Anal. Chem. **2000,** *72,* 1112−1118

Protein Identification with a Single Accurate Mass of a Cysteine-Containing Peptide and Constrained Database Searching

David R. Goodlett,*,† James E. Bruce,‡ Gordon A. Anderson,§ Beate Rist,† Ljiljana Pasa-Tolic,§ Oliver Fiehn,|| Richard D. Smith,§ and Ruedi Aebersold†

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195, Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352, Max-Planck-Institute of Molecular Plant Physiology, Potsdam, Germany, and Merck Research Laboratories, West Point, Pennsylvania 19486

A method for rapid and unambiguous identification of proteins by sequence database searching using the accurate mass of a single peptide and specific sequence constraints is described. Peptide masses were measured using electrospray ionization-Fourier transform ion cyclotron resonance mass spectrometry to an accuracy of 1 ppm. The presence of a cysteine residue within a peptide sequence was used as a database searching constraint to reduce the number of potential database hits. Cysteine-containing peptides were detected within a mixture of peptides by incorporating chlorine into a general alkylating reagent specific for cysteine residues. Secondary search constraints included the specificity of the protease used for protein digestion and the molecular mass of the protein estimated by gel electrophoresis. The natural isotopic distribution of chlorine encoded the cysteine-containing peptide with a distinctive isotopic pattern that allowed automatic screening of mass spectra. The method is demonstrated for a peptide standard and unknown proteins from a yeast lysate using all 6118 possible yeast open reading frames as a database. As judged by calculation of codon bias, low-abundance proteins were identified from the yeast lysate using this new method but not by traditional methods such as tandem mass spectrometry via data-dependent acquisition or mass mapping.

Traditionally, protein sequences were determined by stepwise, chemical degradation of purified proteins or fragments thereof. With the advent of sequence databases that contain complete genomic sequences or large numbers of complete or partial expressed gene sequences (i.e., expressed sequence tags, ESTs),[1−3] the sequences and identities of most proteins from species prominently represented in sequence databases can be determined by correlating experimental data extracted from the protein with sequence databases.[4−6] The many implemented sequence database searching strategies have in common the use of a combination of specific constraints to narrow down a candidate list of matching proteins in a database to a single sequence.[7] Currently, the most restrictive constraints are generated after proteolysis of a protein by mass spectrometric (MS) mass mapping using multiple peptide masses derived from the same protein or tandem mass spectrometric (MS/MS) analysis of single peptides in mixtures.

The constraints provided by collision-induced dissociation (CID) of selected peptides in a tandem mass spectrometer are highly discriminating because CID spectra reflect the amino acid sequence of the peptide analyzed. Tandem MS is frequently practiced with peptides separated by capillary HPLC or capillary electrophoresis (CE) connected on-line to an electrospray ionization (ESI) tandem MS instrument because better sensitivity is achievable via the concentrating effect of a separation than by direct infusion methods. Peptides eluting from the separation system are analyzed using the first-stage mass spectrometer that also selects peptide ions for CID via data-dependent procedures, followed by fragment ion analysis in a second-stage mass spectrometer. Observed CID spectra are used to identify the protein from which the peptide originated, either by automated correlation of uninterpreted CID spectra with a sequence database or by searching sequence databases with complete or partial

* To whom correspondence should be addressed: Department of Molecular Biotechnology, Health Sciences Building K327, Box 357730, University of Washington, Seattle, WA 98195; (tel) 206.616.1021; (fax) 206.685.7301; (e-mail) goodlett@u.washington.edu.

† University of Washington

‡ Merck Research Laboratories.

§ Pacific Northwest National Laboratory.

|| Max-Planck-Institute of Molecular Plant Physiology.

(1) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Gailbert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsin, P.; Tettelin, H.; Oliver, S. G. *Science* **1996**, *274*, 546−549.

(2) Fraser, C. M.; Casjens, S.; Huang, W. M.; Sutton, G. G.; Clayton, R.; Lathigra, R.; White, O.; Ketchum, K. A.; Dodson, R.; Hickey, E. K.; Gwinn, M.; Dougherty, B.; Tomb, J. F.; Fleischmann, R. D.; Richardson, D.; Peterson, J.; Kerlavage, A. R.; Quackenbush, J.; Salzberg, S.; Hanson, M.; van Vugt, R.; Palmer, N.; Adams, M. D.; Gocayne, J.; Weidman, J.; Utterback, T.; Watthey, T.; McDonald, L.; Artiach, P.; Bowman, C.; Garland, S.; Fujii, C.; Cotton, M. D.; Horst, K.; Roberts, K.; Hatch, B.; Smith, H. O.; Venter, J. C. *Nature* **1997**, *390*, 580−586.

(3) Neubauer, G.; King, A.; Rappsilber, J.; Calvio, C.; Watson, M.; Ajuh, P.; Sleeman, J.; Lamond, A.; Mann, M. *Nat. Genet.* **1998**, *20*, 46−50.

(4) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011−5015.

(5) Eng, J.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass. Spectrom.* **1994**, *5*, 976−989.

(6) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.

(7) Patterson, S. D.; Aebersold, R. *Electrophoresis* **1995**, *16*, 1791−1814.

10.1021/ac9913210 CCC: $19.00
Published on Web 02/08/2000

peptide sequences obtained by manual or computer-assisted interpretation of CID spectra.[5,6] The method has the significant advantage that in most cases a CID spectrum from a single peptide species is sufficient to conclusively identify a protein.[8] Consequently, proteins can be identified by correlating a single CID mass spectrum with databases containing incomplete gene sequences as found in EST databases. Furthermore, components of protein mixtures can be identified without the need for purification to homogeneity and proteins can be identified across species, provided that the peptide segment analyzed is conserved between these species. The method has the disadvantage that peptide ions need to be sequentially selected for CID out of a mixture of analytes.[9] In cases where complex protein mixtures are digested, the CID data acquisition rate often fails to generate CID on all possible peptide ions in the time available for analysis. For automated, data-dependent CID the mass spectrometer is generally programmed to give highest priority for CID selection to ions with the highest ion current.[9] Therefore, if complex peptide mixtures are analyzed, lower intensity peptide ions will not be selected for CID even though their ion current may exceed the detection limit of the mass spectrometer, thus effectively reducing the dynamic range for analysis. This effective compression of dynamic range is due in part to the speed at which tandem MS spectra can be acquired. It can be somewhat alleviated, but not completely eliminated, by extending the time for chromatographic separation.[10−14]

The accurately measured masses of peptides in a protein digest represent a different type of constraint for database searching. Their use for protein identification is referred to as peptide mass mapping or fingerprinting. Such peptide mass profiles or fingerprints are determined in a single stage of mass spectrometry without CID. The list of observed peptide masses, together with auxiliary constraints (e.g., including the estimated molecular weight of the unfragmented parent protein and the cleavage specificity of the protease), is then searched against sequence databases using any one of a number of available algorithms.[4,7] Peptide mass mapping therefore identifies proteins without sequence-specific information because the subset of peptide masses created by digestion of a protein with a specific protease defines the N- or C-terminus of each fragment and thus provides a set of constraints unique to a given protein. The more accurately peptide masses are measured and the more peptide masses are detected from the same protein, the more conclusively the protein identity can be determined.[15,16] The peptide mass mapping approach has the advantage over the tandem MS strategy that the mass spectrometer operates in full-scan mode (i.e., in a single stage) for the duration of the experiment and should generally provide greater sensitivity without compression of dynamic range. However, peptide mass fingerprinting is incompatible with searching EST databases because of the low probability that a sufficient number of peptide masses will match a single EST and provide an unambiguous correlation. Recent calculations for proteins expressed by the genomes of *Escherichia coli* and *Saccharomyces cerevisiae* indicate that at 0.1 ppm mass accuracy 96% of the proteins will generate tryptic peptides with one or more unique masses. The remaining 4% of proteins correspond to gene products that are largely duplicates of other proteins, suggesting the feasibility of protein identification based on the mass of a single peptide (Figure 1).[17,18]

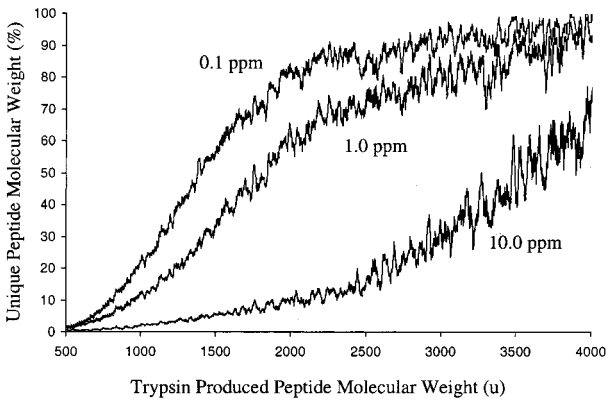

Figure 1. All possible unique peptide molecular weights after digestion of all yeast proteins in the National Center for Biotechnology Information at a mass accuracy of 0.1, 1.0, and 10.0 ppm.

Protein identification using a single accurate peptide mass would combine the advantages of the mass mapping and tandem MS approaches to protein identification while eliminating significant limitations of the respective methods. Just as with the other approaches, inclusion of additional constraints such as the estimated molecular weight of the parent protein, the cleavage specificity of the protease used to digest the parent protein, and the presence of a relatively rare amino acid such as cysteine, methionine, or tryptophan in the peptide sequence would further enhance the stringency of the database search. Among these constraints, the presence of cysteine in a peptide sequence is particularly attractive because the sulfhydryl side chain of cysteine residues is chemically distinct among amino acid residues and its presence significantly constrains the database search while still covering ~92% of the open reading frames in yeast.[19]

Here we report a method allowing protein identification by accurate mass measurement of a single cysteine-containing peptide. Mass spectra were acquired in a single stage of mass spectrometry using electrospray ionization-Fourier transform ion cyclotron resonance mass spectrometry (ESI-FTICR MS) at a mass accuracy of 1 ppm.[20] Cysteine-containing peptides were

(8) Susin, S. A.; Lorenzo, H. K.; Zamzami, N.; Marzo, I.; Brothers, G.; Snow, B.; Jacotot, E.; Costantini, P.; Larochette, N.; Goodlett, D. R.; Aebersold, R.; Pietu, G.; Prevost, M.-C.; Siderovski, D.; Penninger, J. and Kroemer, G. *Nature* **1999**, *397*, 441−446.

(9) Ducret, A.; van Oostveen, I.; Eng, J. K.; Yates, J. R., III; Aebersold, R. *Protein Sci.* **1998**, *7*, 706−719.

(10) Goodlett, D. R.; Wahl, J. H.; Udseth, H. R.; Smith, R. D. *J. Microcolumn Sep.* **1993**, *5*, 57−62.

(11) Davis, M. T.; Lee, T. D. *J. Am. Soc. Mass. Spectrom.* **1996**, *9*, 194−201.

(12) Figeys, D.; Corthals, G. L.; Gallis, B.; Goodlett, D. R.; Ducret, A.; Corson, M. A.; Aebersold, R. *Anal. Chem.* **1999**, *71*, 2279−2287.

(13) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III *Nat. Biotechnol.* **1999**, *17*, 676−682.

(14) Opiteck, G. J.; Lewis, K. C.; Jorgenson, J. W.; Anderegg, R. J. *Anal. Chem.* **1997**, *69*, 1518−1524.

(15) Fenyö, D.; Qin, J.; Chait, B. T. *Electrophoresis* **1998**, *19*, 998−1005.

(16) Zubarev, R. A.; Håkansson, P.; Sundqvist, B. *Anal. Chem.* **1996**, *68*, 4060−4063.

(17) Bruce, J. A.; Anderson, G. A.; Wen, J.; Harkewicz, R.; Smith, R. D. *Anal. Chem.* **1999**, *71*, 2595−2599.

(18) Anderson, G. A.; Bruce, J. E.; Pasa-Tolic, L.; Smith, R. D. *Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, May 31−June 4, 1998; p 1270.

(19) Sechi, S.; Chait, B. T. *Anal. Chem.* **1998**, *70*, 5150−5158.

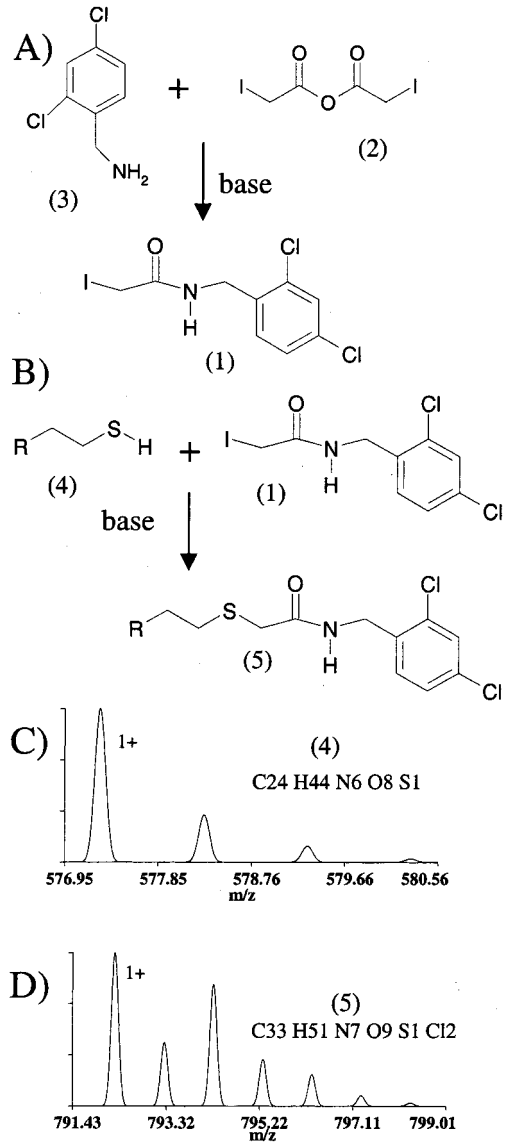

Figure 2. Synthesis of the alkylating IDEnT, 2,4-dichlorobenzyliodoacetamide (A); alkylation of **4**, a hypothetical ($R = C_{24}H_{43}N_6O_8S_1$) cysteine-containing peptide with **1** (B); and the modeled isotopic pattern for $[R + H]^+$ before (C) and after (D) alkylation with **1**.

identified in mixtures by selective alkylation with an isotope distribution encoded tag (IDEnT) that contains an element not normally found in proteins, chlorine.

EXPERIMENTAL SECTION

Reagent Synthesis. Preparation of the cysteine-alkylating IDEnT reagent 2,4-dichlorobenzyliodoacetamide (**1**) was by addition of a 0.5 molar excess of iodoacetic anhydride (**2**) to a 500 mM solution of 2,4-dichlorobenzylamine (**3**) in *N*,*N*-dimethylformamide as shown in Figure 2A. This mixture was stirred in the dark for a total of 2 h. At 30 and 60 min after initial mixing, an amount of *N*,*N*-diisopropylamine in *N*,*N*-dimethylformamide, equimolar to the 2,4-dichlorobenzylamine, was added. After 2 h, the reaction mixture was evaporated to dryness, redissolved in acetonitrile/water, and purified by reversed-phase HPLC. The reagent was stored at −4 °C in 100% acetonitrile, and purity was estimated as 100% by HPLC. The structure and reactivity of the reagent were determined by mass spectrometry. All reagents, peptide standards, and solvents were purchased from Aldrich Chemical Co. (St. Loius, MO)

Peptide Labeling. A 2-fold molar excess of **1** was allowed to react with an alkaline solution (50 mM ammonium bicarbonate) of laminin B1 peptide (Sigma Chemical Co.) for 2 h in the dark. After labeling with the reagent, laminin B1 (RYVVLPRPVCFEKGMNYTVR) was digested with trypsin to produce two fragments of which one (RYVVLPRPVCFEK) was labeled with the IDEnT reagent and one (GMNYTVR) was not. Prior to MS analysis the sample was desalted on a C-18 Ziptip (Millipore Corp., Bedford, MA).

Protein Labeling. Yeast (K1322) cells were inoculated into 5 mL of media containing raffinose without uracil and expanded to 30 mL after 24 h. Cells were then switched to 275 mL of media containing galactose without uracil and harvested at log phase. A lysate was prepared according to previously reported methods.[21] The lysate was fractionated by ion-exchange chromatography on a Q-Sepharose column to reduce the complexity of the mixture such that discrete protein bands were visible after SDS−PAGE separation and sliver staining of a single fraction. Three fractions (1 mg of protein total) eluted at ∼50 mM sodium chloride were pooled, desalted three times with a Centricon Spin column (Millipore Corp.) using 50 mM ammonium bicarbonate, and diluted finally to 370 μL. After desalting, 20 μL (or 90 μg total of protein) of the sample was combined with 20 μL of 3% SDS/0.9 M TRIS, pH 8.5, and 10 μL of 1 nmol/μL dithiothreitol. This mixture was boiled for 5 min and then 10 μL of the IDEnT reagent, **1** (20 nmol/μL), in 100% acetonitrile was added and allowed to react for 1 h in the dark at room temperature. Prior to separation by SDS−PAGE, the samples was subjected to vacuum concentration (3 min) in a Speedvac to remove excess acetonitrile and 16 μL of a solution of glycerol (40%) and bromophenol blue added. Labeled proteins were separated by SDS−PAGE (10% C) and detected by silver staining, and protein bands of interest were excised/digested in the gel with trypsin.[22,23]

Mass Spectrometry Desalted peptide standard samples were analyzed by direct infusion into an ESI-FTICR MS using an 11.5-T magnet. Conditions for operation of the FTICR MS were similar to those reported elsewhere, and external mass calibration was established with an ESI-FTICR mass spectrum using a peptide mixture generated by tryptic digestion of bovine serum albumin.[17] Mass spectra were obtained each 2.5 s using an external quadrupole to store ions prior to injection into the ICR cell.

Complex peptide mixtures from in-gel digestion of yeast proteins were analyzed by microcapillary HPLC (μLC) using a 50 μm × 10 cm capillary column packed with POROS C18 (PerSeptive Biosystems) using a pressure cell (Mass Evolution, Inc., Houston, TX) to slurry pack the column and to load the sample.[24,25]

(20) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1−35.

(21) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. *Mol. Cell. Biol.* **1999**, *19*, 1720−1730.

(22) Laemmli, U. K. *Nature* **1970**, *227*, 680−695.

(23) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.

(24) Lee, N.; Goodlett, D. R.; Marquardt, H.; Geraghty, D. E. *J. Immunol.* **1998**, *160*, 4951−4960.

(25) Mosely, M. A.; Deterding, L. J.; Tomer, K. B.; Jorgenson, J. W. *Anal. Chem.* **1991**, *63*, 1467−1473.

For ESI, the μLC column was connected to a Valco union (Houston, TX) that transferred the column effluent to a New Objective, Inc. (Cambridge, MA) glass needle with a 30-μm tapered tip. Mass spectral data were acquired using an FTICR MS equipped with a 7-T magnet that employed in-trap ion accumulation (i.e., ion injection, pump-down, and excitation/detection) every ~3 s.[26] A new method of calibration was used that enables very high mass measurement accuracy without the requirement for internal calibrants and is described elsewhere.[28] Tandem mass spectral data of protein digests were acquired using an ion trap (ThermoQuest, San Jose) operated in dynamic exclusion mode.[21]

IDEnT Peptide Detection. Peptides alkylated by reaction with 2,4-dichlorobenzyliodoacetamide (IDEnT peptides) were identified either visually or by screening of the data with a computer algorithm that recognized the presence of chlorine in an isotopic distribution. As seen in the modeled data in parts C and D of Figure 2, peptides with masses less than 2000 u had a distinctive isotopic distribution when alkyated with the IDEnT reagent. FTICR spectra were processed with software (i.e., ICR-2LS) developed at Pacific Northwest National Laboratory, to automatically detect and report the masses of all IDEnT-labeled peaks in a spectrum. The experimental distribution was compared with a theoretical distribution for both labeled and unlabeled peptides. A least-squares error was calculated between the experimental and both theoretical distributions. Chlorine-labeled peaks were considered positively identified when experimental isotopic patterns better fit the theoretical chlorine-labeled distributions than the unlabeled distributions.

Protein Identification. Tandem MS spectra of single peptides were analyzed using the software SEQUEST, which generates a list of peptides in a database (e.g., OWL) that match the molecular mass of the unknown peptide on which CID was carried out and then compares the observed CID spectrum of the unknown with that for all possible isobars.[5] A correlation score of 2.0 and a cross-correlation score of 0.1 were set as the minimum reliable values for an initial protein screen. These initial positive protein identifications were confirmed by manually checking the CID data for a fit to the identified sequence.

Proteins were also identified by mass mapping of the data acquired by μLC-ESI-FTICR MS analysis of peptide mixtures. Automated analysis of ESI-FTICR mass spectra using ICR-2LS was coupled with database searching. To identify proteins isolated from *S. cerevisiae*, proteolysis was carried out on all open reading frames according to the rules for protein fragmentation by trypsin (i.e., amide bonds are hydrolyzed leaving lysine or arginine at the new carboxyl terminus). Mass deconvolution used an algorithm called THRASH.[27] The resulting masses were then corrected for the space charge induced frequency shift to increase mass measurement accuracy.[28] The resulting table of neutral masses was then compared to the table of masses consisting of all possible proteolytic (i.e., tryptic) fragments calculated for the entire yeast genome downloaded from the NCBI web site. Program output was structured as a table of experimentally measured masses and predicted peptides within the defined search criteria (e.g., mass measurement error).

For protein identification by accurate mass measurement of a single cysteine-containing peptide, a series of constraints were applied. The mass of each IDEnT-labeled peptide was corrected for the mass of the IDEnT reagent. For the model study, the sequence of the cysteine-containing, laminin B 1 peptide fragment (RYVVLPRPVCFEK) produced by tryptic digestion of laminin B1 peptide was introduced into a database downloaded from the NCBI web site containing all yeast open reading frames. For each IDEnT peptide being studied, a list of all possible peptide isobars to the corrected mass was produced by trypsin digestion of all 6118 yeast proteins. The list of yeast peptides isobaric for each IDEnT peptide allowed for the presence of missed tryptic cleavage sites, but a lysine or arginine at the carboxyl terminus was required. To identify the protein from which a given IDEnT peptide was derived, the following constraints were applied: (1) generate a list of all possible isobaric peptides at an accuracy of 1−2 ppm for each IDEnT peptide with a carboxyl terminal lysine or arginine, (2) retain peptides with at least one cysteine in their sequence (Note: two or more cysteines in an unknown peptide would be distinguished by an enhanced chlorine isotopic pattern and provide a more restrictive constraint than a single cysteine), and (3) retain peptides derived from proteins with a molecular weight that was within ±1000 u of the average mass of proteins found by mass mapping of the same ESI-FTICR MS data set.

RESULTS AND DISCUSSION

Reagent Design and Test. The reagent 2,4-dichlorobenzyliodoacetamide was designed to react with high selectivity to the sulfhyrdyl side chain of the amino acid cysteine on the basis of well-understood alkylation chemistry (Figure 2B).[30] The reagent incorporated chlorine as an isotope distribution encoded tag that was used to detect cysteine-containing peptides within a mixture of peptides predominantly without cysteine.[19,29] Chlorine was chosen as the isotope tag because two naturally occurring isotopes, at 34.969 and 36.966 u, present at abundances of 0.755 and 0.245, provide a isotopic signature readily discernible by mass spectrometry. Since chlorine is not normally found in naturally occurring proteins, it provides a tag with high specificity for cysteine-containing peptides. The isotopic contribution from two atoms of chlorine to the normal peptide isotopic distribution was easily recognized in peptides up to 2000 u (Figure 2C,D), but isotopic modeling was required to confirm the presence of chlorine at higher mass.

Laminin B1 peptide (RYVVLPRPVCFEKGMNYTVR), containing one cysteine and one tryptic cleavage site, served as a model peptide to test the reagent and the method. The peptide was derivatized with a single IDEnT and with high yield (i.e., no unreacted laminin B1 remained). The IDEnT-labeled peptide was treated with trypsin and analyzed by ESI-FTICR MS. Figure 3A shows the ESI-FTICR mass spectrum for the labeled/trypsin-digested laminin B1 and in expanded views the isotope distribution of the unlabeled peptide fragment GMNYTVR (Figure 3B), as well as the IDEnT-labeled peptide fragment RYVVLPRPVCFEK (Figure

(26) Winger, B. E.; Hofstadler, S. A.; Bruce, J. E.; Udseth, H. R.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 566−577.

(27) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, May 31−June4, 1998; p 118.

(28) Bruce, J. E.; Anderson, G. A.; Brands, M. D.; Pasa-Tolic, L.; Smith, R. D. *J. Am. Soc. for Mass Spectrom.*, in press.

(29) Adamczyk, M.; Gebler, . C.; Wu, J. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 1813−1817.

(30) Lundell, N.; Schreitmuller, T. *Anal. Biochem.* **1999**, *266*, 31−47.

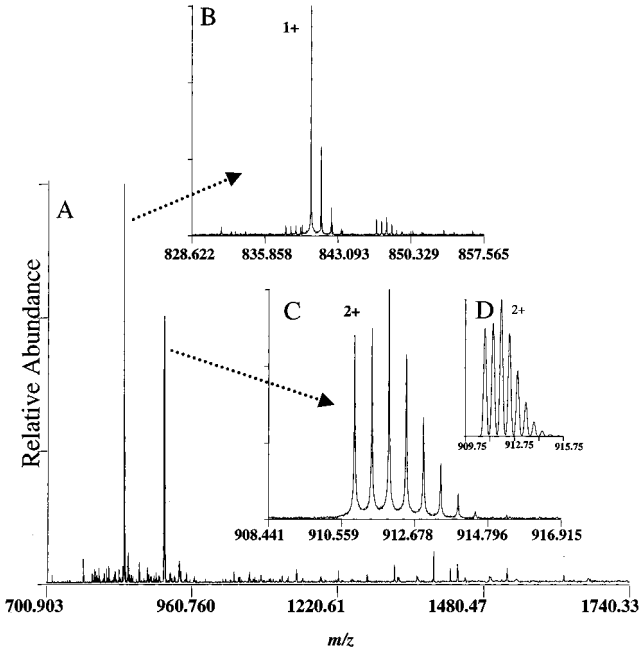

Figure 3. ESI-FTICR mass spectrum of (A) laminin B1 peptide, RYVVLPRPVCFEKGMNYTVR, after reaction with **1** and digestion with trypsin. Expanded *m*/*z* scales show isotope pattern for (B) $[M + H]^{1+}$ ion of laminin B1 tryptic fragment, GMNYTVR, (C) $[M + 2H]^{2+}$ ion of the IDEnT-labeled laminin B1 tryptic fragment, RYVVLPRPVCFEK, and (D) modeled isotopic pattern for peptide in (C) labeled with the IDEnT reagent.

3C). Note that trypsin will hydrolyze neither the arginyl−proline peptide bond nor the amino or carboxyl terminal arginyl−tyrosine or valyl−arginine bonds. This is due to the requirement that the enzyme bind the substrate peptide on both sides of the active site. The presence of chlorine covalently bound to the labeled peptide was readily apparent based on comparison of the observed mass spectrum for the IDEnT-labeled peptide (Figure 3C) and a modeled mass spectrum of the same peptide sequence with the IDEnT label attached (Figure 3D).

Protein Identification. To assess the feasibility of identifying a protein in a database from the accurately measured mass of a single cysteine-containing peptide, the sequence of the IDEnT-labeled laminin B1 tryptic peptide fragment was added to the genomic sequence for *S. cerevisae* which contains 6118 open reading frames (ORF). From these 6118 possible ORFs, 344 855 possible peptide fragments could be generated by complete or incomplete digestion with trypsin. Of these 344 855 peptides, 20 had masses within 2 ppm of the mass measured for the IDEnT-labeled peptide after correction for the weight of the label. This list of 20 peptides indistinguishable by mass measurement alone (Table 1) was reduced to a single and correct answer, the laminin B1 tryptic fragment RYVVLPRPVCFEK, by applying the cysteine constraint. In this case, it was not necessary to consider the molecular weight of the parent proteins to arrive at the correct identification. The protein molecular weight would, however, provide an additional and easily obtainable constraint in cases where proteins separated by gel electrophoresis were analyzed.

Identification of proteins using the accurate mass of a single cysteine-containing peptide was next tested on a complex sample containing proteins from a yeast lysate. However, to validate results from this new approach, identical aliquots of the same

Table 1. Yeast Tryptic Peptides Isobaric with RYVVLPRPVCFEK

protein name[a]	protein mass[b]	peptide isobars	peptide mass	error (ppm)
laminin fragment	2 426.271	***R***YVVLP***R***PV**C**FE***K***	1604.886[c]	0.189
TPI1	26 778.962	FLAS***K***LGD***K***AASEL***R***	1604.889	1.747
YPR143W	52 661.268	D***KKR***T***RK***NAEFG***R***	1604.886	0.07
YDR428C	52 785.179	NLYDAVSNIT***R***LV***K***	1604.889	1.747
TAP42	31 135.369	IELFQ***R***N***K***EIST***K***	1604.889	1.747
CYC2	37 692.826	VQL***K***IFETD***R***QT***K***	1604.889	1.747
MFS1	46 151.311	ESHPVGIL***R***DLIE***K***	1604.889	1.747
YMR291W	64 850.898	EFDLL***R***SISE***K***I***R***	1604.889	1.747
YKR078W	51 708.695	I***R***TAEDEY***R***IVL***K***	1604.889	1.747
TFC6	75 311.526	D***K***IE***R***IYGLN***K***E***K***	1604.889	1.747
SSE1	77 318.483	YLA***K***EEE***KK***QAI***R***	1604.889	1.747
YBR102C	85 484.685	LDEFI***KK***NSD***K***I***R***	1604.889	1.747
STB6	88 779.841	***K***ISADLN***K***IDGLY**R**	1604.889	1.747
FZ01	97 746.957	E***K***NGFNIE***KK***ALS***K***	1604.889	1.747
SEC10	100 279.455	NES***K***IV***KR***VFEE***K***	1604.889	1.747
YLL005C	102 103.872	I***K***ELLFELFYY***K***	1604.885	0.234
S51441	105 161.643	HTVTEL***K***SEIHAL***K***	1604.889	1.747
PEX1	117 202.758	EEV***K***DIIE***R***HLP***K***	1604.889	1.747
RRP5	193 015.955	A***K***D***KKK***VEDLFE***R***	1604.889	1.747
DOP1	194 565.002	LTSSLSPALPAGVHQ***K***	1604.889	1.747

[a] Protein names are as found in the YPD database at www.proteome.com.[34] [b] Calculated protein monoisotopic molecular weight. [c] Observed mass whereas all other masses are calculated monoisotopic.

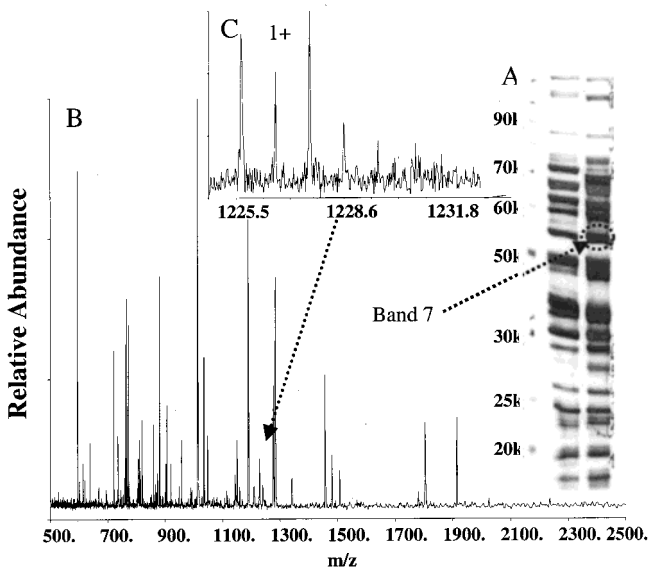

Figure 4. Yeast proteins, alkylated with **1**, separated by SDS gel electrophoresis, visualized by silver staining, and digested by trypsin (A). Example from band 7 of (B) a single acquisition from the μLC-ESI-FTICR MS analysis and (C) a cysteine-containing peptide labeled with the chlorine isotope distribution encoded tag.

sample were first identified by μLC-ESI using data acquired by tandem MS on an ion trap and mass mapping with an FTICR MS. The proteins present in an ion-exchange chromatography fraction of total yeast lysate were separated by gel electrophoresis and the proteins in band 7 (Figure 4A) were analyzed. The tandem MS method identified five proteins with codon bias values ranging from 0.608 to 0.324 (Table 2, column 2). Proteins with calculated codon bias values in this range are expected to be expressed in yeast at a high abundance. Codon bias values indicate the propensity for a gene to utilize the same codon to encode an amino acid even though other codons would insert the identical amino acid into the growing polypeptide chain.[31] In yeast cells, the value

Table 2. Yeast Proteins Identified by Three Different Methods

protein[a]	tandem MS	mass map	IDEnT	codon bias	protein MW[b]
YML056C	+	+	+	0.608	56 357.72
YLR432W	+	+	+	0.599	56 548.73
CYS4	−	+	−	0.444	55 987.36
YHR216W	+	+	−	0.438	56 493.85
YOR374W	+	+	−	0.422	56 688.19
ARO8	+	+	+	0.324	56 142.66
ALD5	−	−	+	0.262	56 585.29
YDR132C	−	−	+	0.040	57 158.03
CLB2	−	−	+	0.004	56 211.64

[a] Protein names are as found in the YPD database at www.proteome.com.[34] [b] Calculated protein monoisotopic molecular weight.

ranges from −0.3 to 1.0 and it is further found empirically that proteins having a large codon bias value (>0.2) are expressed at high levels and proteins expressed at low levels have low codon bias values (<0.1).

Next, a duplicate of band 7 (Figure 4A) was digested with trypsin and the peptide masses measured to ∼1 ppm by μLC-ESI-FTICR MS. The data were analyzed by computer algorithm to identify proteins by mass mapping. Six proteins (Table 2, column 3) were identified including all five of the proteins identified by the data-dependent tandem MS approach. All proteins identified by mass mapping had codon bias values that were >0.3, indicating that these proteins were likely to be expressed at a high abundance in the cell. The proteins identified by this method had a calculated average molecular weight of 56 352.60 ± 191.73 u that was in good agreement with the estimate by gel electrophoresis.

Finally, proteins were identified by the accurately measured masses of single IDEnT-labeled peptides using the same FTICR MS data set used for the mass mapping analysis. Figure 4B shows the complex mixture of ions detected in a single acquisition during the μLC separation of peptides produced by tryptic digestion of the proteins in band 7. All IDEnT-labeled peptides were automatically detected by a computer algorithm that compared the observed isotopic pattern (Figure 4C) with theoretical isotopic patterns for the same mass with and without the IDEnT label. This novel approach identified six proteins (Table 2, column 4), three of which were also identified by the data-dependent tandem MS method and by the mass mapping method. The three additional proteins identified by accurate mass measurement of IDEnT-labeled peptides were YER073W, a protein involved in amino acid metabolism, YDR132C, a protein of unknown function, and CLB2, a protein that participates in signal transduction and known to be phosphorylated and to contain a zinc finger domain. On the basis of codon bias values, YDR132C and CLB2 can be expected to be expressed at low abundance in yeast (Table 2, column 5). Such proteins are not usually identified from complex cell lysates by the data-dependent, tandem MS method, which is one indicator of the better effective sensitivity achieved by using the accurately measured masses of single, IDEnT-labeled peptides to identify proteins.[21,32]

CONCLUSIONS

The labeling of peptides with a nonnatural isotope like chlorine allows them to be readily distinguished from other peptides in a mixture due to the distinctive isotope distribution of chlorine. By designing a reagent containing chlorine that also has a high chemical selectivity for the sulfhydryl side chain of cysteine, peptides containing cysteine can be confidently distinguished in a mixture. The presence of cysteine in the sequence provides a highly constraining search parameter that allows the accurate mass of a single cysteine-containing peptide together with other constraints such as parent protein molecular weight to be used to identify proteins in a database. This novel approach identified more proteins than the data-dependent, tandem MS approach alone (Table 2, column 2). Furthermore, if the FTICR MS data set is analyzed as a whole by combining the mass mapping and IDEnT approaches (Table 2, columns 3 and 4), then nine proteins were identified versus only five by the tandem MS approach alone. None of the proteins identified by the tandem MS approach were missed by such an analysis.

More important though than identification of a large number of proteins is the fact that the method allowed the identification of low-abundance proteins normally missed by a tandem MS method because of the effective loss of dynamic range due to data-dependent analysis. While the two low-abundance proteins identified by the accurate mass approach cannot be confirmed by the two standard methods, there is no reason to expect the method to fail at low signal-to-noise ratios as long as the mass can be accurately assigned. Identifying proteins using FTICR MS and thus avoiding CID provided the expected advantages, namely, (1) high resolution (>10 000 at 1000 u) to analyze more complex mixtures than possible with other instruments (the ion trap data set was notable in that only one IDEnT-labeled peptide was observed)[17] and (2) mass accuracy of <1.0 ppm, allowing more proteins to be identified by better utilizing more of the information in the data set. Such an approach should aid the rapidly expanding field of proteomics where the number of complete genomes is growing and thus the list of expected peptide masses for a given proteome available.[1,2,3,3333−34] It should be noted that identification of proteins by accurate peptide mass is also amenable to searching for posttranslational modifications such as phosphorylation.

Finally, the use of an IDEnT reagent has as its basic tenant the ability to distinguish an analyte encoded with a non-native isotope from those not encoded with the isotope or from those encoded with a different isotope by virtue of their distinctive isotopic signatures in a mass spectrometer.[19,29] This allows the isotopically encoded analytes to be easily identified in a mass spectrometer and provides advantages in a number of applications other than the one presented here for protein identification by sequence database searching. The design of IDEnT reagents with target specificities other than cysteine will be implemented for applications such as de novo peptide sequencing by tandem MS, determination of juxtaposed neighboring groups in cross-linking studies and identification of active sites in enzymes. Many of these

(31) Bennetzen, J. L.; Hall, B. D. *J. Biol. Chem.* **1982**, *257*, 3026−3031.

(32) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.

(33) Wilkins, M. R.; Pasquali, C.; Appel, R. D.; Ou, K.; Golaz, O.; Sanchez, J.-C.; Yan, J. X.; Gooley, A. A.; Hughes, G.; Humphrey-Smith, I.; Williams, K. L.; Hochstrasser, D. *Biotechnology* **1996**, *14*, 61−65.

(34) Hodges, P. E.; McKee, A. H.; Davis, B. P.; Payne, W. E.; Garrels, J. I. *Nucleic Acids Res.* **1999**, *26*, 69−73.

applications will not require high mass accuracy, but simply mass resolution high enough to decode the incorporated IDEnT. The increasing availability of mass spectrometers capable of high resolution and high mass accuracy such as FTICR MS should aid the development of IDEnT reagents in many other areas of macromolecular chemistry and mixture analysis.

ACKNOWLEDGMENT

This work was supported by a grant from the Merck Genome Research Institute (MGRI) to R.A. and by the NSF Science and Technology Center for Molecular Biotechnology. B.R. was the recipient of a fellowship from the Swiss National Science Foundation. O.F. thanks the Max-Planck-Institute of Molecular Plant Physiology for sabbatical support. J.E.B., G.A.A., L.P.-T., and R.D.S. thank the U.S Department of Energy, Office of Biological and Environmental Research, and NINDS through Grant NS39617 for support of this research. The Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy through Contract DE-AC06-76RLO 1830.

Received for review November 17, 1999. Accepted January 13, 2000.

AC9913210