Protein Identification Using Top-Down*

Xiaowen Liu‡, Yakov Sirotkin§, Yufeng Shen¶, Gordon Anderson¶, Yihsuan S. Tsai∥, Ying S. Ting∥, David R. Goodlett∥, Richard D. Smith¶, Vineet Bafna‡, and Pavel A. Pevzner‡§**

In the last two years, because of advances in protein separation and mass spectrometry, top-down mass spectrometry moved from analyzing single proteins to analyzing complex samples and identifying hundreds and even thousands of proteins. However, computational tools for database search of top-down spectra against protein databases are still in their infancy. We describe MS-Align+, a fast algorithm for top-down protein identification based on spectral alignment that enables searches for unexpected post-translational modifications. We also propose a method for evaluating statistical significance of top-down protein identifications and further benchmark various software tools on two top-down data sets from Saccharomyces cerevisiae and Salmonella typhimurium. We demonstrate that MS-Align+ significantly increases the number of identified spectra as compared with MASCOT and OMSSA on both data sets. Although MS-Align+ and ProSightPC have similar performance on the Salmonella typhimurium data set, MS-Align+ outperforms ProSightPC on the (more complex) Saccharomyces cerevisiae data set. Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.008524, 1–13, 2012.

In the past two decades, proteomics was dominated by bottom-up mass spectrometry that analyzes digested peptides rather than intact proteins. Bottom-up approaches, although powerful, do have limitations in analyzing protein species, e.g. various proteolytic forms of the same protein or various protein isoforms resulting from alternative splicing. Top-down mass spectrometry focuses on analyzing intact proteins and large peptides (1–10) and has advantages in localizing multiple post-translational modifications (PTMs) in a coordinated fashion (e.g. combinatorial PTM code) and identifying multiple protein species (e.g. proteolytically processed protein species) (11). Until recently, most top-down studies were limited to single purified proteins (12–15). Top-down studies of protein mixtures were restricted by difficulties in separating and fragmenting intact proteins and a shortage of robust computational tools.

In the last two years, because of advances in protein separation and top-down instrumentation, top-down mass spectrometry moved from analyzing single proteins to analyzing complex samples containing hundreds and even thousands of proteins (16–21). Because algorithms for interpreting top-down spectra are still in their infancy, many recent developments include computational innovations in protein identification.

Because top-down spectra are complex, the first step in top-down spectral interpretation is usually spectral deconvolution, which converts a complex top-down spectrum to a list of monoisotopic masses (a deconvolved spectrum). Every protein (possibly with modifications) can be scored against a top-down deconvoluted spectrum, resulting in a Protein-Spectrum-Match (PrSM). The top-down protein identification problem is finding a protein in a database with the highest scoring PrSM for a top-down spectrum and further output the PrSM if it is statistically significant. There are several software tools for top-down protein identification (Table I).

- ProSightPC—ProSightPC is the most commonly used tool for top-down protein identification (22, 23). ProSightPC searches spectra against a “shotgun annotated” protein database, which is generated by considering all expected PTMs. The “shotgun annotated” protein database is much larger than the original protein database. ProSightPC can identify some (but not all) proteins with unexpected PTMs using advanced search options, such as biomarker search and Δm mode, but it is not designed for identifying truncated proteins with unexpected PTMs that are not represented in the “shotgun annotated” database. ProSightPC is a fast tool that reports the statistical significance of PrSMs.

- PIITA—Unlike ProSightPC, PIITA (19) is a precursor independent method that uses only fragment ions for protein identification. It is capable of identifying protein species with unexpected PTMs on N- or C-termini, but it cannot directly identify protein species with PTMs on both N- and C-termini.
Protein Identification Using Top-Down

| Table I | Characteristics of various top-down protein identification software tools |
|---------|-----------------------------|
| Software | Identification of unexpected modifications | Proteogenomics search against 6-frame translation | Speed | Estimation of statistical significance |
| ProSightPC | +/− | + | Fast/Slow | − |
| PIITA | +/− | − | Fast | − |
| UTag | + | + | Fast | − |
| MS-TopDown | + | − | Slow | − |
| MS-Align+ | + | + | Fast | + |

* +/− indicates various search modes that contribute to bridging the gap between blind and restrictive modes of MS/MS database search.

** In its most advanced mode, ProSightPC can search the annotated top-down database that contains various protein species. However, ProSightPC searches in this mode at an order of magnitude slower.

PIITA is a fast tool that provides FIT scores and Δ scores rather than statistical significance estimates.

- **USTag**—Unique Sequence Tag (USTag) (17) generates long (6 amino acids or longer) peptide sequence tags to identify PrSMs. This approach, although fast, relies on long peptide sequence tags that may be difficult to obtain for some spectra. It also does not provide an estimate of the statistical significance of PrSMs.

- **MS-TopDown**—MS-TopDown (24) is based on spectral alignment (25). MS-TopDown allows one to match top-down spectra to proteins with unexpected PTMs, i.e. without knowing which PTMs are present in the sample. However, MS-TopDown is rather slow when searching against large proteomes and does not provide the statistical significance of PrSMs.

- **MS-Align+**— shares the spectral alignment approach with MS-TopDown, but greatly improves on speed, statistical analysis (providing E-values of PrSMs), and the number of identified PrSMs (e.g. by finding spectral alignments between spectra and truncated proteins). We benchmarked various tools for top-down protein identification on two data sets from Saccharomyces cerevisiae (SC) and Salmonella typhimurium (ST). We demonstrate that MS-Align+ significantly increases the number of identified spectra as compared with MASCOT and OMSSA on both data sets. Although MS-Align+ and ProSightPC have similar performance on the ST data set, MS-Align+ outperforms ProSightPC on the more complex SC data set.

**EXPERIMENTAL PROCEDURES**

**Spectral alignment**—Some MS/MS tools searching for unexpected PTMs transform a spectrum into a prefix residue mass (PRM) spectrum (28) that represents various peptide prefixes supported by the spectrum. For example, a monoisotopic peak with mass \( m \) corresponding to a \( y \)-ion in the MS/MS spectrum is converted to a mass \( M - m \) in the PRM spectrum (\( M \) is the precursor mass). A deconvoluted spectrum with \( k \) monoisotopic masses and precursor mass \( M \) is converted to a PRM spectrum \( S = 2k + 2 \) masses. For a CID spectrum, we include into \( S \) two masses \( m \) and \( M - m \) for each monoisotopic mass \( m \) and two other masses \( 0 \) and \( M - 18 \) (\( M - 18 \) is treated as the PRM of the entire protein). We represent spectrum \( S \) as a list of ordered masses \( a_0 < a_1 < \ldots < a_n \), where \( a_0 = 0 \) and \( a_n = M - 18 \).

A protein \( P \) of length \( m \) is represented as theoretical peak masses \( b_0 < b_1 < \ldots < b_m \), where \( b \) equals the sum of the masses of the first \( i \) residues in \( P \) (we assume that \( b_0 = 0 \) and \( b_m = M' - 18 \) where \( M' \) is the molecular mass of the unmodified protein form). Although, for the sake of simplicity, we ignore intensities of the peaks, intensities can be easily incorporated in the spectral alignment framework.

Given a protein \( P = \{b_0, b_1, \ldots, b_m\} \) and a spectrum \( S = \{a_0, a_1, \ldots, a_n\} \), we define the spectral grid of \( P \) and \( S \) as a rectangle in two dimension formed by four points \((0, 0), (0, m), (n, 0), (n, m)\) with \((n + 1)(m + 1)\) matching points \((b_i, a_j)\) for \( 0 \leq i \leq n \) and \( 0 \leq j \leq m \)

Located within the rectangle. A global spectral alignment between \( P \) and \( S \) is a path from the top left corner to the bottom right corner of the spectral grid consisting of diagonal (−45°), vertical, and horizontal segments (Fig. 1A). Diagonal segments correspond to subpeptides of \( P \) matched to \( S \) without modifications; vertical and horizontal segments correspond to PTMs with positive and negative offsets, respectively. The length of a horizontal or vertical segment represents the absolute value of the offset. For example, the spectral alignment in Fig. 1A has three diagonal segments (corresponding to subpeptides “PR”, “TEI”, and “STRING”), one vertical segment (a modification or mutation with +80 Da offset) and one horizontal segment (a modification or mutation with −114 Da offset).

We often observe top-down spectra mapped to proteins with truncated peptides on N- and/or C-termini. To identify these spectra, we define a semiglobal spectral alignment between \( P \) and \( S \) as a path from a matching point \((b_x, 0)\) on the top border to a matching point \((b_y, a_y)\) with \( x < y \), on the bottom border in the spectral grid, consisting of diagonal, vertical, and horizontal segments.

Based on the positions of \( b_x \) and \( b_y \), we classify semiglobal spectral alignments into four groups: complete \((x = y = m)\), prefix \((x = 0 \text{ and } y < m)\), suffix \((x > 0 \text{ and } y = m)\), and internal \((x > 0 \text{ and } y < m)\) (Fig. 1). A complete alignment is characterized by a path from the top left corner to the bottom right corner of the spectral grid, corresponding to a global spectral alignment between the spectrum and the entire protein. A prefix/suffix/internal alignment corresponds to a global spectral alignment between the spectrum and a prefix/suffix/ internal subprotein of the protein. An alignment is called a nonshift alignment if it has no offsets and contains only one diagonal segment. Otherwise, it is called a shift alignment.

A matching point is on a diagonal segment if the distance between the point and the segment is within a given error tolerance. The score of an alignment is the number of matching points on the diagonal.
segments in the alignment path. For example, the path in Fig. 1A represents an alignment with a score of 12. Spectral alignment algorithms find a highest-scoring alignment between a deconvoluted spectrum and a protein \( P \). Frank et al., 2008 (24) described a dynamic programming algorithm for finding an optimal global alignment between \( P \) and \( S \) with \( F \) offsets in \( O(nmF) \) time. This algorithm with minor modifications still works for finding an optimal semi-global spectral alignment between a deconvoluted spectrum and a protein.

**Speeding up Spectral Alignment**—The spectral alignment algorithm in (24) (e.g. its MS-TopDown implementation) typically takes 3 s to align a single spectrum against a single protein (on a typical desktop machine with a 2.2 GHz CPU). Aligning a small data set with 1000 spectra against a small proteome with 1000 proteins takes \( 3 \times 10^6 \) seconds (\( \sim 800 \) h) and makes spectral alignment computationally intensive. To speed up protein identification, we quickly filter out proteins in the proteome that cannot possibly attain high-scoring PrSMs for a given spectrum, thus reducing the number of candidate proteins.

Given a –45° line crossing the spectral grid, we define its weight as the number of matching points on this line. This definition of weight is different from the score of a spectral alignment, which is defined as the number of matching points on its corresponding path. We rank all crossing lines in the decreasing order of their weights and output top

**Fig. 1. Spectral alignment examples.** A, A global spectral alignment (complete alignment) between a spectrum and a protein \( P = \text{PRTEINSTRING} \). The blue path from the top left corner to the bottom right corner represents the alignment of protein \( P \) and the spectrum arising from \( P \) with two PTMs: +80 Da on “T” and −114 Da between “I” and “S” (deletion of ‘N’). The circles along the path denote the matching points on the path. B, A prefix alignment between a spectrum and protein \( P \). The path from the top left corner to a matching point on the bottom border represents an alignment between the spectrum and a prefix “PRTEINSTRI” of \( P \). C, A suffix alignment between a spectrum and protein \( P \). The path from a matching point on the top border to a matching point on the bottom border represents an alignment between the spectrum and a subsequence “EINSTRI” of \( P \).

**Fig. 2. The top five crossing lines in the spectral grid of the protein \( P = \text{PRTEINSTRING} \) and the spectrum in Fig. 1A.** The blue crossing line has the largest weight 7. The alignment in Fig. 1A is composed of three diagonal segments from the red, green, and blue crossing lines. The two black crossing lines, with weights 3 and 2, are not in the path of the best alignment.

\( \alpha \) crossing lines (Fig. 2). The weights of crossing lines can be quickly computed using spectral convolution (see (29) for the definition of spectral convolution).
Protein Identification Using Top-Down

For a spectral alignment with a score \( s \) and \( k \) diagonal segments, there exists a crossing line in the spectral grid with weight at least \( s/k \). Thus, if all crossing lines in the spectral grid have low weights, the corresponding protein can be safely filtered out, eliminating the need to run spectral alignment against this protein. We use this filter to reduce the number of candidate proteins, thus speeding up protein identification. Given a spectrum, the diagonal score of a protein is the maximum weight among all crossing lines in the spectral grid. We rank all proteins in the proteome based on their diagonal scores and filter out all but top \( \beta \) high-scoring proteins that are further subjected to spectral alignment (Fig. 3). The default value for \( \beta \) is 20.

If a crossing line contains a diagonal segment that is in an optimal spectral alignment, the crossing line usually has a relatively heavier weight compared with a random crossing line. Therefore, we first select top \( \alpha \) crossing lines (default value \( \alpha = 20 \)) for a protein spectrum pair, and use a dynamic programming algorithm to find an optimal path containing only diagonal segments on the spectral grid. We rank all proteins in the proteome based on their diagonal scores and filter out all but top \( \beta \) high-scoring proteins that are further subjected to spectral alignment (Fig. 3). The default value for \( \beta \) is 20.

![Distribution of diagonal scores when comparing a single spectrum (with 187 monoisotopic masses) against all proteins in the Salmonella typhimurium proteome. The mean value of the diagonal scores (across all proteins) is 5 whereas the diagonal score for the correct PrSM is 54.](image)

**RESULTS**

Data Sets—Two top-down data sets from Saccharomyces cerevisiae and Salmonella typhimurium were used for benchmarking:

**S. cerevisiae (SC) Data Set (16)**—A lysate was quickly extracted from SC cell with the use of pressure cycling technology and in the presence of a protease inhibitor. The lysate obtained was directly separated on an LC system that was coupled online to an LTQ-Orbitrap (Thermo Fisher Scientific). A total of 30,760 FT-MS/MS spectra were acquired during the 600-min LC separation. Both FT-MS and MS/MS spectra were collected at a resolution of 30,000. The charge states of the MS/MS spectra range from 1 to 30 and the precursor masses range from 800 to 20 KDa (supplemental Fig. S2A).

**S. typhimurium (ST) Data Set (19)**—The proteins extracted from ST were reduced with dithiothreitol and alkylated with iodoacetamide, and desalted on a C4 spin column. The protein mixture obtained was separated with a NanoAcuity
HPLC system that was coupled online to an LTQ-Orbitrap (Thermo Fisher Scientific). MS and MS/MS spectra were collected at a resolution of 60,000 and 30,000, respectively. The experiment was repeated using gas-phase fractionation. The detailed experiment procedure can be found in (19). A total of 14,041 FT-MS/MS spectra were acquired. The charge states of the MS/MS spectra range from 1 to 24 and the precursor masses range from 1K to 20K Da (supplemental Fig. S2B).

Data Preprocessing—Thermo raw files were first converted to mzXML files using ReAdW (http://tools.proteomcenter.org/software.php), and each MS/MS spectrum was converted to a deconvoluted spectrum (a monoisotopic mass list) using MS-Deconv (31). A total of 11,030 and 4439 deconvoluted spectra had a precursor mass \( < 11350 \) Da and at least 10 fragment peaks in the SC and ST data sets, respectively. We focused our attention on these deconvoluted spectra because shorter peptides are well handled by the existing bottom-up approaches and spectra with a small number of peaks seldom have good PrSMs. Below these 11,030 and 4439 spectra are referred to as SC* and ST* data sets, respectively. Some of the spectra had very similar precursor mass (up to 15 ppm error tolerance), suggesting that some of them may correspond to the same protein species. For error tolerance 15 ppm, there were 4462 and 1825 various precursor masses in the two sets of spectra.

Protein Identification—MS-Align+ was applied to search the deconvoluted spectra against the corresponding proteomes using the parameter settings described in supplemental Table S4. The SC and ST protein databases (downloaded from www.yeastgenome.org and NCBI) had 5885 and 4527 proteins, respectively. Error tolerance for fragment and precursor ions was set to 15 ppm. Because various deconvolution software tools often report monoisotopic masses that deviate from correct masses by \( < 1 \) Da, this type of shift error was allowed in spectral alignment. We emphasize the difference between the shift errors (that represents large but fixed \( \pm 1 \) Da shifts from correct masses) and typical (small) errors in peak position, e.g. 15 ppm error from correct masses. To reflect shift errors, a theoretical peak and a spectral peak with mass \( m \) were classified as matching peaks if the mass of the theoretical peak was matched to either \( m - 1 \), or \( m \), or \( m + 1 \) within 15 ppm. Carbamidomethylation of cysteine was used as a fixed modification for ST* data set, and no fixed modifications for SC* data set. MS-Align+ allows users to select some variable modifications to boost the number of identified proteins with particular frequent modifications. Protein N-terminal acetylation, protein N-terminal methionine loss, and protein N-terminal methionine loss plus acetylation were used as variable modifications for the two data sets. In addition, two mass shifts were allowed in spectral alignment (N- and C-terminal truncated peptides were not counted as mass shifts). Thus, every identified PrSM corresponded to up to 4 modifications other than the specified fixed and variable modifications. MS-Align+ reported only the PrSM with the best \( E \)-value for each spectrum. The running time of MS-Align+ for analyzing SC* and ST* data sets on a PC with a 2.67 GHz CPU and 12 GB memory was 724 and 495 min.

To estimate False Discovery Rate (FDR), we also searched the spectra in SC* and ST* data sets against decoy SC and ST databases, respectively (32). The decoy databases were generated by shuffling each protein sequence in the original data sets. Spectrum level FDR was estimated based on the distributions of the \( E \)-values of the identified PrSMs in the target and decoy databases (supplemental Figs. S3, S4). With spectrum level 1% FDR, MS-Align+ identified 4059 spectra (290 proteins) and 2217 spectra (180 proteins), corresponding to about 36.8% and 49.9% identification rates, from SC* and ST* data sets, respectively (supplemental Table S1). Most unidentified spectra were characterized by small numbers of deconvoluted peaks (typically less than 30) and small diagonal scores (typically less than 10) (Fig. 4).

Comparison with MASCOT, OMSSA, PIITA, and ProSightPC—Karakacak et al., 2009 (26) benchmarked MASCOT and ProSightPC with respect to analyzing top-down spectra. In this paper, MS-Align+ was compared with MASCOT, OMSSA, PIITA, and ProSightPC.
Comparison with MASCOT and OMSSA—Both MASCOT and OMSSA are popular software tools for bottom-up protein identification (27, 33) that are sometimes used for analyzing top-down spectra.

Using the same target/decoy approach, with spectrum level 1% FDR, MASCOT identified 3111 spectra (217 proteins) and 571 spectra (76 proteins), and OMSSA identified 2874 spectra (212 proteins) and 464 spectra (81 proteins), from SC* and ST* data sets, respectively. More than 91% of the spectra and more than 83% of the proteins were also reported by MS-Align+. (Fig. 5 and supplemental Table S5), indicating that most of them were correctly identified. In addition, MS-Align+ identified a large number of spectra, especially in ST* data set, that were missed by MASCOT and OMSSA. Fig. 6 illustrates that MS-Align+ outperforms MASCOT and OMSSA and that the performance of MASCOT and OMSSA deteriorates for spectra with precursor masses ≥5K Da.

There were 301 spectra identified by MASCOT or/and OMSSA, but missed by MS-Align+. From SC* and ST* data sets. The simple scoring function of MS-Align+, which is based on peak counting, may be another reason why the spectra were missed by MS-Align+. The distribution of E-values of the spectra reported by MS-Align+ is shown in supplemental Fig. S5.

Comparison with PIITA—We compared PIITA and MS-Align+ on ST* data set (running PIITA on SC* data set resulted in a software error). Using the same target/decoy approach, with spectrum level 1% FDR, PIITA coupled with MS-Deconv identified 1589 spectra (169 proteins) from ST* data set, a significant improvement over Mascot and OMSSA. MS-Align+ and PIITA shared 1524 identified spectra and 144 identified proteins (supplemental Fig. S6 and supplemental Table S5). Interestingly, PIITA found some PrSMs missed by MS-Align+ and vice-versa (supplemental Tables S2 and S3). The possible reasons why PIITA identified some PrSMs missed by MS-Align+ are (1) PIITA considers six ion types: b, b-H₂O b-NH₃, y, y-H₂O and y-NH₃, whereas MS-Align+ considers only b and y-ions; (2) PIITA does not depend on the precursor mass for protein identification, whereas MS-Align+ uses the precursor mass to convert y-ion peaks to PRM peaks. Therefore, PIITA is more sensitive when (1) the spectrum contains many peaks from ions with neutral losses or (2) the precursor mass is incorrect. PIITA identified 434 PrSMs from ST* data set which were missed by MS-Align+. For each PrSM, PIITA reports only the unmodified form, not the modified form, of the identified protein. Therefore, we can not compare the precursor mass of the spectrum with the molecular mass of the modified form of the protein. In the 434 PrSMs, no precursor masses can be matched to the molecular masses of the unmodified forms of the proteins within 0.5 Da (supplemental Table S3). On the other hand, PIITA can not identify PrSMs with both N- and C-terminal PTMs and may fail to identify PrSMs with internal PTMs. Therefore, MS-Align+ identified many PrSMs missed by PIITA.

Comparison with ProSightPC—ProSightPC 2.0 was used in the comparison. Target and shuffled protein databases required by ProSightPC were generated from the same FASTA sequences for other experiments using the default parameters of ProSightPC. Coupling with MS-Deconv, two search modes of ProSightPC: absolute mass search and (more powerful) biomarker search were used to analyze SC* and ST* data sets. The running time of ProSightPC for processing SC* and ST* data sets using biomarker search on a PC with a 2.16 GHz CPU and 3 GB memory was 890 and 199 min, respectively. The running time for each run of absolute mass search was less than 90 min. With spectrum level 1% FDR, ProSightPC identified 123 and 2230 spectra from SC* data set with absolute mass search and biomarker search, respec-
A total of 2299 (175 proteins) spectra were reported by combining the results from the two search modes (Fig. 7 and supplemental Table S5). MS-Align+ identified all but 10 of the spectra (Fig. 7A). Using the same approach, ProSightPC identified 2230 spectra (158 proteins) from ST* data sets (absolute mass search reported 2083 spectra and biomarker search reported 892 spectra). ST* data set is somehow “simpler” than SC* data set because it contains fewer truncated and modified proteins. MS-Align+ identified 1744 (78.2%) of the spectra identified by ProSightPC (Fig. 7B).

ProSightPC has various search modes and parameters. The user can increase its search space by setting a very wide window for precursor ions or using an annotated protein database containing many forms of a protein. With most advanced search modes, the running time of ProSightPC increases dramatically. For example, the “biomarker

Fig. 6. Distributions of precursor masses of the spectra in SC* and ST* data sets and of the spectra identified by Ms-Align+, MASCOT, OMSSA, PIITA, and ProSightPC. A, SC* data set. B, ST* data set.
Protein Identification Using Top-Down

**Fig. 7.** Venn diagrams of the spectra and protein species identified by MS-Align+ and ProSightPC. A, Spectra identified in SC* data set. ProSightPC shares 2289 (99.6%) spectra with MS-Align+. B, Protein species identified in SC* data set. ProSightPC shares 813 (89.3%) protein species with MS-Align+. C, Spectra identified in ST* data set. ProSightPC shares 1744 (78.2%) spectra with MS-Align+. D, Protein species identified in ST* data set. ProSightPC shares 403 (70.1%) protein species with MS-Align+.

First, we tested ProSightPC using annotated databases. We downloaded a standard annotated top-down database of yeast from https://prosightpm2.northwestern.edu/. The annotated database contains 912,107 protein forms, about 32 times larger than the unannotated database with 28,238 protein forms generated from the FASTA sequences. The running time of ProSightPC for searching SC* data set against the annotated database was about 4 h and 160 h for absolute mass search and biomarker search (Δm option was not used for biomarker search), respectively. Using the same E-value cutoffs as in the experiment with the unannotated database, ProSightPC identified 2310 spectra from 181 proteins (326 spectra from absolute mass search and 2125 spectra from biomarker search). MS-Align+ identified all but 16 of the spectra in about 12 h, a 13 times speed-up compared with ProSightPC. We also downloaded a SWISSPROT flat file of yeast S288C from the SWISSPROT database. By importing the flat file to ProSightPC with its default parameters, we generated an annotated database with 6629 proteins and 4,437,468 protein forms, which is about 158 times larger than the unannotated database. We only tested absolute mass search for this database, which took about 8 h. Using the same E-value cutoff as in the experiment with the annotated database, ProSightPC identified 157 spectra from 11 proteins, of which MS-Align+ identified 148 spectra. Compared with the unannotated database, ProSightPC coupled with the annotated databases is slower and does not significantly improve the number of identifications. The reason may be that SC* data set contains many truncated and modified proteins, which are not included the annotated database.

Second, we tested the advanced search modes with a large search window for precursor ions. Because the running time for these parameter settings was estimated to be very large (77 days), a smaller database of 500 yeast proteins was used instead of the complete yeast database with 5885 proteins. The small database was composed of 305 proteins identified by MS-Align+ or ProSightPC (supplemental Table S5) and 195 randomly selected proteins. Using the small database, the running time of ProSightPC is estimated to be 11 times smaller as compared with the complete database. We changed the precursor ion error tolerance from 1999 Da to 600K Da for absolute mass search (other parameters were set the same as in the previous experiment). With this parameter setting, no proteins were filtered out by precursor ion error. The running time of ProSightPC was about 10 h (estimated running time for the complete database is about 110 h). With 1% spectrum level FDR, 1217 PrSMs (122 proteins) were identified. We also changed the precursor ion error tolerance from 15 ppm to 100 Da for biomarker search. The running time of ProSightPC was about 7 days (estimated running time for the complete database is about 77 days). With 1% spectrum level FDR, 2526 PrSMs (174 proteins) were identified. By combining the results of the two search modes, ProSightPC identified 2822 spectra from 198 proteins. Using the same small database, MS-Align+ identified 4448 spectra from 298 proteins in 3 h. ProSightPC shared 2654 (94%) identified PrSMs with MS-Align+. MS-Align+ identified about 57% more PrSMs while being 56 times faster. We point out that this approach is not optimal and is chosen only to bypass the extreme time requirements of ProSightPC. Indeed, when a small database is populated mainly with identified proteins, the FDR estimates become biased. Our goal is to roughly compare the relative performance of the two tools rather than to estimate the absolute number of proteins these tools can identify. We emphasize that some advanced modes of ProSightPC, although increasing the number of identified PrSMs, may be extremely time-consuming. For example, the biomarker search mode with a precursor ion error tolerance of 100 Da is estimated to take more than 150 days to search SC* data set against both the target and decoy protein databases. The reason why MS-Align+ missed some spectra identified by ProSightPC is that the filtering step of MS-Align+ may miss some possible PrSMs. ProSightPC also failed to report many spectra identified by MS-Align+, especially for SC* data set. Among the 4059 PrSMs identified by MS-Align+ from
SC* data set, 3778 PrSMs are from truncated proteins (supplemental Table S1). Because most of the spectra in SC* data set are truncated proteins, absolute mass search of ProSightPC may fail to identify the corresponding PrSMs (see Fig. 6A). This observation suggests that MS-Align+ and ProSightPC may complement each other in top-down searches.

Coupling MS-Align+ and ProSightPC with Thrash—MS-Align+ is usually coupled with MS-Deconv and is optimized for the frequent ±1 Da shift errors that are common in MS-Deconv. ProSightPC, on the other hand, is coupled with Thrash. To analyze the performance of MS-Align+ decoupling from MS-Deconv, we coupled MS-Align+ and ProSightPC with Thrash instead. The general conclusion is that switching from Thrash to Ms-Deconv improves the performance of both MS-Align+ and ProSightPC (similar observation was made in (31)).

Using its default parameters, Thrash reported 5136 and 2097 merged deconvoluted spectra with precursor mass ≥ 2500 Da and at least 10 peaks from the SC and ST data sets, respectively. Using the same target/decoy approach, with spectrum level 1% FDR, MS-Align+ (coupled with Thrash) identified 2031 and 934 spectra from the merged spectra from the SC and ST data sets, respectively. By combining the results of absolute mass search and biomarker search, ProSightPC identified 617 spectra (61 spectra from absolute mass search and 562 spectra from biomarker search) from the SC data set, and 900 spectra (815 spectra from absolute mass search and 418 spectra from biomarker search) from the ST data set. MS-Align+ and ProSightPC shared 598 and 707 identified spectra in the SC and ST data sets, respectively.

Because MS-Deconv recovered single spectra and Thrash recovered merged spectra, we mapped the single spectra to the merged spectra and compared the identified single spectra and the identified merged spectra. For MS-Align+, a total of 2592 (87.4%) identified merged spectra can be mapped to identified single spectra, and 4343 (69.2%) identified single spectra can be mapped to identified merged spectra. For ProSightPC, 1319 (86.9%) identified merged spectra can be mapped to identified single spectra, and 2295 (50.6%) identified single spectra can be mapped to identified merged spectra (supplemental Table S4). The reason might be that MS-Deconv performed better than Thrash for single spectra, and that Thrash performed better than MS-Deconv when several spectra can be merged to report a single deconvoluted spectrum.

Validation of E-Values Reported by MS-Align+ and ProSightPC—Both MS-Align+ and ProSightPC report an E-value for each identified PrSM, which is used for estimating the statistical significance of the PrSM. The E-value is computed based on an estimated p value of the PrSM and the size of the protein database. In difference from ProSightPC, which uses the number of proteins as the size of the database, MS-Align+ uses the number of proteins with the same mass (with error tolerance) as the precursor mass of the spectrum as the size of the database. Although the size of a protein database is easy to compute, the estimation of p value is a difficult problem. For example, the p values reported by some bottom-up database search tools may differ from the textbook definition of the notion of p value by orders of magnitude (30). To test if the p values reported by MS-Align+ and ProSightPC are accurate, we estimate the theoretical p values by searching the spectra against a giant random protein database as described in Kim et al., 2008 (30). Suppose a PrSM between spectrum S and protein P has no PTMs and its score is t. We generate a random protein database with 10^6 proteins, each of which has a molecular mass similar to P. Then we search S against the random protein database and count the number a of PrSMs with score ≥ t. Thus, the textbook estimate of the theoretical p value is a × 10^–6. The accuracy of reported p values is examined by comparing the reported p values and the estimated theoretical p values.

MS-Align+ used a generating function approach to compute p values, and reported 1348 PrSMs with p value ≤ 10^–4 without PTMs in ST* data set. We computed the estimated theoretical p value for each PrSM by searching a giant random protein database with 1 million proteins. Even with such a giant protein database, one is unable to reliably estimate theoretical p values in the cases of p values below 10^–6 (because one expects to find less than a single database hit in these cases, often resulting in the estimated p value equal to 0). Thus, we had to limit our attention to only 144 PrSMs (out of 1348) with a reported p value in the interval [10^–6, 10^–4] and an estimated theoretical p value ≥ 10^–6. We computed the ratios between the reported p values and the estimated theoretical p values for these 144 PrSMs (Fig. 8). In the case of accurate p values one expects to see log-ratios of theoretical to reported p values equal to 0. Although the histogram in Fig. 8 shows a peak at interval [−0.5, 0.5] (close match between theoretical and reported p-values), 10.4% of the PrSMs have...
Protein Identification Using Top-Down

Table II
Comparison of p-values reported by MS-Align+ and ProSightPC and theoretical p-values estimated by searching against a giant random protein database. MS-Align+ reported more matched peaks than ProSightPC because MS-Align+ treats an observed mass and a theoretical mass with ± 1 Da shift as a matched mass pair whereas ProSightPC does not. The average log-ratio (base 10) between the reported and the estimated theoretical p-values is 0.17 for MS-Align+ and 1.20 for ProSightPC. In the worst case, the log-ratio between the reported p-value and the theoretical p-value is nearly 0.61 for MS-Align+ (Spectrum 9) and nearly 2.09 for ProSightPC (Spectrum 5).

| Spectrum | Precursor mass | # peaks | MS-Align+ | ProSightPC |
|----------|----------------|---------|-----------|------------|
|          |                |         | # matched fragments | Reported p-value | Estimated theoretical p-value | # matched fragments | Reported p-value | Estimated theoretical p-value |
| 1        | 5377.08        | 89      | 7         | 9.4E-5     | 1.05E-4           | 8               | 1.00E-7     | 2.00E-6           |
| 2        | 11206.33       | 107     | 11        | 8.28E-5    | 9.50E-5           | 8               | 5.77E-5     | 4.00E-6           |
| 3        | 8113.32        | 53      | 6         | 4.21E-4    | 4.89E-4           | 5               | 1.16E-4     | 5.00E-6           |
| 4        | 9044.94        | 135     | 13        | 1.36E-6    | 2.00E-6           | 9               | 9.70E-6     | 1.00E-6           |
| 5        | 8970.92        | 149     | 12        | 3.58E-6    | 2.00E-6           | 8               | 1.23E-4     | 1.00E-6           |
| 6        | 8420.76        | 83      | 7         | 1.12E-4    | 7.50E-5           | 5               | 1.02E-3     | 1.31E-4           |
| 7        | 11546.28       | 67      | 7         | 3.61E-5    | 4.30E-5           | 4               | 1.01E-2     | 1.22E-4           |
| 8        | 11206.32       | 33      | 7         | 3.86E-5    | 4.60E-5           | 5               | 5.72E-5     | 3.00E-6           |
| 9        | 10060.49       | 29      | 6         | 1.24E-5    | 5.00E-5           | 5               | 1.89E-5     | 6.00E-6           |
| 10       | 6378.58        | 17      | 7         | 7.74E-5    | 1.26E-4           | 4               | 6.40E-6     | 2.00E-6           |

MS-Align+ identified 31 proteins without NME and 88 proteins with NME, including seven proteins that did not follow the NME specificity rule (34).

Signal Peptides—MS-Align+ identified 28 proteins from ST* data set with a short truncated prefix (the truncated prefix had 15–35 residues). Although 14 proteins had a truncated prefix with the canonical “AXA” motif for signal peptides, the other 14 proteins did not. For the 14 protein without AXA motif, SignalP (36) reported 12 signal peptides (supplemental Table S7), out of which nine signal peptides matched to the truncated prefixes reported by MS-Align+, and the other three signal peptides did not.

Erroneous Annotations of Translation Start Sites—MS-Align+ found in ST* data set 12 protein species (12 proteins) with a truncated prefix that either ends in “M” or exposes M as the first amino acid in the truncated protein (Table III), including eight proteins with short truncated prefixes (from 2 to 9 residues) typical for mis-annotated bacterial start sites (37). However, the remaining 4 proteins had long truncated prefixes (over 80 amino acids) pointing to potential alternative start sites (38, 39). All start sites but one in Table III follow the NME rule, thus reinforcing our conclusion that the translation start sites in these proteins represent either mis-annotated or alternative start sites.

PTMs on Internal Residues—MS-Align+ identified many proteins with common PTMs, e.g. proteins with oxidation (+16 Da) and methylation (+14 Da), and several proteins with uncommon PTMs. In SC* data set, MS-Align+ reported proteins with a mass shift of +38 Da and proteins with a mass shift of +183 Da (the protease inhibitor mixture used in preparing the sample may contain 4-(2-Aminoethyl)benzenesulfonyl fluoride hydrochloride that introduces this modification. See http://www.unimod.org/modifications_view.php?editid1=276). In ST* data set, MS-Align+ reported proteins with Carboxamidomethyl DTT on cysteine (the cysteine modification is +209 Da and the modification relative to a carboxamidomethyl-
lated cysteine is +152 Da), proteins with a disulfide bond on two cysteines (−116 Da compared with two cysteines with carbamidomethylation), and proteins with a −12 Da mass shift (one possible explanation is replacing an isoleucine with a threonine).

Searching Six-Frame Translation—We also generated the six-frame translation of the ST genome and searched the top-down spectra in ST* data set against the database. Similar searches have recently been done using ProSightPC (40). By searching the (unannotated) six-frame translation, MS-Align was able to identify nearly all proteins identified in searches of the (annotated) proteome thus emphasizing potential proteogenomics applications of top-down mass spectrometry. Moreover, MS-Align identified a protein species of NP_462331.2 which contains four more amino acids than the annotated protein sequence at the N terminus (Fig. 9), the likely indication of an annotation error.

Analyzing Unidentified Spectra: Evidence of Room for Improvement—Highly accurate precursor mass measurements of top-down spectra allow one to consider a hypothesis that a protein matches to an unidentified spectrum based only on the precursor mass. Indeed, as supplemental Fig. S7A shows, a surprisingly large number of spectra have precursor masses either equal to the theoretical precursor masses of proteins in the ST proteome (118 spectra) or differ from these masses by 131 Da (89 spectra). The peaks at 0 (matching precursor and theoretical masses) and at 131 (matching precursor and theoretical mass minus NME) are the tallest peaks in supplemental Fig. S7A whereas the noise level corresponds to ~25 spectra. Thus, we estimate that roughly 118 + 89 - 25 - 25 =

### Table III

Proteins identified from ST* data set with a truncated prefix that either ends in “M” or exposes “M” as the first amino acid in the truncated protein. The first 10 amino acids in the identified protein species and the truncated prefixes are shown in the third column. The symbol “-” indicates the N-terminus of the protein.

| Protein Length of truncated prefix | First 10 amino acids |
|-----------------------------------|----------------------|
| NP_461253.1 putative inner membrane protein 3 | -MRM.SYQFGESRVD |
| NP_459378.1 putative cytoplasmic protein 2 | -MV.MKNLIAELL |
| NP_460769.1 putative cytoplasmic protein 13 | VTIPLIQSKS.MFCVIYRSSK |
| NP_459993.1 hypothetical protein STM018[Phage Gifsy-2] 7 | -MNYEDKL.MYQMFIEEG |
| NP_462709.2 small heat shock protein 2 | -MI.MRNFDSLPLY |
| NP_460044.1 putative hydrogenase membrane component precursor 304 | PSDKISARGM.GESNPVTGNT |
| NP_460424.2 putative cytoplasmic protein 5 | -MDQLY.MTVQDYLLK |
| NP_490541.1 putative cytoplasmic protein 245 | RVPQHGDA.MTIAQQLEQKG |
| NP_459230 outer membrane protein H precursor 87 | SDRTKLEKD.VSQRQTF |
| NP_462207.1 ATP-dependent zinc protease 573 | LTDNMDIL.HAKMDKMY |
| NP_461361.1 putative cytoplasmic protein 3 | -MPE.MKEQLVEIA |
| NP_459322.1 putative cytoplasmic protein 9 | -MVNFDKS.MPTAI |

Fig. 9. A protein species identified by MS-Align via the search against the six frame translation of the ST genome. In the reported protein species, four residues “AKQS” (underlined) at the N terminus precede the annotated start site and suggest an annotation error.
157 unidentified spectra can potentially form PrSMs with proteins matching their precursor masses. Similar arguments based on supplemental Fig. S7 illustrate that these spectra may correspond to roughly 85 + 74 - 21 - 21 = 117 protein species. Some of these protein species are identified by PIITA/ProSightPC and missed by MS-Align+. These observations illustrate that more elaborate scoring functions may be needed for top-down MS/MS database searches.

CONCLUSIONS

Our analysis demonstrated that MS-Align+ favorably compares with other tools in identifying PTMs, particularly in the blind mode. In addition, it is fast and it computes more accurate p values than ProSightPC. As a result, MS-Align+ reported many new protein identifications (e.g. proteins with N- and C-terminal truncations and proteins with internal PTMs) from the SC and ST data sets.

There are several computational challenges in top-down mass spectrometry that require further studies. For example, a faster and sensitive filtering method is needed for speeding up the protein identification without significant loss of sensitivity. In MS-Align+, multiple PTMs can be identified only if there are several peaks in the spectrum supporting each PTM. Such peaks may be missing in some spectra thus raising a problem of identifying multiple PTMs with a limited number of matched peaks. Another important problem is to develop computational tools for automatically determining and localizing PTMs.

Acknowledgments—We thank Dr. Julian Whitelegge and Dr. Puneet Souda for allowing us to use their ProSightPC server.

REFERENCES

1. Loo, J. A., Edmonds, C. G., and Smith, R. D. (1990) Primary sequence information from intact proteins by electrospray ionization tandem mass spectrometry. Science 248, 201–204
2. Reid, G. E., and McLuckey, S. A. (2002) 'Top down' protein characterization via tandem mass spectrometry. J. Mass Spectrometry 37, 665–675
3. Sore, S. K., Ge, Y., Oh, H., and McLafferty, F. W. (2002) Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. Proc. Natl. Acad. Sci. U. S. A. 99, 1774–1779
4. Dorrestein, P. C., Zhai, H., Taylor, S. V., McLafferty, F. W., and Begley, T. P. (2004) The biosynthesis of the thiazole phosphate moiety of thiamin (vitamin B1): The early steps catalyzed by thiazole synthase. J. Am. Chem. Soc. 126, 3091–3096
5. Whitelegge, J., Halgand, F., Souda, P., and Zabrouskov, V. (2006) Top-down mass spectrometry of integral membrane proteins. Expert Rev. Proteomics 3, 585–596
6. Dorrestein, P. C., Van Lanen, S. G., Li, W., Zhao, C., Deng, Z., Shen, B., and Kelleher, N. L. (2006) The bifunctional glycerol transferase/phosphatase OzM belong to the HAD superfamily that diverts 1,3-bisphosphoglycerate into polyketide biosynthesis. J. Am. Chem. Soc. 128, 10386–10387
7. McLafferty, F. W., Breuker, K., Jin, M., Han, X., Infusini, G., Jiang, H., Kong, X., and Begley, T. P. (2007) Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. FEBS J. 274, 6256–6268
8. Suti, N., and Kelleher, N. L. (2007) Decoding protein modifications using top-down mass spectrometry. J. Proteome Res. 6, 1890–1897
9. Whitelegge, J. P., Zabrouskov, V., Halgand, F., Souda, P., Bassilian, S., Yan, W., Wolinsky, L., Loo, J. A., Wong, D. T., and Faull, K. F. (2007) 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. 268, 190–197
10. Zabrouskov, V., and Whitelegge, J. P. (2007) Increased coverage in the transmembrane domain with activated-ion electron capture dissociation for top-down Fourier-transform mass spectrometry of integral membrane proteins. J. Proteome Res. 6, 2205–2210
11. Garcia, B. A. (2010) What does the future hold for top down mass spectrometry? J. Am. Soc. Mass Spectrom. 21, 193–202
12. Meng, F., Cargile, B. J., Patrie, S. M., Johnson, J. R., McLoughlin, S. M., and Kelleher, N. L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. Anal. Chem. 74, 2923–2929
13. Meng, F., Du, Y., Miller, L. M., Patrie, S. M., Robinson, D. E., and Kelleher, N. L. (2004) Molecular-level description of proteins from saccharomyces cerevisiae using quadrupole FT hybrid mass spectrometry for top down proteomics. Anal. Chem. 76, 2852–2858
14. Patrie, S. M., Ferguson, J. T., Robinson, D. E., Whipple, D., Rother, M., Metcalf, W. W., and Kelleher, N. L. (2006) Top down mass spectrometry of < 50-kDa proteins from Methanosarcina acetivorans using quadrupole FTMS with automated octopole collisionally activated dissociation. Mol. Cell. Proteomics 5, 14–25
15. Sharma, S., Simpson, D. C., Tolic, N., Jaitly, N., Mayampurath, A. M., Smith, R. D., and Pasa-Tolici, L. (2007) Proteomic profiling of intact proteins using WAX-RPLC 2-D separations and FTICR mass spectrometry. J. Proteome Res. 6, 602–610
16. Shen, Y., Hixson, K. K., Camp, D. G., Purvine, S. O., Moore, R. J., and Smith, R. D. (2008) Mass spectrometry analysis of proteome-wide proteolytic post-translational degradation of proteins. Anal. Chem. 80, 5819–5828
17. Shen, Y., Tolic, N., Hixson, K. K., Purvine, S. O., Anderson, G. A., and Smith, R. D. (2008) De novo sequencing of unique sequence tags for discovery of post-translational modifications of proteins. Anal. Chem. 80, 7742–7754
18. Wynne, C., Fenselau, C., Demirev, P. A., and Edwards, N. (2009) Top-down identification of protein biomarkers in bacteria with unsequenced genomes. Anal. Chem. 81, 9633–9642
19. Tsai, Y. S., Scherl, A., Shaw, J. L., MacKay, C. L., Shaffer, S. A., Langridge-Smith, P. R., and Goodlett, D. R. (2009) Precursor ion independent algorithm for top-down shotgun proteomics. J. Am. Soc. Mass Spectrom. 20, 2154–2166
20. Vellaichamy, A., Tran, J. C., Catherman, A. D., Lee, J. E., Kellie, J. F., Sweet, S. M. M., Zamdborg, L., Thomas, P. M., Ahlf, D. R., Durbin, K. R., Valaskovic, G. A., and Kelleher, N. L. (2010) Size-sorting combined with improved nanocapillary liquid chromatography-mass spectrometry for identification of intact proteins up to 80 kDa. Anal. Chem. 82, 1234–1244
21. Roth, M. J., Parks, B. A., Ferguson, J. T., Boyne, M. T., and Kelleher, N. L. (2008) "Proteotyping": population proteomics of human leukocytes using top-down mass spectrometry. Anal. Chem. 80, 2857–2866
22. LeDuc, R. D., Taylor, G. K., Kim, Y. B., Januszyk, E. T., Bynum, L. H., Sola, J. V., Garavelli, J. S., and Kelleher, N. L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Res. 32, W340-W345
23. Zamdborg, L., LeDuc, R. D., Glowacz, K. J., Kim, Y. B., Viswanathan, V.,
Spaulding, I. T., Early, B. P., Bluhm, E. J., Babai, S., and Kelleher, N. L. (2007) ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **35**, W701-W706

24. Frank, A. M., Pesavento, J. J., Mizzen, C. A., Kelleher, N. L., and Pevzner, P. A. (2008) Interpreting top-down mass spectra using spectral alignment. *Anal. Chem.* **80**, 2499–2505

25. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562–1567

26. Karabacak, N. M., Li, L., Tiwari, A., Hayward, L. J., Hong, P., Easterling, M. L., and Agar, J. N. (2009) Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Mol. Cell. Proteomics* **8**, 846–856

27. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964

28. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumbery, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77**, 4626–4639

29. Pevzner, P. A., Dancik, V., and Tang, C. L. (2000) Mutation-tolerant protein identification by mass spectrometry. *J. Computational Biol.* **7**, 777–787

30. Kim, S., Gupta, N., and Pevzner, P. A. (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **7**, 3354–3363

31. Liu, X., Inbar, Y., Dorrestein, P. C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J. P., Bafna, V., and Pevzner, P. A. (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins: A combinatorial approach. *Mol. Cell. Proteomics* **9**, 2772–2782

32. Elias, J. E., and Gygi, S. P. (2010) Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **604**, 55–71

33. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567

34. Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., and Meinnel, T. (2006) The proteomics of N-terminal methionine cleavage. *Mol. Cell. Proteomics* **5**, 2336–2349

35. Gupta, N., Benhamida, J., Bhargava, V., Goodman, D., Kain, E., Kereman, I., Nguyen, N., Ollikainen, N., Rodriguez, J., Wang, J., Lipton, M. S., Romine, M., Bafna, V., Smith, R. D., and Pevzner, P. A. (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Gen. Res.* **18**, 1133–1142

36. Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**, 953–971

37. Gupta, N., Tanner, S., Jaitly, N., Adkins, J. N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R. D., and Pevzner, P. A. (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Gen. Res.* **17**, 1362–1377

38. Touriol, C., Bornes, S., Bonnal, S., Audigier, S., Prats, H., Prats, A.-C., and Vagner, S. (2003) Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol. Cell* **95**, 169–178

39. Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamanoto, T., and Sugano, S. (2007) Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol. Cell. Proteomics* **6**, 1000–1006

40. Ferguson, J. T., Wenger, C. D., Metcalf, W. W., and Kelleher, N. L. (2009) Top-down proteomics reveals novel protein forms expressed in *Methanosarcina acetivorans*. *J. Am. Soc. Mass Spectrom.* **20**, 1743–1750

In order to cite this article properly, please include all of the following information: Liu, X., Sirotkin, Y., Shen, Y., Anderson, G., Tsai, Y. S., Ting, Y. S., Goodlett, D. R., Smith, R. D., Bafna, V., and Pevzner, P. A. (2012) Protein Identification Using Top-Down. *Mol. Cell. Proteomics* **11**(6):M111.008524. DOI: 10.1074/mcp.M111.008524.