Laboratory for Computational Proteomics

www.FenyoLab.org

E-mail: Info@FenyoLab.org

Facebook: *NYUMC Computational Proteomics Laboratory*

Twitter: *@CompProteomics*

Anal. Chem. **1997,** *69,* 3995−4001

A Strategy for Rapid, High-Confidence Protein Identification

Jun Qin,[†] David Fenyö,[†] Yingming Zhao,[†] William W. Hall,[‡] David M. Chao,[§] Christopher J. Wilson,[§] Richard A. Young,[§] and Brian T. Chait*[,†]

The Rockefeller University, 1230 York Avenue, New York, New York 10021, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142 and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and University College Dublin, Belfield, Dublin 4, Ireland

A procedure is described for rapid, high-confidence identification of proteins using matrix-assisted laser desorption/ionization tandem ion trap mass spectrometry in conjunction with a genome database searching strategy. The procedure involves excision of copper-stained bands or spots from electrophoretic gels, in-gel trypsin digestion of the proteins, single-stage mass spectrometric analysis of the resultant mixture of tryptic peptides, followed by tandem ion trap mass spectrometric analysis of selected individual peptides, and database searching of the relevant genomic database using the program PepFrag. The scheme provides sensitive, real-time protein identification as well as facile identification of modifications. A single operator can unambiguously identify 5−10 proteins/day from an organism whose genome is known at a level of >0.5 pmol of protein loaded on a gel. The utility of the technique was demonstrated by the identification and characterization of a band from a human HTLV-I preparation and 11 different proteins from a yeast RNA polymerase II C-terminal repeat domain-affinity preparation. The technology has great potential for postgenome biological science, where it promises to facilitate the dissection and anatomy of macromolecular assemblages, the definition of disease state markers, and the investigation of protein targets in biological processes such as the cell cycle and signal transduction.

Genome sequencing projects are producing an unprecedented information resource for biologists. Efficient utilization of this remarkable resource demands the development of new tools for rapidly analyzing mature proteins and for correlating them with their genes and ultimately their functions. One particularly powerful new set of tools for rapidly identifying and characterizing proteins utilizes mass spectrometric techniques such as matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) mass spectrometry (MS) in combination with genome database searching strategies.[1−15]

Two general strategies have been developed for identifying proteins by MS. In both, the proteins of interest are separated (e.g., by gel electrophoresis) and individually subjected to proteolysis with an enzyme of known specificity (e.g., trypsin), and the molecular masses of the resulting peptides accurately and rapidly determined by MS. In the first strategy, these experimentally determined masses (i.e., the tryptic map) are compared with the calculated masses of all tryptic peptides that can be theoretically produced from sequences corresponding to all of the proteins in the genomic database of the organism under study.[1−5,12] The protein yielding the best match between the experimental and theoretical peptides is identified. MALDI time-of-flight MS has been the preferred technique for this peptide mapping approach because it is sensitive and allows for the measurement of the component peptides resulting from the digest without prior HPLC separation or extensive cleanup. Although this method is fast and simple, its success can be compromised by the presence of more than one protein in the gel spot or by extensive posttranslational modifications of the protein of interest and errors in the database sequence. In addition, the observation of too few peptides in the MS map from a given protein may preclude its identification.

A second strategy, involving tandem mass spectrometry (MS/MS),[16] has been developed to circumvent these difficulties.[6−15] Here, a particular tryptic peptide is selected and dissociated in the mass spectrometer to produce a fragmentation mass spectrum

* E-mail: chait@rockvax.rockefeller.edu.

[†] The Rockefeller University.

[‡] University College Dublin.

[§] Whitehead Institute for Biomedical Research and Massachusetts Institute of Technology.

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

(2) Pappin, D. D. J.; Hijrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−332.

(3) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338−345.

(4) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64; *Protein Sci.* **1994**, *3*, 1347−1350.

(5) Yates, J. R., III; Speichner, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397−408.

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

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

(8) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426−1436.

(9) Griffin, P. R.; MacCoss, M. J.; Eng, J. K.; Blevins, R. A.; Aaronson, J. S.; Yates, J. R., III *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1546−1551.

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

(11) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466−469.

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

(13) Figeys, D.; Ducret, A.; Yates, Y. R., III; Aebersold, R. *Nat. Biotechnol.* **1996**, *14*, 1579−1583.

(14) Neubauer, G.; Gottschalk A.; Fabrizio, P.; Seraphin, B.; Luhrmann, R.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 385−390.

(15) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *Proc. Nat. Acad. Sci. U.S.A.* **1996**, *93*, 14440−14445.

(16) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad Sci U.S.A.* **1986**, *83*, 17, 6233−6237.

S0003-2700(97)00488-5 CCC: $14.00

that is characteristic of the sequence of the peptide. The database search for the protein uses the molecular mass (M_r) of the tryptic peptide together with its fragmentation spectrum. Although the M_r of the peptide by itself only moderately constrains the search—and can lead to large numbers of possible proteins—a good match of the fragmentation spectrum often identifies a unique protein. The identification can be verified by checking how many of the remaining tryptic peptides have measured masses that are in accord with hypothetical tryptic peptides from the putative protein. To further increase the confidence of the call, fragmentation mass spectra of one or more additional tryptic peptides are obtained. Because protein identification incorporating mass spectrometric fragmentation and database searching requires only single (or at most a few) tryptic peptides from any given protein, this strategy can confidently identify multiple proteins in mixtures[14,17] and is highly tolerant of posttranslational modifications or errors in the database. Nanospray ionization[18] (i.e., low flow rate electrospray ionization) combined with triple-stage quadrupole MS has been shown to work well for this latter strategy.[10,11,13−15] Alternatively, postsource decay in MALDI time-of-flight MS analysis[19] has been applied for this purpose.[20−24]

Although protein identification by simple peptide mapping using MALDI time-of-flight MS is sensitive and fast, the confidence level of the identification may be insufficient for an assured call. The additional constraints obtained through ESI tandem MS provides more reliable identification than simple peptide mapping but suffers from reduced sample throughput. Alternatively, additional constraints may be obtained through the use of postsource decay in MALDI time-of-flight mass spectrometry.[19−24]

In this paper, we present a procedure that combines the positive virtues of the above two strategies for the identification of proteins. This procedure utilizes unique properties of our newly developed MALDI ion trap mass spectrometer[25−27] and provides sensitive, confident protein identification with high throughput.

EXPERIMENTAL SECTION

Isolation of HTLV-1 Proteins. The HTLV-I transformed lymphocyte cell (N5-CR) was maintained in RPM I medium containing 10% fetal calf serum. Supernatants forming large-scale cultures (5000 mL) were clarified by centrifugation at 60000*g* for 90 min. Pellets were resuspended and centrifuged on linear 25−65% sucrose gradients. Purified virus bands were collected and pelleted. The latter were resuspended in NTE buffer (0.01 M Tris-HCl pH 7.4, 0.001 M EDTA, 0.1 M NaCl).

Isolation of CTD-Binding Proteins. Details of the isolation of the CTD-binding proteins are described in ref 28.

Procedures for Sample Preparation. Protein samples were resuspended in 1× Bio-Rad Tris−glycine sample buffer (Bio-Rad Laboratories, Hercules, CA) plus 1/10 vol of 10% (w/v) SDS solution. The proteins were separated on a 4−15% gradient Bio-Rad Ready Gel and electrophoresed at 200 V. After protein separation, the gel was soaked in deionized H_2O for 1 min and then 1× Bio-Rad copper stain solution for 5 min, with constant shaking. The stained gel was washed for 1 min in H_2O. Because the protein bands are negatively stained, the gel appears opaque greenish blue and the protein bands relatively clear. The bands are most easily visualized on a black background.

The protein bands were cut out using a stainless steel scalpel and transferred to 0.5 mL Eppendorf tubes. To each tube was added 0.4 mL of 1× Bio-Rad copper destain solution (Bio-Rad Laboratories), and the tubes were vortexed for 5 min. After discarding the wash liquid, this step was repeated for 2 min. By this stage, the gel pieces have turned from faint greenish blue to clear. H_2O (0.4 mL) was added, the mixture vortexed for 1 min, and the liquid discarded. Digestion buffer (0.4 mL 50 mM Tris-HCl, pH 8.0) was added, the mixture vortexed for 1 min, and the liquid discarded. Modified trypsin (20 μL of 25 ng/μL in 50 mM Tris-HCl, pH 8.0, Boeringer Mannhein, Indianapolis, IN) was added to each tube (i.e., to each gel piece). The gel pieces were squashed with a plastic pipet tip that had been sealed closed using heat from an open flame. Proteins in the squashed gel pieces were digested for 2 h at 37 °C.

Extraction of the peptides from the gel pieces was facilitated by sonication for 3 min, followed by removal of the liquid with a gel-loading tip. The liquid was transferred to a fresh Eppendorf tube, taking care not to inadvertently transfer any small gel pieces into the tube. An extraction solution (20 μL of ACN−0.5% TFA/H_2O 1:1 (v/v)) was added to the original crushed gel pieces, the mixture sonicated for 3 min, and the liquid removed and pooled with the first extract. Finally, 10 μL of 100% ACN was added to the crushed gel pieces, the mixture sonicated for 2 min, and the liquid removed and pooled with the first two extracts. The pooled solutions were evaporated to dryness (SpeedVac, Savant, Farmingdale, NY) at medium heat. For the MS measurements, the dried samples were redissolved in 5−10 μL of acetonitrile−0.5% TFA/ H_2O 1:1 (v/v).

Matrix-Assisted Laser Desorption Ion Trap MS. The design and performance of our custom MALDI ion trap mass spectrometer has been described.[25−27] It is composed of an external MALDI ion source and a modified Finnigan ITMS electronics kit. Laser desorption/ionization was carried out at wavelength of 355 nm with 10 ns duration pulses. MS and MS/MS spectra were taken as described previously.[25] The matrix solution consisted of 2,5-dihydroxybenzoic acid (DHB) in 1:1 (v/v) ACN/H_2O (2× dilution of a saturated solution of DHB in 1:1 (v/v) ACN/H_2O). The MS samples were prepared by mixing on the sample probe 1 μL of sample solution with 1 μL of matrix solution. The instrument was mass calibrated once a week.

Procedure for Protein Identification. (1) The peptide mixture produced by in-gel trypsin digestion of a protein was analyzed directly by MALDI ion trap MS without prior chromatographic separation or further treatment. (2) After the MS peptide

(17) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III *Anal Chem.* **1997**, *69*, 767−776.
(18) Wilm, M.; Mann, M. *Anal. Chem.* **1996**, *68*, 1−8.
(19) Kaufmann, R.; Spengler, B.; Lutzenkirchen, F. *Rapid Commun. Mass Spectrom.* **1993**, *7*, 902−910.
(20) Griffin, P. R.; MacCoss, M. J.; Eng, J. K.; Blevins, R. A.; Aaronson, J. S.; Yates, J. R., III *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1546−1551.
(21) Patterson, S. D.; Thomas, D.; Bradshaw, R. A. *Electrophoresis* **1996**, *17*, 877−891.
(22) O'Connell, K. L.; Stults, J. T. *Electrophoresis* **1997**, *18*, 349−359.
(23) Larsson, T.; Norbeck, J.; Karlsson, H.; Karlsson, K. A.; Blomberg, A. *Electrophoresis* **1997**, *18*, 418−423.
(24) Matsui, N. M.; Smith, D. M.; Clauser, K. R.; Fichmann, J.; Andrews, L. E.; Sullivan, C. M.; Burlingame, A. L.; Epstein, L. B. *Electrophoresis* **1997**, *18*, 409−417.
(25) Qin, J.; Steenvoorden, R. J. J. M.; Chait, B. T. *Anal. Chem.* **1996**, *68*, 1784−1791.
(26) Qin, J.; Chait, B. T. *Anal. Chem.* **1996**, *68*, 2102−2107.
(27) Qin, J.; Chait, B. T. *Anal. Chem.* **1996**, *68*, 2108−2112.
(28) Wilson, C. J.; Mann, M.; Imbalzano, A. N.; Schnitzler, G. R.; Kingston, R. E.; Young, R. A. *Cell* **1996**, *84*, 235−244.

map was inspected, one peptide ion species was isolated and fragmented by collision-induced dissociation to obtain an MS/MS spectrum. (3) The masses of the precursor and fragment ions were searched against a database using the program PepFrag[29] (see below), and candidate proteins were identified. The search was carried out with constraints that include the cleavage specificity of the digesting enzyme, the originating species of the protein, and the systematics of MALDI ion trap collision-induced dissociation of peptides.[30,31] (4) Other peptides in the measured MS peptide map were assigned to the identified candidate protein. (5) Of these newly assigned peptides, one was chosen and fragmented, and the fragmentation pattern checked against the candidate protein sequence to verify the identification. (6) For those peptides that could not be assigned to the identified protein, one peptide was selected and fragmented, and steps 3−5 were repeated until the majority of intense peaks were assigned. Steps 1−6 were all performed in real time.

The Protein Identification Program PepFrag. The program PepFrag,[29] which was developed in our laboratory, allows for the searching of protein or nucleotide sequence databases (SWISS-PROT, PIR, GENPEPT, OWL, or dbEST) using a combination of information from MS peptide maps and MS/MS spectra of proteolytic peptides. The databases have been taxonomically divided to allow for faster searches and to minimize the number of unrelated hits. The experimental conditions (enzyme specificity, approximate protein mass, place in phylogenic tree of the species, and modifications of amino acids) can be specified in the search. The result of the search is a list of proteins, each of which contains a peptide that matches the measured mass of a proteolytic peptide as well as the measured masses of MS fragments of the peptide. In addition, other search constraints can be specified, if such information is available. These constraints include specification of the MS fragmentation systematics, masses of other proteolytic peptides that are assumed to belong to the protein of interest, and partial amino acid composition. PepFrag is publicly available over the Internet at URL http://chait-sgi.rockefeller-.edu.[29]

RESULTS AND DISCUSSION

Preferential Cleavage of Peptide Ions in Tandem MALDI Ion Trap MS: (1) A Highly Effective Constraint for Protein Identification and (2) High Sensitivity. We found previously that fragmentation of peptide ions by tandem MALDI ion trap MS is highly selective and that Arg-containing peptides undergo facile, preferential cleavage adjacent to amino acid residues with acidic side chains,[25,30,31] producing exclusively b- and/or y-type ions.[32,33] Lys-containing peptides also undergo preferential fragmentation adjacent to Asp/Glu although the selectivity is not as high as for Arg-containing peptides.[31] Because most tryptic peptides contain an Arg/Lys residue at the C-terminal, we have included such preferential cleavages as a selectable constraint in our search program, PepFrag[29] (see Experimental Section) and have found that application of this constraint greatly facilitates the unambiguous identification of proteins. For example, in an experiment designed to test the present methodology, we separated proteins from a human T cell leukemia virus type 1 (HTLV-I) preparation by SDS−PAGE and obtained a MS tryptic peptide map (Figure 1a) for a prominent band with an apparent M_r of ∼22 kDa (Experimental Section). Some 20 peaks appear in the m/z range between 1200 and 2800 (Figure 1a). To identify the protein in the band, we obtained an MS/MS spectrum of the ion with m/z 2147.6 Da (Figure 1b). Two product ions that arise from preferential cleavage at the C-termini of Asp/Glu residues, dominate the MS/MS spectrum (in addition to noninformative fragment ions produced by the facile loss of small neutral molecules, e.g., H_2O, NH_3, and CO_2). We therefore input into PepFrag (Figure 1c) the M_r of the precursor peptide and the m/z values of the two informative fragment ions, together with the constraint that the fragments were b- and/or y-type generated at the C-terminal of Asp/Glu residues. We also specified the type of enzyme used (trypsin) and allowed for partial enzymatic degradation of the protein (with up to two internal Arg/Lys residues retained in the fragments). We did not specify the species from which the protein originated to allow for the possibility of adventitious impurities from the host cell and did not constrain the search with the apparent SDS−PAGE M_r to allow for the possibility of proteolytic processing of the protein.

The search results (using the program PepFrag) summarized in Figure 1c identify a single gene product—the gag polyprotein of HTLV-I (albeit from three different strains of the virus). The specificity of the cleavage reaction is seen to provide a highly effective constraint for protein identification. The results are especially impressive considering that the search was carried out for all proteins in the SwissProt database using a conservative mass tolerance (±2 Da) for both the precursor and product ions. We have subsequently found, through an analysis of more than 200 different protein bands, that it is nearly always possible to identify a protein (whose sequence is present in the database) with just two fragment ions generated at the C-termini of Asp/Glu in a single tryptic peptide and that it is usually possible to identify a protein with more than four fragments generated at unspecified amino acid residues. In the latter case, the fragment ions need not constitute a sequence tag[7] as long as they are b- and/or y-type ions—i.e., gaps are well tolerated. Because b- and y-type ions dominate the MALDI ion trap MS/MS spectra,[25,30,31] the intense fragmentation peaks normally correspond to these ion species. One potentially complicating factor in the interpretation of the fragmentation spectra is the occurrence of b*- and y*-type fragment ions—i.e., b and y ions that have undergone loss of a H_2O/NH_3 moiety. This additional fragmentation produces little ambiguity in practice because its occurrence is readily recognized by the presence of pairs of fragment peaks spaced 17−18 Da apart (see, e.g., the pair at m/z 1975.9 and 1959.1).

The average abundance of Asp/Glu residues in proteins is ∼12%.[34] Thus, tryptic peptides with mass of >1000 Da often contain an Asp/Glu residue and the probability for the presence of Asp/Glu increases as a function of increasing peptide mass. For the purpose of protein identification, we therefore find it advantageous to obtain MS/MS spectra of tryptic peptide ions with higher masses ($1500 < m/z < 3500$). We have found that singly protonated peptide ions in this mass range can be efficiently fragmented in the MALDI ion trap.[25] An added benefit of

(29) Fenyö, D. http://chait-sgi.rockefeller.edu 1996. Fenyö, D.; Zhang, W.; Chait, B. T.; Beavis, R. C. *Anal. Chem.* **1996**, *68*, 721A−726A.
(30) Qin, J.; Chait, B. T. *J. Am. Chem. Soc.* **1995**, *117*, 5411−5412.
(31) Qin, J. Ph.D. Thesis, Rockefeller University, 1996.
(32) Roepstorff, P.; Fohlman, J. *J. Biomed. Environ. Mass Spectrom.* **1984**, *11*, 601.
(33) Biemann, K. Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Methods Enzymol.* **1990**, *193*, 886−887.
(34) McCaldon, P.; Argos, P. *Proteins* **1988**, *4*, 99−122.

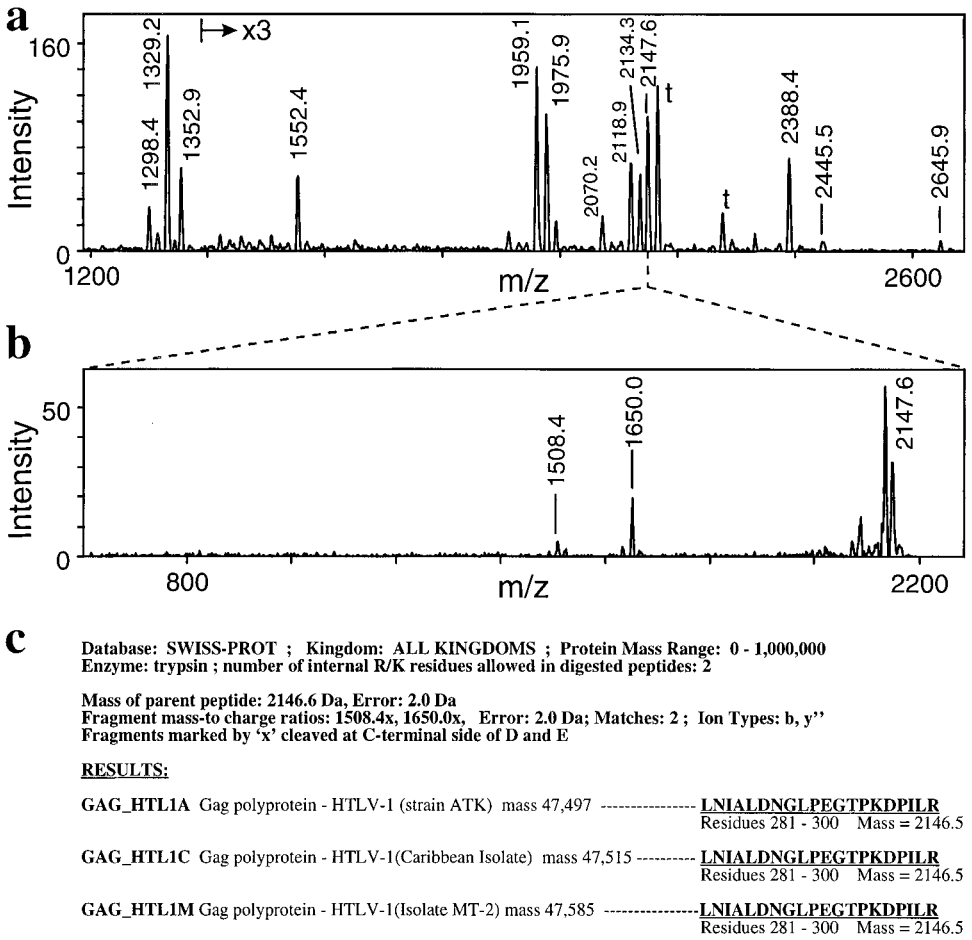

Figure 1. (a) MALDI ion trap mass spectrum of the products of in-gel trypsin digestion of the 22 kDa SDS−PAGE protein band from the HTLV-I preparation (Experimental Section). (b) Tandem MALDI ion trap mass spectrum of the peptide ion with *m/z* 2147.6. (c) Protein identification search results obtained with the program PepFrag using the information in Figure 1b. The protein is identified as the gag polyprotein from HTLV-I.

analyzing higher mass tryptic peptide ions derives from their reduced statistical occurrence relative to lower mass tryptic peptides.

Selective fragmentation at Asp/Glu also significantly enhances the sensitivity of the MS/MS measurement. Because relatively few dissociation channels are open, signal dilution effects normally experienced in ESI triple quadrupole or MALDI postsource decay tandem MS is avoided. This concentration of fragmentation is particularly important when only small amounts of sample are available—i.e., where the number of ions available for fragmentation is severely limited. It can be seen from Figure 1b that dilution of the fragmentation spectrum into additional channels would quickly compromise our ability to observe the fragmentation peaks. The selectivity at Asp/Glu appears especially strong for singly charged ions (unpublished observations) and is most obvious in the slow (ms) decomposition measurements made in the ion trap. Finally, in contrast to ESI, MALDI yields mainly singly charged peptide ions, leading to an additional reduction in signal dilution losses and easier mass spectral interpretation.

Protein Identification with MALDI Tandem Ion Trap MS with High Confidence and High Throughput. The protein identification procedure outlined in the Experimental Section ensures the identification of proteins with a high level of confidence. After the protein is tentatively identified using the MS/MS data (Figure 1b), the measured peptide map is compared with the map calculated for this putative protein and an attempt

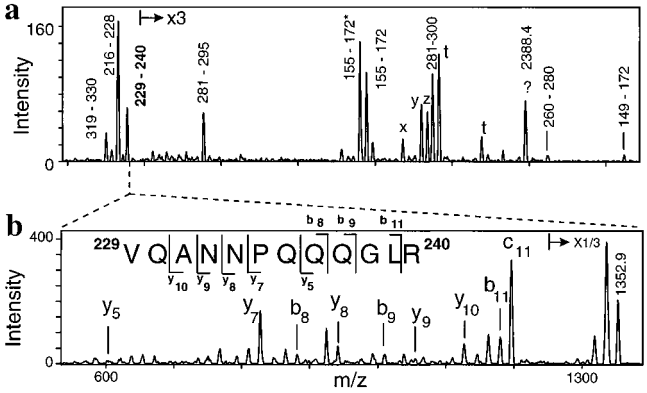

Figure 2. (a) Assignment of the observed MS peaks to peptides from the identified gag polyprotein of HTLV-I. (b) Tandem MALDI ion trap mass spectrum of the peptide ion with *m/z* 1352.8, confirming its assignment as gag 229−240 and the progenitor protein as gag.

is made to assign the various peptide peaks (Figure 2a). Ions that cannot be assigned could arise either from other proteins present in the band or from modifications of the already identified protein. In cases where most of the ions can be assigned to the putative protein, the likelihood is high that the identification is correct. However, to further increase the confidence level of the call, we test the hypothesis that the protein has been correctly identified by obtaining an additional MS/MS spectrum from a second peptide that has been assigned to the putative protein

(Figure 2b). Two MS/MS spectra plus a peptide map normally yield an unambiguous identification.

Our utilization of copper staining/destaining (see Experimental Section) and the elimination of the need for chromatographic separation make the sample preparation procedure fast. The whole process—from staining the gel to the point when samples are inserted into the mass spectrometer—requires < 4 h. For convenience, we typically process 10 proteins at a time. The initial MS and first MS/MS measurements require <30 min. Another 30 min is needed to search the database, to interpret the data, and to obtain the second MS/MS spectrum. In this way, protein identification is performed at an average rate of 1 protein/h by a single operator. Such high throughput is crucial for large-scale biological research projects (see below). The present protocols allow us to unambiguously identify real-world protein samples at a rate of 5−10/day from genomes that have been fully sequenced (e.g., *Saccharomyces cerevisiae*).

Protein Identification with MALDI Tandem Ion Trap MS in Real-Time. The extremely low sample consumption and pulsed nature of MALDI allow us to stop the experiment at any time, analyze the data in real-time, and design the next experiment by following leads provided by the previous experiment(s). It is not necessary to attempt to obtain MS/MS spectra for all peptide ions in the map as may be optimum for a continuous ionization technique like electrospray or nanospray, where the spray time is limited. The ability to continue to measure a single sample for as long as is needed to obtain an unambiguous result is crucial for the success of the present iterative protein identification procedure. For example, after we obtained the peptide map shown in Figure 1a, we took the MS/MS spectrum of the peptide with M_r 2146.6 Da (Figure 1b) and stopped the experiment to search the database. The search (<1 min) identified the gag protein (Figure 1c). After assigning several of the observed peptides in the peptide map to the gag protein (Figure 2a), we resumed the experiment and took another MS/MS spectrum of the assigned peptide with M_r 1351.9 Da to confirm the putative identification (Figure 2b).

To assure that we have identified all of the abundant proteins present in a band, we attempt to assign every intense peak in the peptide map. Inspection of Figure 2a shows that we were unable to assign several of the peaks to the gag protein, including the ion with m/z 2118.9. To investigate the origin of this ion, we obtained its MS/MS spectrum (data not shown). The subsequent database search failed to return an identification. The failure to identify a protein indicates that the peptide may be posttranslationally modified. However, no suggestion of the most commonly occurring posttranslational modifications (including phosphorylation and glycosylation) was apparent from the MS and MS/MS spectra (see below). We thus deduced that the protein may have been processed proteolytically. A relaxed search allowing nonspecific processing returns a peptide that belongs to the same gag gene product but is proteolytically processed between Leu-130 and Pro-131. This finding is in concert with in vivo processing of the HTLV-1 gag polyprotein, which has been observed to yield three mature proteins—i.e., gag P19 (residues 1−130), gag P24 (residues 131−344), and gag P15 (residues 345−429).[35−37] The peptide with m/z 2118.9 arises from the processed N-terminal peptide of gag P24.

Amino Acid Modifications Identified by MALDI Ion Trap MS. Posttranslational modifications or chemical modifications introduced in gel electrophoresis can be problematic for protein identification because modification information is not inherent to the DNA sequence database. We have found that commonly occurring posttranslational modifications such as phosphorylation and glycosylation can be readily identified in MALDI ion trap MS because phosphopeptides and glycopeptides each have clear signatures in the MALDI ion trap mass spectra.[31,38] Protonated phosphopeptides undergo facile loss of ∼98 Da in single-stage MALDI ion trap MS so that peaks separated by ∼98 Da indicate the presence of a phosphopeptide. The presence of a suspected phosphopeptide can be readily confirmed by acquiring its MS/MS spectrum. The observation of a dominant product ion with a mass ∼98 Da less than the precursor ion confirms the presence of a phosphate group on the peptide. A similar phenomenon occurs for glycopeptides, which readily dissociate at the glycosidic bonds, generating a series of ions in the single-stage spectrum with mass differences of 162 (Hex), 203 (HexNAc), or 291 Da (sialic acid). The observation of these ions indicates the presence of a glycopeptide. Again, the hypothesis can be readily tested by MS/MS.

A commonly occurring chemical modification is the oxidation of Met during electrophoresis or storage of the dried gels. Peptide ions containing oxidized Met readily lose the methyl sulfoxide moiety in the ion trap to give a signature pair of peaks 64 Da apart.[38,39] If the Met residue is not completely oxidized to the sulfoxide, a triplet of peaks can be observed separated by −48 and +16 Da from the unmodified peptide ion. In Figure 2a, the ions at m/z 2070.2 (labeled x), 2117.8 (labeled y), and 2134.3 (labeled z) constitute such a triplet. Again, the peptide containing oxidized Met can be readily verified by an MS/MS experiment, in which the product spectrum shows a dominant product ion with a mass 64 Da less than the precursor ion (data not shown). The ability to readily identify modifications of proteins by MALDI ion trap MS further increases the success rate and confidence of protein identification.

Protein Identification with MALDI Tandem Ion Trap MS: A Rapid and Effective Tool for Identifying Protein Components of Complex Macromolecular Assemblages. We have applied the above-described methodology to identify proteins interacting with the RNA polymerase II C-terminal repeat domain (CTD) (reviewed in ref 40). The CTD has been shown to interact with components of the RNA polymerase II holoenzyme.

A highly purified preparation of CTD-binding protein complex was prepared by a combination of CTD affinity and ion exchange chromatography. Protein components of this CTD-binding fraction were separated by SDS−PAGE, subjected to proteolysis with trypsin, and analyzed by MS. As an example, Figure 3a shows the MS tryptic map from a gel band that migrated with an apparent M_r of 35 kDa. The MS/MS spectrum of the m/z 2445.7 ion (Figure 3b) unambiguously identifies its progenitor protein as Srb5, a component of a functional preinitiation complex that has

(35) Seiki, M.; Hattori, S.; Hirayama, Y.; Yoshida, M. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 3618−3622.
(36) Oroszlan, S.; Sarngadharan, M. G.; Copeland, T. D.; Kalyanaraman, V. S.; Gilden, R. V.; Gallo, R. C. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 1291−1294.
(37) Copeland, T. D.; Oroszlan, S.; Kalyanaraman, V. S.; Sarngadharan, M. G.; Gallo, R. C. *FEBS Lett.* **1983**, *162*, 390−395.
(38) Qin, J.; Chait, B. T. *Anal. Chem.* **1997**, *69*, 4002−4009.
(39) Jiang, X.; Smith, J. B.; Abraham, E. C. *J. Mass Spectrom.* **1996**, *31*, 1309−1310.
(40) Koleske, A. J.; Young, R. A. *Trends Biochem. Sci.* **1995**, *20*, 113−116.

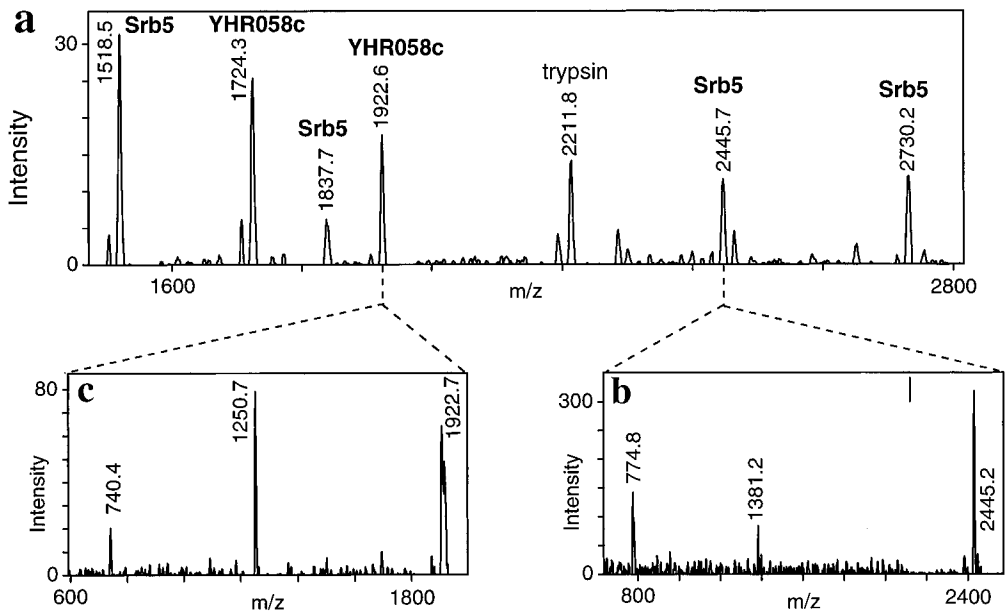

Figure 3. (a) MALDI ion trap mass spectrum of the products of in-gel trypsin digestion of an electrophoretic band with an apparent M_r of 35 kDa from a purified preparation of yeast RNA polymerase II CTD-binding proteins (Experimental Section). (b) Tandem MALDI ion trap mass spectrum of the peptide ion with *m/z* 2446, identifying its progenitor protein as Srb5. (c) Tandem MALDI ion trap mass spectrum of peptide ion with *m/z* 1923, identifying its progenitor protein as the open reading frame yielding the gene product YHR058c.

Table 1. List of Identified Proteins

M_r				
PAGE	calcd	gene name	NCBI identifiers	protein description
97	78.5	SRB4	172693	RNA polymerase II supressor protein Srb4
35	34.3	SRB5	172536	RNA polymerase II supressor protein Srb5
200	191.0	RPB1	1419221	RNA polymerase II subunit 1
130	139.0	RPB2	1293711	RNA polymerase II subunit 2
114	123.3	RGR1	218472	glucose repression regulatory protein Rgr1
58	54.8	ACT3	436808	actin-like protein Act3
56	60.4	RRN7	567927	RNA polymerase I specific initiation factor Rrn7
31	25.2	YBR193c	311669	protein of unknown function
35	32.8	YHR058c	487956	protein of unknown function
70	64.2	YPR070w	805050	protein of unknown function
150	110.0	YDR359c	849181	protein of unknown function

previously been found to be required for efficient transcription initiation.[41] Although three other peaks in the peptide map were also identified as arising from Srb5 (Figure 3a), the peaks at *m/z* 1922.6 and 1724.3 could not be assigned to this protein. MS/MS of the ion at *m/z* 1923 (Figure 3c) identified the presence of a second protein in the band—i.e., the open reading frame yielding the hypothetical gene product YHR058c. This identification was confirmed by MS/MS of a second peptide at *m/z* 1724.3 from the same gene product. This example demonstrates that two proteins can be readily identified in a single electrophoretic band, even when relatively few peptides are available for MS/MS analysis.

In this manner, 11 proteins were identified over a period of two days by a single operator (Table 1). Five (RPB1, RPB2, RGR1,[42] SRB4, SRB5) are known components of the RNA polymerase II holoenzyme, confirming the efficacy of the technique. Proteins corresponding to four open reading frames (with unknown function) and ACT3 and RRN7 were also identified (Table 1). ACT3 is an actin-related protein for which genetic evidence suggests a role in transcriptional regulation.[43] The identification of ACT3 as a CTD-binding protein provides biochemical evidence for a link between class II transcription and actin-related functions. The identification of RRN7, a regulatory protein known to be involved in class I transcription,[44] suggests a common regulatory mechanism that is conserved between class I and class II transcription. Further work will be required to test these hypotheses.

CONCLUSIONS

We have devised a procedure for identifying proteins that combines the robustness, simplicity, and high sensitivity of MALDI-MS and the specificity and efficiency of ion trap tandem MS. The scheme provides fast, sensitive, high-confidence real-time protein identification as well as facile identification of modifications. A single operator can currently identify unambiguously 5−10 proteins/day (from an organism whose genome has been sequenced) at a sensitivity level of >0.5 pmol of protein

(41) Thompson, C. M.; Koleske, A. J.; Chao, D. M.; Young, R. A. *Cell* **1993**, *73*, 1361−1375.
(42) Li, Y.; Bjorklund, S.; Jiang, Y. W.; Kim, Y. J.; Lane, W. S.; Stillman, D. J.; Kornberg, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 10864−10868.
(43) Jiang, Y. W.; Stillman, D. J. *Genes Dev.* **1996**, *10*, 604−619.
(44) Keys, D. A.; Vu, L., Steffan, J. S.; Dodd, J. A.; Yamamoto, R. T.; Nogi, Y.; Nomura, M. *Genes Dev.* **1994**, *8*, 2349−62.

loaded on a gel. The utility of the technique was demonstrated by the rapid identification of 11 proteins in CTD-affinity preparations. The technology has great potential for postgenome biological science where it promises to facilitate the dissection and anatomy of macromolecular assemblages[14,17] (e.g., we are currently defining the total complement of proteins in the yeast nuclear pore complex), the definition of disease state markers, and the investigation of protein targets in biological processes such as cell cycle and signal transduction.

ACKNOWLEDGMENT

This work was supported in part by grants from the NIH (RR00862) and the NSF (DBI-9630936).

Received for review May 12, 1997. Accepted July 16, 1997.[⊗]

AC970488V

[⊗] Abstract published in *Advance ACS Abstracts*, September 1, 1997.