Top-down and bottom-up mass spectrometry methods can generate gas phase fragments and use these to identify proteins. Top-down methods, in addition, can provide the mass of the protein itself and therefore additional structural information. Despite the conceptual advantage of top-down methods, the market share advantage belongs to bottom-up methods as a result of their more robust sample preparation, fragmentation, and data processing methods. Here we report improved fragmentation and data processing methods for top-down mass spectrometry. Specifically we report the use of funnel-skimmer dissociation, a variation of nozzle-skimmer dissociation, and compare its performance with electron capture dissociation. We also debut BIG Mascot, an extended version of Mascot with incorporated top-down MS² search ability and the first search engine that can perform both bottom-up and top-down searches. Using BIG Mascot, we demonstrated the ability to identify proteins 1) using only intact protein MS¹, 2) using only MS², and 3) using the combination of MS¹ and MS². We correctly identified proteins with a wide range of masses, including 13 amyotrophic lateral sclerosis-associated variants of the protein Cu/Zn-superoxide dismutase, and extended the upper mass limit of top-down protein identification to 669 kDa by identifying thyroglobulin.  

Molecular & Cellular Proteomics 8:846–856, 2009.

Ionization and detection of intact proteins have been possible for 20 years (1, 2). This breakthrough, combined with the development of methods of intact protein fragmentation (3–9) and database searching (10–12), gave rise to top-down mass spectrometric approaches of protein analysis (13–19). Although top-down proteomics excels in the area of protein variant and post-translational modification (PTM)¹ characterization (14, 20, 21), comprehensive protein identification is an area for improvement (22–24). Top-down proteomics is hampered by the lack of a sample preparation technique as generally applicable as in-gel digestion (25), the lack of a fragmentation technique as robust as collisionally activated dissociation (CAD) of peptides (26), and the lack of compatibility of top-down MS data with all popular bottom-up software (27), including Mascot (28), Sequest (29), Protein Prospector (30), and Peptide Search (31). Recent improvements in these areas include the following: 1) separation techniques compatible with top-down analysis (18, 32); 2) fragmentation techniques, including nozzle-skimmer dissociation (3, 33–35), electron transfer dissociation (7, 8), prefolding dissociation (9), and activated ion electron capture dissociation (ECD) (5), which extended the protein upper mass limit to 229 kDa for top-down methods (9); and 3) a search engine as advanced as any bottom-up software; ProSight PTM (11, 12, 36) was developed and integrated with a comprehensive MS database within the Xcalibur software suite (Thermo Finnigan, San José, CA).

Our objectives here are (i) to help develop an extended version of Mascot that incorporates top-down MS² search ability and (ii) to infer its search sensitivity and selectivity. Here we present BIG Mascot, the first general use search engine that can perform both bottom-up and top-down MS² searches. Mascot has been one of the most widely used database search engines because of its probability-based scoring method (28), its superior performance (37, 38), and its fast operating algorithm. However, the applicability of Mascot to top-down work flows is severely limited as the commercially available version has a precursor ion mass limit of 16 kDa. Using a version of Mascot referred to as BIG Mascot that was modified to allow precursor ions of up to 110 kDa, we demonstrate that either intact protein MS¹ or MS² by themselves yield sufficient information for identification of many proteins. To identify variants of a single cSNP, coding single nucleotide polymorphism; CAD, collisionally activated dissociation; ECD, electron capture dissociation; SOD1, Cu/Zn-superoxide dismutase; ALS, amyotrophic lateral sclerosis; SORI, sustained off-resonance irradiation; MSDB, Mass Spectrometry Protein Sequence Database; SNAP, Sophisticated Numerical Annotation Procedure; IPI, International Protein Index.

From the ‡Department of Chemistry and Volen Center for Complex Systems and §Department of Computer Science, National Center of Behavioral Genomics, Brandeis University, Waltham, Massachusetts 02454, ¶University of Massachusetts Medical School, Worcester, Massachusetts 01655, and **Bruker Daltonics Inc., Billerica, Massachusetts 01821

Received, March 6, 2008, and in revised form, December 8, 2008 Published, MCP Papers in Press, December 15, 2008, DOI 10.1074/mcp.M800099-MCP200

¹ The abbreviations used are: PTM, post-translational modification; MS¹, precursor ion mass spectrum; MS², product ion mass spectrum;
Identification of Proteins Using Top-down Mass Spectrometry

protein, including mutations or PTMs, however, the masses of both the intact protein and the fragments are required. Finally by identifying the protein thyroglobulin using a single step analysis, we extend the mass of intact protein identification beyond the previous limit of 229 kDa (9) to 669 kDa.

EXPERIMENTAL PROCEDURES

Sample Preparation—Amyotrophic lateral sclerosis (ALS)-related (39, 40) human Cu/Zn-superoxide dismutase (SOD1) variants (A4V, L38V, G41S, G72S, G76Y, G85R, D90A, G93A, D124V, D125H, E133del, and S134N) were isolated and purified from baculovirus-transfected Sf21 insect cells as described previously (41). Storage solutions were 100 μg of protein in 100 μl of 20 mM potassium phosphate, pH 7.2, 150 mM potassium chloride. SOD1 intrasubunit binding to the ZipTip, it was judged to be disulfide bonds were reduced in 50 mM tris(2-carboxyethyl)phosphine, pH 7.2, 150 mM potassium chloride. SOD1 intrasubunit disulfide bonds were reduced in 50 mM tris(2-carboxyethyl)phosphine, 5 mM EDTA, and 5 μl of protein was desalted using a C18 ZipTip (Millipore, Billerica, MA). Wetting (acetonitrile; two repeats), equilibration (0.1% formic acid; two repeats), binding, washing (HPLC water; two repeats), and elution (30–40 μl of 50% acetonitrile, 0.1% formic acid; one repeat) steps were performed according to the manufacturer’s instructions except that the initial wash utilized 10 mM ammonium bicarbonate, pH 8.0, to remove any nonspecifically bound metal ions (i.e. potassium). The concentration of SOD1 variants was nominally 10 μM following this procedure, but because of incomplete binding to the ZipTip, it was judged to be ~2 μM. Following the ZipTip procedure, SOD1 variants were found to be N-terminally acetylated and without any metal.

BSA (66 kDa), bovine thyroglobulin (669 kDa as a dimer under native conditions (42–45)), and equine myoglobin (17 kDa) were purchased from Sigma-Aldrich. UCHEL-1 and DJ-1 were expressed and purified as described elsewhere (for UCHEL-1, see Ref. 46; for DJ-1, see Ref. 47). Proteins were desalted using an Amicon Ultra-4 centrifugal membrane filter (Millipore) with 5-kDa nominal molecular mass limit (SOD1 variants were found to aggregate under these same conditions, hence the difference in desalting methodologies). An initial wash using 10 mM ammonium bicarbonate was performed followed by centrifugation. Three more wash steps were performed using HPLC grade water. Amicon-desalted samples were diluted to a final concentration of 1 μM (ubiquitin and myoglobin) or 10 μM (thyroglobulin) in 50% acetonitrile, 0.1% formic acid. A separate thyroglobulin aliquot was diluted in 10 mM ammonium acetate (to maintain dimeric quaternary structure), pH 6.8, for additional analysis.

MS1 and MS2 Analyses—Samples were directly infused using a syringe pump and nanosprayed into a dual ion funnel (48) (Apollo II) ion source connected to a hybrid quadrupole FT-ICR mass spectrometer (9.4 teslas, apex Qe-94, Bruker Daltonics Inc., Billerica, MA). The capillary temperature was set at 160 °C. The instrument is equipped with a hollow cathode (49) for ECD. External calibration of masses was determined from deisotoped and reconstructed mass spectra (Data Analysis, Bruker Daltonics Inc.). For fragmentation, funnel-skimmer dissociation or ECD was used. ECD involved a filament bias of 1–4 V and pulse duration of 2–4 ms. What we have coined as funnel-skimmer dissociation is analogous to nozzleskimmer dissociation (51) and involved modulating the source declustering potential. Because the ambient pressure in the first stage of the dual stage funnel is ~3 torr, the momentum necessary to collisionally activate ions exiting the source capillary cannot be achieved, thereby preventing CAD at the location analogous to the traditional “nozzle.” Instead the voltage of the ion optic labeled skimmer 1 is raised significantly to accelerate ions into the second stage of the ion funnel stack where the pressure is a few millitorrs. Although this effectively creates a shallower direct current voltage gradient across the first stage of the radio frequency funnel, no impact on sensitivity is observed. Performing CAD in the second funnel stage is advantageous because the funnel has a wide acceptance angle, which mitigates scattering effects. Thus we favor the more descriptive term funnel-skimmer dissociation because dissociation takes place in the absence of a nozzle and in the presence of an ion funnel. The voltage of skimmer 1 was increased 2-fold (80 V) or more (120 V for BSA and 100 V for thyroglobulin). Skimmer 1 voltages of 40 and 80 V correspond to 3.5 (skimmer 1 voltage – funnel 2 voltage) of 33 and 83 V under our conditions. Ion optics voltages downstream of the skimmer 1 electrode were not modified for these experiments.

Searches—BIG Mascot was used for all database searches described. Its graphical user interface and options are identical to that of Mascot. BIG Mascot benefits from a precursor ion mass limit of 110,000 Da instead of 16,000 Da. This feature enabled us to submit the intact protein mass in the place of precursor mass in Mascot generic files. As a result of a BIG Mascot search, the user obtains one “peptide” identified per protein with a sequence coverage of 100%. In this study we accepted the highest ranking protein as the BIG Mascot identification of our “unknown.” We did not filter our results according to significance thresholds reported by BIG Mascot.

Because of the vestiges of the Mascot graphical user interface, BIG Mascot requires that an enzyme be specified. The enzyme definition encodes a set of operations that will be performed in a sequence-specific manner, e.g.,“trypsin” will cleave a primary sequence following Arg and Lys residues. We therefore developed an enzyme definition named “IntactProtein” that essentially does nothing other than cleave the N-terminal methionine. Specifically we altered the enzyme definitions file of the BIG Mascot server to include the following code: “Cleave: M, Restrict: ABCDEFGHIJKLMNOPQRSTUVWXYZ, Cterm.” We report significance thresholds from searches with IntactProtein because our proteins were intact.

For searches in which either the protein is larger than 110,000 Da, the intact protein mass is not detected, or the protein is modified in a way that is not accounted for in the database, we used a search we term MS2-only search. Normally BIG Mascot requires a precursor ion mass input. We bypassed this requirement by specifying a nominal 10,000–Da precursor mass and ±1% (maximum) error tolerance and selecting “None” as the enzyme specificity. This setup allows an MS2-only search of all sequences and subsequences of proteins larger than 9,900 Da (10,000 – 100 Da difference is because of the tolerance window). In other words, the nominal 10-kDa precursor mass specified acts as the lower mass limit of protein (MS2-only search with 10-kDa nominal precursor mass would not identify 8-kDa ubiquitin, thus we used a 5-kDa nominal mass for ubiquitin). In summary, MS2-only searches do not require knowledge of the intact protein mass but a nominal minimum protein mass. Searches using None as the enzyme definition yielded the same scores and matches but much larger significance thresholds for BIG Mascot scores (between 60 and 70 instead of 10–20) because the number of sequences searched increases immensely because of the inclusion of subsequences. Note that MS2-only searches specifying a precursor mass
Manual isotopic peak determination was also performed on all SOD1 high resolution analysis of spectra by Horn) algorithm introduced by nonlinear fit to the data in a similar fashion to the THRASH (thorough case averagine (52)) to calculate isotopic distribution to obtain a uses a given mass, charge, and mean molecular constitution (in our presentation Procedure (SNAP) version 2.0, Bruker Daltonics Inc.). SNAP ance were set to our observed limits of protein identifica-

of 10–110 kDa (the limit of BIG Mascot) yielded thyroglobulin as the first hit; the critical parameter for MS²-only searches is therefore that the specified precursor mass is lower than the database-derived mass.

MSDB (released August 31, 2006) was augmented with sequences of 102 different ALS-associated variants of SOD1 and the sequence of GPLGS-tagged UCHL-1. All database searches utilized monoisotopic masses obtained via automated data reduction of the highly complex FT-ICR MS² spectral data (Sophisticated Numerical Annotation Procedure (SNAP) version 2.0, Bruker Daltonics Inc.). SNAP uses a given mass, charge, and mean molecular constitution (in our case average (52)) to calculate isotopic distribution to obtain a nonlinear fit to the data in a similar fashion to the THRASH (thorough high resolution analysis of spectra by Horn) algorithm introduced by Horn et al. (35). This fit is used to calculate the monoisotopic mass. Manual isotopic peak determination was also performed on all SOD1 (wild type and variants) MS² data for verification. Default error tolerances were set to ±2 Da for MS¹ data and ±0.5 Da for ECD or ±0.2 Da for funnel-skimmer dissociation. Default instrument definition was set as FTMS-ECD for ECD and ESI-FTICR for funnel-skimmer dissociation searches. MS¹ tolerance of ±2 Da does not reflect the accuracy of the instrument (10 ppm or less) but instead accounted for the partial reduction of disulfide bonds in SOD1 that we observed. N-terminal acetylation was specified as a fixed modification for all SOD1 variants.

Following these MS² searches, an additional set of CAD “internal fragment” MS² searches were performed. Instrument fragment type definitions of BIG Mascot were edited to allow for internal fragments of by-type as large as 30 kDa to be searched in addition to b- and y-type fragments. MS² error tolerance was more stringent at 0.05 Da for these internal fragment searches to avoid random matches. To match this stringency internal calibration of the CAD MS² data using six to eight high intensity b or y fragment ions was performed. The m/z values corresponding to the theoretical monoisotopic masses and observed charge state of each fragment were calculated manually from the BIG Mascot fragmentation report of the protein.

Using the same parameters, we performed additional BIG Mascot automatic decoy database searches of randomized sequences. Each random sequence has the same composition and weight as the database sequence. As a result, BIG Mascot displays the number of random sequences matched to data with a score above significance thresholds (identity and homology) and generates a separate report for decoy database search. As we have a small number of spectra, it is not useful to display false discovery rates (for our case the false discovery rate was either 0.0% (no significant random hits) or undefined (BIG Mascot search result has no significant hits)). Nevertheless in this study we report the score of the highest scoring random sequence to compare with the score of the BIG Mascot-identified protein sequence.

RESULTS

Evaluation of the Upper Mass Limit of Top-down MS—A frequently raised criticism of top-down MS is the mass limit imposed by the largest detectable protein mass, which in routine operation of FTMS instruments is 70–100 kDa (53). We therefore calculated the masses of all known proteins in the IPI (54) human database version 3.31 to determine the fraction below this upper mass limit (see Fig. 1). An in-house, manually written software was used to calculate the monoisotopic and average mass of each entry and export entry numbers and corresponding masses as a text file. This file was imported into an Excel spreadsheet and plotted. Fig. 1 illustrates that 80% of human proteins weigh less than 70 kDa and are therefore currently amenable to top-down MS. Moreover the median mass of proteins in the human IPI database is 31.4 kDa, and the highest density region (mode) of the database is around 14.5 kDa. Methods are presented below that identify the protein thyroglobulin, a protein larger than all but 0.6% of all human primary sequences.

Evaluation of the Hypothetical Requirements for the Identification of Proteins Using only MS² Data—To explore the prospect of protein identification using only fragmentation (no intact protein mass), we investigated the required quality of fragmentation data for BIG Mascot protein identification using MS²-only searches. Fig. 2 illustrates that for the proteins that we analyzed a single sequence tag with seven sequential fragments was sufficient to identify a protein when the mass accuracy of analysis is 100 ppm, whereas at 1 ppm accuracy four or five sequential fragments were sufficient. We observed that fewer fragments or lower mass accuracy decreases the specificity of identification, whereas the size of the fragments does not significantly affect the specificity of identification (1–8 kDa; data not shown). Notably consecutive C- and N-terminally derived fragments in this size range are generally well represented in experimental data. Because of the astronomical number of possible combinations, we could not per-
form an exhaustive evaluation of non-sequential fragments. A cursory evaluation indicated that higher search specificity could be achieved with similar numbers of non-sequential fragments such that a search with sequential fragments at 1 ppm yields specificity similar to that with non-sequential fragments searched at 10 ppm. This could be due to extra information encoded between two separate sequence tags (b13, b14, and b15 sequence tag yields residues 14 and 15, whereas b13, b16, and b18 sequence tag also yields residue 16).

**Protein Identification using FT-ICR-MS\(^1\) and -MS\(^2\) Data—** Although the masses of intact proteins are in many cases the most informative feature of “top-down” data, intact protein mass is not sufficient for the identification of structural isomers (which have the same mass). The ultimate test would be to distinguish the locations of the same mutations occurring at different parts of the protein. Thus, for evaluating combined performance of BIG Mascot and our top-down methods including funnel-skimmer dissociation (Fig. 3), we acquired FT-ICR-MS\(^1\) and -MS\(^2\) data for 13 different ALS-related variants of human SOD1. Despite reduction with tris(2-carboxyethyl)phosphine, disulfide-oxidized hSOD1 was sometimes observed, resulting in less efficient dissociation and an uncertainty of 2 Da in intact protein mass.

We augmented the MSDB to include ALS-causing SOD1 variants, including four distinct His → Arg variants, three different Gly → Ser variants, three distinct Ala → Thr variants, etc. As a result, identification of a given variant required both intact protein mass and protein fragmentation data. Fig. 4 illustrates spectra of wild type and G93A SOD1 MS\(^1\), MS\(^2\), and zoomed MS\(^2\) data (see supplemental Figs. 2 and 3 for G93A BIG Mascot search results). All 13 variants were uniquely identified and ranked first using ECD, and all except D76Y ranked first using funnel-skimmer dissociation (Table I). Selectivity in identifying some SOD1 variants depended upon mass accuracies commonly achieved by Fourier transform mass spectrometers, for example D124V funnel-skimmer dissociation data (±0.01-Da MS\(^1\) and ±0.01-Da MS\(^2\) error tolerances). From the SORI-ECD results of the 153-residue wild type SOD1, BIG Mascot matched 66 masses including 41 c-, four y-, and 21 z-type ions.

**Mascot** determines the significance threshold by considering the database size searched and absolute probabilities. Our BIG Mascot searches with an enzyme definition of None yielded significance thresholds that are between scores of 60 and 70, values higher than most of the ion scores, while at the same time correct protein ranked first. In contrast, IntactProtein enzyme definition reported the significance level between scores of 10 and 20.

**Protein Identification Using Only FT-ICR-MS\(^2\) Data—** We developed a way to search only fragmentation data using BIG Mascot (“Experimental Procedures”). This method was successful for the identification of all wild type proteins (Table I). For protein variants, these types of searches often (70% of the time) identified the correct gene while failing to identify the correct protein variant (35% correct protein variant). This illustrates the importance of MS\(^1\) information for accurate identification of protein variants. The case of UCHL-1 illustrates the utility of MS\(^2\) for searches of proteins with unknown modifications. Funnel-skimmer dissociation of UCHL-1 yielded high quality MS\(^1\) and MS\(^2\) spectra, yet no protein was identified when using a 2-Da MS\(^1\) tolerance. However, when searched using MS\(^2\) data only as described under “Experimental Procedures,” UCHL-1 was the single significant hit.
and 32 y- and no b-type fragment ions were matched. The protein mass differed by 411 Da from the database prediction, and we were then able to deduce that the UCHL-1 sequence contained an N-terminal addition (GPLGS-) that was an artifact of recombinant expression and purification techniques.

The case of thyroglobulin illustrates the utility of MS²-only searches for the identification of intact proteins that are either heterogeneous or too large to detect. Moreover the current version of BIG Mascot does not allow precursor masses larger than 110 kDa, and therefore MS²-only searches are necessary to overcome this limitation for any protein larger than 110 kDa whether or not the intact protein is detected. We attempted to detect intact thyroglobulin without success. We observed only low signal to noise, unresolved signal at 2500–4000 m/z (data not shown) using electrospray, and no signal using MALDI-TOF (data not shown). Thyroglobulin is not only large in size but is also heterogeneous in composition with many PTMs including iodination, glycosylation, and disulfide bonds (44). We performed funnel-skimmer dissociation from two different solvents (Fig. 5). Using 50% acetonitrile, 0.1% formic acid (Fig. 5, top panel) and performing MS²-only searches we identified internal fragments of thyroglobulin resulting from the cleavage of the peptide bond between Val²²⁰³ and Pro²²⁰⁴ (supplemental Figs. 4–6). Specifically BIG Mascot identified fragments between Pro²²⁰⁴ and the C terminus, suggesting that thyroglobulin fragmented at the N terminus of Pro²²⁰⁴ and then fragmented further to give the sequence tag. Other fragments matched included a b-ion series that terminated at Pro²²³⁴ (supplemental Fig. 6). The second spectrum, acquired using 10 mΜ ammonium acetate, pH 6.8, yielded a considerably different spectrum with some higher weight fragments (Fig. 5). Use of this spectrum still yielded the identification of thyroglobulin (supplemental Fig. 7) but from different regions of sequence (Thr²²⁰⁰–Gln¹⁰⁹⁸) consistent with the observation that solvent additives effect patterns of protein cleavage (9).

Including internal fragments in BIG Mascot searches increases the chance of random matches for the following reasons. 1) There are comparatively more hypothetical internal fragments per proteins than N or C terminus-containing fragments (for wild type SOD1, a 153-residue protein, there are 11,325 possible internal fragments in contrast to 898 b- and y-type fragments including ammonia and water losses). 2) As discussed in Zhai et al. (34), many internal fragment sequences within the sequence of a given protein have identical masses and pose an ambiguity problem. 3) A general problem, but one that is amplified by internal fragment searches, is that BIG Mascot defines the fragment mass tolerance in mass units, as opposed to ppm, because at 500 Da a 0.05-Da error is 100 ppm, whereas at 15,000 Da a 0.05-Da error is 3.3 ppm. Therefore, the proper choice of mass tolerance for a given fragment results in the loss of sensitivity for fragments above that mass and false positives for fragments below that mass. For these reasons, incorrect and ambiguous internal fragment matches occur and sometimes yield a number of matches
higher than the number of experimental ions Mascot used (Table II). To reduce the number of random internal fragment matches, we therefore internally calibrated our data and performed searches with tighter fragment mass tolerances of 0.05 Da for instrument fragment definitions that included by-type internal fragments.

Including internal fragments in searches did not alter the ranks of the correctly identified proteins and increased BIG Mascot scores except in two cases (D125H and D90A SOD1). In the case of D76Y the rank of the variant increased to first place. We observed a significant improvement in proportion of assigned peaks especially for SOD1 variants, BSA, and thyroglobulin.

In the case of BSA, in addition to internal fragments, we observed 11 a-ions. Whenever a-ions were observed, the corresponding b-ions were observed at higher intensity. We did not observe significant a-ions in other proteins. Because of these and because addition of each ion series introduces the chance of random matches, we do not report a-ion searches except in the case of BSA and do not recommend the inclusion of a-ions for the search of unknown proteins.

**Comparison of Mascot and ProSight PTM Scoring Algorithms**—ProSight PTM is the gold standard for top-down database searching and modification characterization. Thus, we compared the Mascot and ProSight PTM (version 1.0) scoring algorithms, specifically how each would score identical MS1 and MS2 data using an identical database, by calculating the probabilities of a random match reported by each algorithm. The value of BIG Mascot-assigned probability of a random match was calculated from the ion scores using the following formula:

$$\text{Score}^1 / (1000^1 / (1000^2 \log(p)))$$

where $x = (1/111.1) \times 2 \times (M_a \times 2)$. Here $f$ is the number of fragment values submitted, $n$ is the number of random fragment ion hits, and $M_a$ is mass accuracy. The probability reported in Fig. 6 does not relate to the number of protein

---

**Table I**

**BIG Mascot MS² search results summary**

For human proteins, only human entries were searched; for the others, mammalian sequences were searched. Red color indicates that other proteins of identical mass (±2 ppm) are in the database; blue color indicates that the correct variant shares the rank with other variants; an asterisk indicates that the protein ranking first was another SOD1 variant. NF, not found; NS, not searched; N/A, not applicable. §, indicates SORI-ECD.
forms in the database. The number of protein forms is neither specified nor used for the calculation of the ion scores of Mascot or the calculation of Poisson probability, thus giving a measure that is independent of the database. As a result of the probability calculations explained above, BIG Mascot consistently provided a higher probability of a false match (was more stringent) compared to ProSight PTM (Fig. 6).

**DISCUSSION**

Here we debut both funnel-skimmer dissociation and the BIG Mascot top-down database search engine. Using these tools we 1) identified 13 variants of the same protein, 2) improved the upper mass limit of top-down methods to the 669 kDa, 3) demonstrated that funnel-skimmer dissociation generates frequent internal fragmentation and interresidue cleavage C-terminal to proline, and 4) identified proteins using either intact protein mass, protein fragments (MS²-only), or a combination of the two. Finally we compared the BIG Mascot and ProSight PTM top-down data base search engines and demonstrated that the BIG Mascot scoring algorithm is currently too stringent. Here we used protein variants to evaluate the specificity of our methods and identified SOD1 variants correctly out of a possible 102 distinct variants, a feat we could not achieve using bottom-up methods.

Funnel-skimmer dissociation occurs following the skimmer 1 region and under a pressure of a few millitorrs in analogy to prefolding dissociation (9). Our setup, however, includes 1) an ion funnel instead of a nozzle and 2) an additional funnel and skimmer stack following the first skimmer where activation occurs. Funnel-skimmer dissociation is an in-source fragmentation technique that has the following advantages. 1) It does not require automated real time ion selection software, 2) it has a high efficiency, and 3) it allows for dissociation of first generation fragment ions (pseudo-MS³). Limitations of in-source dissociation techniques include 1) the inability to purify the precursor within the mass spectrometer, therefore requiring separation or purification, LC, and/or ion mobility separation, and 2) a higher percentage of unassigned fragments, including internal fragments, compared to ECD. We found ECD to be more specific but it exhibited lower fragmentation efficiency, which could potentially be increased using activation techniques (56).

We showed fragmentation of larger proteins using funnel-skimmer dissociation and demonstrated fragmentation efficiency similar to the nozzle-skimmer results with BSA (51) and the additional identification of significant numbers of internal fragments (Table II). We have provided the dissociation of thyroglobulin (669 kDa as a dimer). Thyroglobulin is reported to be a very stable molecule, held together by not only noncovalent forces but also many intrachain disulfide and dityrosine bonds (44). The fact that we observed a heterogenous unresolved spectrum is in accord with electrophoretic analy-

---

**Fig. 5.** Funnel-skimmer dissociation spectrum of thyroglobulin sprayed from two different solvent compositions. Assignments of m/z and charges were made using SNAP2 reflecting monoisotopic masses. Each sequence shown is the best sequence part identified. Intens., intensity; FA, formic acid; FSD, funnel-skimmer dissociation.

---

### Identification of Proteins Using Top-down Mass Spectrometry

Molecular & Cellular Proteomics 8.4
TABLE II
Summary of BIG Mascot MS² searches of CAD spectra when internal fragment search was enabled

| Protein    | Rank | Score (ID/used) | Fragments | Decoy score |
|------------|------|-----------------|-----------|-------------|
| Ubiquitin  | 1    | 322             | 134/213   | 10          |
| DJ1        | 1    | 73              | 64/83     | 3           |
| UCHL1      | 1    | 137             | 234/204   | 2           |
| Myoglobin  | 1    | 84              | 112/118   | 1020*       |
| Albumin    | 1    | 87              | 419/130   | 4           |
| Thyroglobulin | 1    | 24              | 181/71    | 16          |
| WT SOD     | 1    | 120             | 124/135   | 1020*       |
| D125H SOD  | 1    | 7               | 35/73     | 1           |
| S134N SOD  | 1    | 52              | 116/105   | 4           |
| D90A SOD   | 1    | 12              | 40/77     | 1020*       |
| G93A SOD   | 1    | 162             | 329/259   | 2           |
| A4V SOD    | 1    | 22              | 92/136    | 5           |
| G85R SOD   | 1    | 63              | 173/198   | 1020*       |
| E133del SOD | 1   | 58              | 95/88     | 9           |
| G41S SOD   | 1    | 68              | 54/62     | 6           |
| D76Y SOD   | 1    | 34              | 57/78     | 4           |
| L38V SOD   | 1    | 67              | 111/118   | 6           |
| G72S SOD   | 1    | 48              | 124/164   | 1020*       |
| D124V SOD  | 1    | 34              | 113/203   | 0           |

Fig. 6. Comparison of probabilities reported by Mascot and ProSight PTM. Probability of a random match for each set of funnel-skimmer dissociation data reported in Table I was calculated using BIG Mascot scores (empty circles) and by the Poisson method described previously (36) (filled circles).

2 Sigma-Aldrich certificate of analysis for product number T1001.

Identification of Proteins Using Top-down Mass Spectrometry

An instrument definition was created to allow internal fragments as large as 30 kDa to be searched in addition to b- and y-ions. This was used to search our CAD data generated by funnel-skimmer dissociation. Because the number of masses searched against is increased significantly, we included decoy database searches and report the score of the highest scoring protein hit. The score of 1020 is caused by a glitch that we observed when no fragments were assigned to any hits, and therefore scoring changes over to peptide mass fingerprint mode. Thus the random hit has zero score in MS² mode scoring.

Internal fragments are often overlooked, although they are observed in both in-source and in-cell fragmentation methods (34). Overall internal fragments are important because their identification increases the confidence in correct protein match by 1) decreasing the number of unmatched ions and 2) shortening the time to identify the protein (see Fig. 1 and supplemental Fig. 1). Mass of fragments alone can give information specific enough to uniquely identify a protein but is insufficient for distinguishing different variants of the protein (see Fig. 2 and Table I). Combination of MS¹ and MS² enables the complete and accurate identification of variants and PTMs. In the bottom-up ap-

Molecular & Cellular Proteomics 8.4 853
Identification of Proteins Using Top-down Mass Spectrometry

proach, confident protein identification using peptide mass alone is generally not possible (64) for the following reason: tryptic digestion of proteins removes the broad, natural distribution of masses (Fig. 1) and replaces it with a narrow distribution of many more peptides distributed mostly between 400 and 2000 Da. Using more specific proteases such as Lys-C widens this distribution and offers a compromise between top-down and bottom-up approaches (referred to as “middle-down”) (20, 50, 65).

We found that probability estimates reported by BIG Mascot were always lower with respect to ProSight PTM scores (11) for top-down MS searches (Fig. 6). The question then arises as to whether Mascot is too stringent or ProSight PTM is not stringent enough. We consider Mascot to be too stringent for the following reasons. 1) The ProSight PTM scoring mechanism is published, has been peer reviewed, and is based upon well established statistical principles. 2) On more than one occasion (A4V and D125H SOD1 CAD; S134N SOD1 and UCHL1 ECD) BIG Mascot assigned a non-significant score to a protein that it ranked first out of a 1,943,018-protein database. Based upon our results, we conclude that the Mascot scoring algorithm requires adjustment for top-down data (it was originally calibrated empirically using bottom-up data).

We suggest that the following improvements need to be made to further improve the BIG Mascot search engine: 1) parts per million MS$^2$ search error tolerance units should be included in the parameters, 2) the upper mass limit needs to be extended to equal the largest protein in the database, 3) scoring needs to be revised to reflect a realistic specificity of top-down MS data in matching the protein reported, and 4) a glitch that causes false identifications to yield a score of 1020 needs to be fixed. To test and improve either database search engine, studies with larger data sets are needed.

We suggest that BIG Mascot will prove useful for attempts to automate the identification of proteins. Strengths of the approach include: 1) its sensitivity, the ability to identify the correct protein within all mammalian subsequences; 2) its specificity, its ability demonstrated by distinguishing very similar protein variants; 3) the contribution of MS$^3$ to the scoring system within the significance threshold calculation; and 4) the ability to search for internal fragments that are significant especially for in-source dissociation techniques.

Note Added in Proof—Following acceptance of the manuscript, BIG Mascot was branded as MascotTD.

Acknowledgments—We thank Drs. Quyen Hoang and Chitta Das for providing DJ-1 and UCHL-1 samples, Scott Daniels from Bruker Daltonics Inc. for help with technical questions, Dr. Nathalie Y. R. Agar for help with Adobe Illustrator and thoughtful discussions, John Cottrell and David Creasy from Matrix Science for providing a version of BIG Mascot for us, and the entire Agar laboratory for thoughtful comments.

* This work was supported by United States Department of Defense Award W81XWH-04-0158 and Amyotrophic Lateral Sclerosis Society of America (ALSA) Grant 1392 from the (to J. N. A.).

[5] The on-line version of this article (available at http://www.mpongline.org) contains supplemental material.

§ Supported by the ALSA and the ALS Therapy Alliance.

†† To whom correspondence and requests for data sets should be addressed: Dept. of Chemistry and Volen Center for Complex Systems, MS 015, Brandeis University, 415 South St., Waltham, MA 02454. Tel.: 781-736-2425; Fax: 781-736-2405; E-mail: agar@brandeis.edu.

REFERENCES

1. Yamashita, M., and Fenn, J. B. (1984) Electrospray ion source. Another variation on the free-jet theme. J. Phys. Chem. 88, 4451–4459

2. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246, 64–71

3. Luo, J. A., Udseth, H. R., Smith, R. D., and Futrell, J. H. (1998) Collisional effects on the charge distribution of ions from large molecules, formed by electrospray-ionization mass spectrometry. Rapid Commun. Mass Spectrom. 2, 207–210

4. Gauthier, J. W., Trautman, T. R., and Jacobson, D. B. (1991) Sustained off-resonance irradiation for collision-activated dissociation involving Fourier-transform mass-spectrometry-collision-activated dissociation technique that emulates infrared multiphoton dissociation. Anal. Chem. Acta 246, 211–225

5. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. J. Am. Chem. Soc. 120, 3265–3266

6. Zubarev, R. A., Horn, D. M., Fridriksson, E. K., Kelleher, N. L., Kruger, N. A., Lewis, M. A., Carpenter, B. K., and McLafferty, F. W. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations. Anal. Chem. 72, 563–573

7. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 101, 9528–9533

8. Coon, J. J., Ueberheide, B., Syka, J. E. P., Dryhurst, D. D., Ausio, J., Shabanowitz, J., and Hunt, D. F. (2005) Protein identification using sequential ion/ion reactions and tandem mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 102, 9463–9468

9. Han, X. M., Jin, M., Breuker, K., and McLafferty, F. W. (2006) Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. Science 314, 109–112

10. Martz, E., O’Connor, P. B., Roeptoroff, P., Kelleher, N. L., Wood, T. D., McLafferty, F. W., and Mann, M. (1996) Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. Proc. Natl. Acad. Sci. U. S. A. 93, 8264–8267

11. Taylor, G. K., Kim, Y. B., Forbes, A. J., Meng, F. Y., McCarthy, R., and Kelleher, N. L. (2003) Web and database software for identification of intact proteins using “top down” mass spectrometry. Anal. Chem. 75, 4081–4086

12. LeDuc, R. D., Taylor, G. K., Kim, Y. B., Januszky, T. E., Byrum, 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

13. Strader, M. B., VerBerkmoes, N. C., Tabb, D. L., Connelly, H. M., Barton, J. W., Bruce, B. D., Pelletier, D. A., Davison, B. H., Hettich, R. L., Larimer, F. W., and Hurst, G. B. (2004) Characterization of the 705 ribosome from Rhodopseudomonas palustris using an integrated “top-down” and “bottom-up” mass spectrometric approach. J. Proteome Res. 3, 965–978

14. Roth, M. J., Forbes, A. J., Boyne, M. T., Kim, Y. B., Robinson, D. E., and Kelleher, N. L. (2005) Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. Mol. Cell. Proteomics 4, 1002–1008

15. Patie, 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 <60-kDa proteins from Methanococcinae acetivorans using quadrupole FTMS with automated octopole collisionally activated dissociation. Mol. Cell. Proteomics 5, 14–25

16. Sutli, N., Roth, M. J., Mizzen, C. A., Kelleher, N. L., and Pesavento, J. J. (2006) Gene-specific characterization of human histone H2B by electron capture dissociation. J. Proteome Res. 5, 233–239
Identification of Proteins Using Top-down Mass Spectrometry

17. Whitelegge, J., Halgand, F., Souda, P., and Zabrouskeep, V. (2006) Top-down mass spectrometry of integral membrane proteins. Expert Rev. Proteomics 3, 585–596

18. Sharma, S., Simpson, D. C., Tolic, N., Jaitly, N., Mayampurath, A. M., Smith, R. D., and Pasa-Tolic, L. (2007) Peptideomic profiling of intact proteins using WAX-RPLC 2-D separations and FTICR mass spectrometry. J. Proteome Res. 6, 602–610

19. Parks, B. A., Jiang, L., Thomas, P. M., Wenger, C. D., Roth, M. J., Boyne, M. T., Burke, P. V., Kwast, K. E., and Kelleher, N. L. (2007) Top-down proteomics on a chromatographic time scale using linear ion trap Fourier transform hybrid mass spectrometers. Anal. Chem. 79, 7984–7991

20. Kelleher, N. L., Lin, H. Y., Valaskovic, G. A., Aaserud, D. J., Fridriksson, E. K., and McLafferty, F. W. (1999) Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. J. Am. Chem. Soc. 121, 806–812

21. Breuker, K., Jin, M., Han, X. M., Jiang, H., and McLafferty, F. (2008) Top-down identification and characterization of biomolecules by mass spectrometry. J. Am. Soc. Mass Spectrom. 19, 1045–1053

22. Reid, G. E., and McLuckey, S. A. (2002) ‘Top down’ protein characterization via tandem mass spectrometry. J. Mass Spectrom. 37, 663–675

23. Kelleher, N. L. (2004) Top-down proteomics. Anal. Chem. 76, 196A–202A

24. Bogdanov, B., and Smith, R. D. (2005) Proteomics by FTICR mass spectrometry: top down and bottom up. Mass Spectrom. Rev. 24, 168–200

25. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectroscopic sequencing of proteins from silver-stained polyacrylamide gels. Anal. Chem. 68, 850–858

26. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986) Protein sequencing by tandem mass-spectrometry. Proc. Natl. Acad. Sci. U.S.A. 83, 6233–6237

27. Sadygov, R. G., Cociorva, D., and Yates, J. R. (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat. Methods 1, 195–202

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

29. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989

30. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (±0.1 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal. Chem. 71, 2871–2882

31. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. 66, 4390–4399

32. Simpson, D. C., Ahn, S., Pasa-Tolic, L., Bogdanov, B., Mottaz, H. M., Vilkov, A. N., Anderson, G. A., Lipton, M. S., and Smith, R. D. (2006) Using size exclusion chromatography-RPLC and RPLC-CIEF as two-dimensional separation strategies for protein profiling. Electrophoresis 27, 2227–2237

33. Eyres, S. J., Speir, J. P., Kruppa, G. H., Gierasch, L. M., and Kaltashov, I. A. (2000) Protein conformational stability probed by Fourier transform ion cyclotron resonance mass spectrometry. Rapid Commun. Mass Spectrom. 17, 1759–1768

34. Wilson, M. A., Collins, J. L., Hod, Y., Rhee, D., and Pitsko, G. A. (2003) The 1.1-A resolution crystal structure of DJ-1, the protein mutated in autosomal recessive early onset Parkinson's disease. Proc. Natl. Acad. Sci. U.S.A. 100, 9256–9261

35. Kim, T., Tolmachev, A. V., Harkevick, R., Prior, D. C., Anderson, G., Udslett, H. R., Bailey, T. H., Rakov, S., Futrell, J. H., and Smith, R. D. (2000) Design and implementation of a new electrodynamic ion funnel. Anal. Chem. 72, 2247–2255

36. Tsymb, Y. O., Witt, M., Baykut, G., Kjeldsen, F., and Nakamangelo, P. (2003) Combined infrared multiphoton dissociation and electron capture dissociation with a hollow electron beam in Fourier transform ion cyclotron resonance mass spectrometry. Rapid Commun. Mass Spectrom. 17, 1759–1768

37. Gorkovsko, M. V., Masselon, C. D., Nikolaev, E. N., Udslett, H. R., Pasa-Tolic, L., and Smith, R. D. (2004) Considerations for electron capture dissociation efficiency in FTICR mass spectrometry. Int. J. Mass Spectrom. 234, 131–136

38. Loo, J. A., Edmonds, C. G., and Smith, R. D. (1991) Tandem mass-spectrometry of very large molecules—serum albumin sequence information from multiply charged ions formed by electrospray ionization. Anal. Chem. 63, 2488–2499

39. Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995) Determination of monosaccharide masses and ion populations for large biomolecules from resolved isotopic distributions. J. Am. Soc. Mass Spectrom. 6, 229–233

40. Kelleher, N. L., Senko, M. W., Siegel, M. M., and McLafferty, F. W. (1997) Unit resolution mass spectra of 112 kDa molecules with 3 Da accuracy. J. Am. Soc. Mass Spectrom. 8, 380–383

41. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulos, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. Proteomics 4, 1985–1988

42. Ropstorp, P., and Fohman, I. (1990) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed. Mass Spectrom. 11, 601

43. Horn, D., Ge, Y., and McLafferty, F. (2000) Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins. Anal. Chem. 72, 4778–4784

44. Loo, J. A. (1995) Observation of large subunit protein complexes by electrospray ionization mass spectrometry. J. Mass Spectrom. 30, 180–183

Molecular & Cellular Proteomics 8.4
58. Loo, J. A. (1997) Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom. Rev.* 16, 1–23
59. Wang, Y., Schubert, M., Ingendoh, A., and Franzen, J. (2000) Analysis of non-covalent protein complexes up to 290 kDa using electrospray ionization and ion trap mass spectrometry. *Rapid Commun. Mass Spectrom.* 14, 12–17
60. Huang, Y. Y., Triscari, J. M., Tseng, G. C., Pasa-Tolic, L., Lipton, M. S., Smith, R. D., and Wysocki, V. H. (2005) Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal. Chem.* 77, 5800–5813
61. Loo, J. A., Edmonds, C. G., and Smith, R. D. (1993) Tandem mass spectrometry of very large molecules. 2. Dissociation of multiply charged proline-containing proteins from electrospray ionization. *Anal. Chem.* 65, 425–438
62. Gardner, M., and Brodbelt, J. (2008) Impact of proline and aspartic acid residues on the dissociation of intermolecularly crosslinked peptides. *J. Am. Soc. Mass Spectrom.* 19, 344–357
63. Mann, M., Hojrup, P., and Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338–345
64. Conrads, T. P., Anderson, G. A., Veenstra, T. D., Pasa-Tolic, L., and Smith, R. D. (2000) Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* 72, 3349–3354
65. Wu, S. L., Kim, J., Hancock, W. S., and Karger, B. (2005) Extended range proteomic analysis (ERPA): a new and sensitive LC-MS platform for high sequence coverage of complex proteins with extensive post-translational modifications—comprehensive analysis of β-casein and epidermal growth factor receptor (EGFR). *J. Proteome Res.* 4, 1155–1170