides with multiple covalent modifications (24, 25). Such methods allow quantifying the ratios among modifications occurring on the same peptide.

Although advances in protein separation and top-down methods are becoming increasingly powerful (10, 16, 20), bottom-up methods remain more widely used and still afford higher throughput (26, 27). When analyzed by bottom-up methods, many proteoform-specific modifications occur on distant parts of the corresponding proteins and upon protease digestion will be found on different peptides. Furthermore, most of the peptides may be shared among proteoforms (11–13). Here we focus on quantifying the stoichiometry among such proteoforms from peptides quantified by bottom-up LC-MS/MS across multiple samples.
One validated approach to quantifying the stoichiometry among such proteoforms relies on external standards (28, 29). This method can afford high accuracy even for complex proteoforms (29). However, their wider use has been limited by expense and applied only to special cases that allow chemical modification of cell lysates, e.g. phosphorylation (30) and acetylation (31, 32). In the absence of external standards, the quantification of complex proteoform stoichiometries remains very challenging because the ratios between proteoform-specific peptides do not necessarily reflect the ratios between the corresponding proteoforms (33); precursor ion areas corresponding to the same phospho-site in the same sample can differ over 100-fold depending on the choice of protease (34). This discrepancy arises because a measured peptide level (precursor ion area) depends not only on the abundance of the corresponding protein(s) but also on extraneous factors including protein digestion, peptide ionization efficiency, the presence of other co-eluting peptides, and chromatographic aberrations (34–36). These extraneous factors break the equivalence between the abundance of a peptide and its precursor ion area and thus make protein quantification much more challenging than DNA quantification by sequencing. This problem is compounded when PTM peptides have been enriched, and thus their intensities scaled by unknown enrichment-dependent factors.

**EXPERIMENTAL PROCEDURES**

We used two experimental approaches to derive bench-mark estimates of proteoform stoichiometries: (1) We mixed alkylated proteoforms of the dynamic universal proteomics standard (UPS2) into known ratios and (2) We quantified stoichiometries among histone PTMs by parallel reaction monitoring (PRM) relative to peptide standards with known absolute abundances. Each of these methods is described in the sub-sections below, and the resulting bench-mark estimates of proteoform stoichiometries were used to evaluate Hiquant inferences.

Quantifying Alkylated proteoforms of UPS2—Our first test of Hiquant on experimental data with bench-marked proteoforms was based on alkylated proteoforms of UPS2 that we mixed into predetermined ratios. To achieve such mixtures, we started by splitting a UPS standard into two equal parts, part A and part B. Then:

- Part A was reduced with TCEP, and Cys were alkylated with iodoacetamide.
- Part B was reduced with TCEP, and Cys were alkylated with vinylpyridine.

The alkylated proteoforms from both part A and part B were diluted to a total volume of 320 μl and spiked into yeast samples according to Table I.

This range of mixing ratios, spanning an order of magnitude, was chosen as it includes the range of relative changes commonly measured reliably by mass-spectrometry. The fractional site occupancy between the proteoforms varied within 10-fold in each direction, the range over which Hiquant is likely to give the most accurate results. As discussed below, the ratios among proteoforms whose levels differ by many orders of magnitude may be better inferred from the precursor ion areas of their unique peptides.

After the spike in of UPS alkylated proteoforms, each sample was digested by trypsin overnight and labeled with the corresponding TMT tag according to manufacturer’s protocol. The labeled samples were mixed into a set and processed as described previously (8). Briefly, the set-sample was injected from an auto-sampler into the trapping column (75 μm column ID, 5 cm length, and packed with 5 μm beads with 20 nm pores; from Michrom Biosources, Inc.) and washed for 15 min; the sample was eluted to analytic column (Waters columns with 75 μm ID, 15 cm length, and packed with HSS T3 1.8 μm beads) with a gradient from 2 to 32% of buffer B (0.1% formic acid in ACN) over 180 min gradient and fed into LTQ Orbitrap Elite (Thermo Fisher, San Jose, CA). The instrument was set to run in TOP 20 MS/MS mode method with dynamic exclusion. After MS1 scan in Orbitrap with 60K resolving power, each ion was submitted to an HCD MS/MS scan subsequently. All quantification data were derived from HCD spectra.

In the particular case of UPS2, trypsin digestion resulted in enough peptides to constrain the Hiquant model (Fig. 1) and to support accurate inference (Fig. 2B, 2C). More complex proteoforms may require more peptides than those quantified from a trypsin digestion. In such cases, to better constrain the inference, one might increase the number of peptides used by Hiquant by compiling peptides quantified from several digestions, each of which using a different protease (34, 36). Because Hiquant uses only the relative quantification and is insensitive to systematic biases, including proteoform-specific digestion efficiency, it can easily accommodate peptides quantified from digestions with different proteases or from different enrichment methods.

---

1 The abbreviations used are: PTM, post-translational modifications; PSM, peptide spectrum matches; FDR, false discovery rate; PRM, parallel reaction monitoring.
Quantifying Homologous Proteins and Proteoforms

Analysis of Mass-Spec Spectra—Mass/charge spectra were analyzed by MaxQuant (37) (version 1.4.1.2) or SEQUEST HT run via the Proteome Discoverer (64 bit version 1.4.0.288, Thermo). All searches were run on a Windows server 2008 64 bit operating system with 64 CPU blades and 256 GB of RAM with the following general parameters. Parent ion mass tolerance was set to 20 ppm, mass tolerance for MS/MS ions was set to 0.02 Da for HCD and to 0.6 Da for CID spectra. For all searches, minimal peptide length was specified as 6 amino acids and maximal peptide length as 50 amino acids. The peptide charge state was limited to /H11001 7 for searches with MaxQuant. Searches were performed against the yeast uniprot database downloaded from www.yeastgenome.org in October 2014 containing 6,750 entries, the.fasta sequences of the UPS standard, and common contaminants. Searches had trypsin specificity, allowing 2 missed cleavages. Asn and Gln deamidation, Cys alkylation by either iodoacetamide or vinylpyridine, and Met oxidation were included as variable modifications in the search parameters. There were no fixed modifications.

The search results from all search engines were filtered at 1% false discovery rate (FDR) on both protein and on peptide levels using the Percolator (Version 2.05 Build Date May 6 2013). The results exported for further analysis included all peptide spectrum matches (PSM) that were assigned to one or more proteins and passed the statistical significance filter. These results were outputted in the “evidence.txt” file for MaxQuant and in a peptide-level-results text file for Proteome Discoverer.

PRM of Histone PTMs Relative to Heavy Peptide Standards—To evaluate Hlquant inferences, we also used PTM stoichiometries estimated from a parallel reaction monitoring (PRM) assay quantifying each peptide relative to an internal heavy standard. Creech et al. have published a detailed description of the assay (29), and here we will only briefly summarize it. Drug-treated cells were collected by centrifugation. After cell lysis, histones were extracted with sulfuric acid and were precipitated with trichloroacetic acid. Samples comprised of 10 μg, a 5-fold reduction compared with Creech et al. (29), were propionylated, desalted, and digested overnight with trypsin. After a second round of propionylation, the samples were desalted using C18 Sep-Pak Cartridge (Waters). A mix of isotopically labeled synthetic peptides was spiked-in to each sample prior to MS analysis. Peptides were separated on a C18 column (EASY-nLC 1000, Thermo Scientific) and analyzed by MS in a PRM mode (Q Exactive-plus, Thermo Scientific) as described previously (29). Detailed SOPs for P100 and GCP assays, including synthetic peptide master mixture formulation, can be found at https://panoramaweb.org/labkey/wiki/LINCS/Overview%20Information/page.view?name=sops.

Inference Model for Hlquant—To infer proteoform stoichiometry, we use a simple model that is illustrated in Fig. 1A with proteoforms of histone H3 and in supplemental Fig. S1 with paralogous ribosomal proteins and phosphoproteoforms of pyruvate dehydrogenase. Hlquant explicitly models peptide levels measured across conditions (x) as a superposition of the protein levels (p), scaled by unknown peptide-specific biases/nuisances (z). These coupled equations can be written in a matrix form whose solution infers the methylation stoichiometry independently from the nuisances (z). B, The general form of the model for K proteoforms (or homologous proteins) with M peptides quantified across N conditions can be formulated and solved. In many, albeit not all, cases an optimal and unique solution can be found, even in the absence of unique peptides; see supplemental Fig. S1 and Supplemental Information.

![Model for inferring stoichiometries among proteoforms and paralogous proteins independently from peptide-specific biases.](http://www.mcponline.org/)
as the system in Fig. 1 shows. HIquant solves this system and infers the protein levels ($P$) independently from the extraneous noise ($Z$; coming from protein-digestion, peptide-ionization differences, sample loss during enrichment, and even coisolation interference); $Z$ is also inferred as part of the solution and discarded. A related superposition model has been used before with peptides quantified at one condition (13). However, for a single condition, the model cannot quantify the proteins independently from the nuisance $Z$ because all problems described by system 1 in Fig. 1 are underdetermined, i.e. have infinite number of solutions (Proof 1; Supplemental Information). Thus, for a single condition, the model cannot take advantage of the robust corresponding-ion pairs, i.e. ratios between ions with the same chemical composition. In contrast, HIquant infers ratios across proteins and their PTMs solely from the corresponding-ion ratios. This is possible because when $N > 1$, the system in Fig. 1B often has a unique solution up to a single scaling constant, even when all peptides are shared, e.g. the problem defined by the design matrix in supplemental Fig. S1C. We characterize the conditions under which HIquant has a unique solution for the abundances of individual proteoforms and derive algorithms that use convex-optimization to find the optimal solution given the data; see Malioutov and Slavov (38) and Supplemental Information.

VALIDATING INFERENCE OF PROTEOFORM STOICHIOMETRY

Our model (Fig. 1B) aims to make proteoform quantification insensitive to many systematic biases. For example, incomplete cleavage of a peptide, e.g. only 5% of the peptide is released during enzyme digestion, is fully absorbed into the corresponding nuisance and does not affect inferred protein levels if the cleavage is 5% for all conditions/samples. Analogously, if coisolation interference compresses the fold-
changes of a peptide, the systematic component of the compression is fully absorbed by the nuisances. Unlike systematic biases, random noise in the data is not absorbed by the nuisances; it can degrade the quality of the inference. To assess the reliability of the inferred proteoform abundances, HLquant carefully evaluates the inference and assigns confidence levels. The evaluation uses inference features, such as fraction of explained variance, eigenvalue spectrum spacing and noise sensitivity; see Supplemental Information.

We sought to experimentally evaluate HLquant’s ability to infer the proteoform stoichiometry in samples for which proteoform stoichiometries are accurately determined by other methods. The first method included creating and mixing alkylated proteoforms. The second method included quantifying histone H3 proteoforms relative to heavy peptide standards with known abundances.

Validation Based on Spiked-in Alkylated Proteoforms—We aimed to create proteoform mixtures with known stoichiometries so that they can be used to assess the accuracy of stoichiometries inferred by HLquant. To this end, the dynamic universal proteomics standard (UPS2) was split into two equal parts, A and B. In part A, cysteines were covalently modified with iodoacetamide, and in part B with vinylypyridine as described in the methods and shown in Fig. 2A. We mixed part A and B in predefined ratios (n) and spiked each mixing ratio into an yeast sample. All samples were labeled with TMT, and the relative peptide levels quantified from the reporter ions at the MS2 level.

These alkylated UPS proteoforms have mostly shared peptides (peptides not containing cysteine) and only one or a few unique peptides per proteoform (peptides containing cysteine). HLquant modeled the relative levels of these peptides as shown in Fig. 1 and solved the model to infer the stoichiometries of the alkylated proteoforms (n), which should correspond to the mixing ratios. Indeed, we find that the actual mixing ratios (n) for all quantified proteoforms correlate strongly to the inferred ratios (n) as shown in Fig. 2B. To examine the accuracy of the inferred ratios more closely, we sought to generate a distribution of errors between the inferred and the expected mixing ratios (n/n) for many HLquant problems, more than the number of UPS2 proteoforms. To do so, we took advantage of the fact that the A/B ratios of the alkylated proteoforms for different UPS2 proteins should be the same because the proteins were mixed simultaneously and therefore in equal proportions. Thus, we created 1500 HLquant problems by sampling with replacement shared peptides (not containing cysteine) and unique peptides (containing cysteine), and then inferred the n ratios between the iodoacetamide and vinylypyridine modified proteoforms, one ratios for each TMT channel and 6 per HLquant problem. The corresponding distribution of relative errors is shown in Fig. 2C along with the errors for ratios estimated from the abundances of the precursor ions. The median error is below 11% for ratios inferred by HLquant and a substantially larger error for the ratios between precursor ion areas of unique peptides (Fig. 2C).

Validation Based on Histone 3 Proteoforms Quantified by PRM—Next, we sought to evaluate the ability of HLquant to infer stoichiometries of more complex PTM proteoforms, those of histone H3. We rigorously quantified endogenous proteoform stoichiometries by a previously developed assay based on external standards (MasterMix) with known concentrations (29). For the test, we used peptides quantified by parallel reaction monitoring (PRM) across 7 drug perturbations. Fractional site occupancies were estimated based on the external standards as described before by Creech et al. (29). Independently, the same stoichiometries were inferred by HLquant only from the relative levels of the indigenous peptides, without using the MasterMix concentrations. The comparison of these estimates indicates good agreement (Fig. 3), supporting the ability of HLquant to infer fractional site occupancy even when the same site may be modified by different PTMs. The estimates from the external standards and from HLquant are very close but also show some systematic deviations. Those deviations may arise because of incomplete protein digestion that is hard to control for with peptide standards, measurement noise corrupting the solution inferred by HLquant or proteoforms not explicitly included in the model. The abundances of some proteoforms with quantified peptides is over 1000-fold lower than the abundance of the main proteoforms. They and their corresponding peptides were omitted from the HLquant inference because their quantification requires unrealistically high accuracy of relative quantification; see Supplemental Information and Discussion.

DISCUSSION

The idea of using ratios between chemically identical ions is a cornerstone of quantitative proteomics (28, 39). It has been used for two decades in the context of relative quantification of proteins based on unique peptides (40) and even applied to the special case of inferring phosphorylation site occupancy (11, 33). Our work expands and generalizes this idea to all peptides, to stoichiometries of complex proteoforms, and to unlimited number of conditions. Crucially, HLquant allows accurate, efficient, and numerically stable inference resulting in reliability estimates.

HLquant requires and depends upon accurate relative quantification. This limitation is largely and increasingly mitigated by technological developments allowing accurate estimates of corresponding ion ratios. Such improvements include instrumentation advances (41, 42), interference free tandem mass tags (43), and enhanced peptide quantification from mass-spectra (27). However, these technological developments on their own do not allow accurate estimates of PTM site occupancy from bottom-up LC-MS/MS (34). HLquant’s dependence on the accuracy of relative quantification in-
creases with increasing difference in the abundance of proteoforms. If the levels of two proteins differ by more than 3–6 orders of magnitude, this difference is likely better inferred from the precursor ion areas of the unique peptides. The associated noise (because of variability in protein digestion and ionization) is generally below 100-fold (36) and thus smaller than the signal. HIquant’s utility is particularly relevant when proteins and proteoforms have comparable abundances (within 10–100 fold difference) but distinct functions (44, 45) and thus accurate quantification is essential for quantifying relatively small differences in abundance. Quantifying such proteoforms is an exciting frontier essential for understanding post-transcriptional regulation (46, 47) and defining cell-types from single cell proteomes (48).

The general form of HIquant described in Fig. 1C indicates that HIquant is not limited to proteoforms, even broadly defined. Rather, HIquant can be applied to any set of proteins sharing a peptide. Here we emphasize the application to proteoforms because existing bottom-up methods are better suited for quantifying the stoichiometry between proteins with low homology that generate many unique peptides. For proteins with multiple unique peptides, some of the peptide-specific bias (from variation in protein-digestion and peptide-ionization efficiency) is likely to be averaged out and reduced. However, this bias is a more serious problem for proteoforms with only one or only a few unique peptides (34). For such proteoforms, HIquant can allow estimating stoichiometries accurately using only ratios between chemically identical ions.

Supplemental Information—Supplemental information includes Extended Experimental Procedures, Mathematical Proofs, and Supplemental Figs. can be found in the Supplemental Information. The supplemental website for interactive data analysis can be found at: https://web.northeastern.edu/slavov/2014_HIquant/ Python code implementing HIquant is available at: https://github.com/nslavov/HIquant.

Acknowledgments—We thank S. Carr, A. Ivanov, S. MacNamara, Y. Katz, and MA. Blanco for help, critical discussions and feedback. N.S. started this work while a postdoc with Alexander van Oudenaarden and thanks him for generous support.

DATA AVAILABILITY

The data are available at ProteomeXchange (accession ID: PXD008557, url: http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD008557) via MassIVE (accession ID: MSV000081857, url: https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=38f67398ac7a4547b68f343ca1aff905).

* This work was supported by grants to N.S. from NIGMS of the NIH under Award Number DP2GM123497, a SPARC grant from the Broad Institute to N.S. and S.C, and a NEU Tier 1 grant to N.S.

†† To whom correspondence should be addressed: Department of Bioengineering, Northeastern University, Boston, MA 02115. E-mail: nslavov@alum.mit.edu.

Author contributions: D.M., T.C., B.B., and N.S. performed research; D.M., E.A., J.D.J., and N.S. contributed new reagents/analytic tools; D.M. and N.S. analyzed data; J.D.J. and N.S. wrote the paper; N.S. designed research.

REFERENCES


