A Method for the Chemical Generation of N-Terminal Peptide Sequence Tags for Rapid Protein Identification

ARTICLE *in* ANALYTICAL CHEMISTRY · APRIL 2000

Impact Factor: 5.64 · DOI: 10.1021/ac9911847 · Source: PubMed

8 AUTHORS, INCLUDING:

Sjouke Hoving
Novartis
29 PUBLICATIONS **1,054** CITATIONS
SEE PROFILE

Peter James
University of Queensland
120 PUBLICATIONS **4,247** CITATIONS
SEE PROFILE

Anal. Chem. **2000,** *72,* 1006−1014

A Method for the Chemical Generation of N-Terminal Peptide Sequence Tags for Rapid Protein Identification

Sjouke Hoving, Martin MÜnchbach, Holger Schmid, Luca Signor, Anton Lehmann, Werner Staudenmann, Manfredo Quadroni, and Peter James*

Protein Chemistry Laboratory, Universitätsstrasse 16, Swiss Federal Institute of Technology, 8092 Zürich, Switzerland

We describe a method for generating multiple small sequences from the N terminal of peptides in unseparated protein digests by stepwise thioacetylation and acid cleavage. The mass differences between a series of N-terminally degraded peptides give short sequences of defined length. Such short "sequence tags" together with the mass of the parent peptide can be used to identify the protein in a database. The sequence ladders are generated without the use of chain terminators or sample aliquoting and the degradation reagents are water soluble so that the chemistry can be carried out on peptides immobilized on C-18 reversed-phase supports without any peptide loss due to washing with organic solvents as occurs in Edman type sequencing. The entire procedure can be automated, and we describe a prototype device for the parallel analysis of multiple samples. We demonstrate the effectiveness of this chemical tagging method in a comparison with Edman sequencing, peptide mass fingerprinting, and MS/MS analysis of crude protein fractions obtained from an HPLC separation of the *Escherichia coli* ribosome complex which consists of 57 proteins. We show that chemical tagging is a viable first-pass high-throughput identification method to be used prior to an in depth MS/MS analysis.

The development of high-throughput DNA sequencing and computer algorithms for the rapid assembly of the random sequence fragments into large contiguous sequences has resulted in an exponential growth in database size. Currently, 17 bacterial and 2 eukaryotic genomes have been completed, with another 62 bacterial and 9 eukaroytic genomes due before the turn of the century. These, together with extensive Expressed Sequence Tag (ESTs, partial mRNA sequences[1]) libraries, allow the entire potential protein complement of organisms to be defined (the Proteome). Genome-wide studies of gene expression are now possible at the mRNA level,[2] since the development of DNA microchips and arrays,[3] differential display PCR,[4] and serial analysis of gene expression.[5]

The level of protein expression and activity is modulated to a great extent by a wide variety of posttranslational modifications, such as phosphorylation, and by the relative rate of synthesis and degradation and is, to a large extent, independent of the level of mRNA expression.[6] In order for proteome-level analysis to be viable, rapid and sensitive protein identification methodologies must be developed. The earliest methods, such as sequence determination by Edman degradation or compositional analysis such as amino acid analysis, are still useful but are increasingly being replaced by mass spectrometry based methods. The first advance was the development of database-searching algorithms to identify proteins on the basis of the masses of the peptides produced by sequence-specific chemical or proteolytic digestion.[7,8,9,10,11] This was subsequently extended to the use of orthogonal digest data,[12] partially interpreted MS/MS,[13] and raw MS/MS data for DNA database searching.[14] Peptide mass fingerprinting is very rapid (1 sample per 30 s in fully automated mode) but the data is only useful for searching protein databases. MS/MS searching is much more effective for DNA database searching however data accumulation is relatively slow at ∼10 min per sample. We therefore decided to develop an intermediate solution, in which data accumulation is very rapid but in which some sequence data is also generated to allow DNA database searching and which gives a much higher confidence level in the search result.

* Corresponding author: (phone) 0041 1632 2919; (fax) 0041 1632 1591; (e-mail) peter.james@bc.biol.ethz.ch.

(1) Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M. Polymeropoulos, M. H.; Xiao, H.; Merril, C. R.; Wu, A.; Olde, B.; Moreno, R. F.; Kerlavage, A. R.; McCombie, W. R.; Venter, J. C. *Science (Washington, D.C.)* **1991**, *252*, 1651−6.

(2) Lashkari, D. A.; DeRisi, J. L.; McCusker, J. H.; Namath, A. F.; Gentile, C.; Hwang, S. Y.; Brown, P. O.; Davis, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 13057−62.

(3) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. *Science (Washington, D.C.)* **1995**, *270*, 467−70.

(4) Liang, P.; Pardee, A. B. *Science (Washington, D.C.)* **1992**, *257*, 967−71.

(5) Velculescu, V. E.; Zhang, L.; Vogelstein, B.; Kinzler, K. W. *Science (Washington, D.C.)* **1995**, *270*, 484−7.

(6) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. *Mol. Cell. Biol.* **1999**, *19*, 1720−30.

(7) 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*, 5011−5.

(8) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64.

(9) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass. Spectrom.* **1993**, *22*, 338−45.

(10) Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−32.

(11) Yates, J. R. D.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397−408.

(12) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Protein Sci.* **1994**, *3*, 1347−50.

(13) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−9.

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

10.1021/ac9911847 CCC: $19.00
Published on Web 02/04/2000

Ladder sequencing, in which a sequence is read by the mass differences between sequential degradation products, was developed first as an enzymatic technique[15] and then subsequently as a modified Edman type chemical degradation.[16,17,18] The use of exopeptidases to generate ladders is limited by the extreme variability of the activity of the protease toward the substrate and is only useful for individual isolated peptides.[19,20] Chemical sequencing using phenylisothiocyanate with a small percentage of phenylisocyanate as a chain-terminating reagent has been used by Chait et al.[16] to sequence long isolated peptides in a modified commercial Edman sequencer. The main disadvantages of the method are the loss of peptide during the washing steps, which limits the sensitivity, as well as the terminating reagent, which removes the alpha N terminus as a charge carrier, thereby diminishing the relative effectiveness of ionization. The volatile trifluoroethylisothiocyanate analogue developed by Pappin et al. removes the need for washes with organic solvent but requires that the parent peptide be added back in aliquots in order to generate the ladder which again causes losses through sample handling. We therefore chose to develop a method that could be carried out in aqueous solution with an unseparated mixture of peptides immobilized on a reversed-phase material and that leaves the N terminal unmodified so that it may act as a proton acceptor.

EXPERIMENTAL SECTION

Materials. Acetonitrile, mercaptoacetic acid, *N*-ethylmorpholine, pyridine, thioacetylthioethane, and trifluoroacetic acid (for protein sequence analysis) were purchased from Fluka AG (Buchs, Switzerland). α-Cyano-4-hydroxycinnamic acid, *N*-methylpiperidine, and *S*-(thiobenzoyl)thioglycolic acid were purchased from Aldrich GmbH (Buchs, Switzerland). Trizma base was purchased from Sigma Chemical Co. (Buchs, Switzerland). Ammonium hydrogen carbonate, β-mercaptoethanol, sodium sulfide, and sodium sulfate were purchased from Merck AG (Darmstadt, Germany). Acetic acid, carbon tetrachloride, diethyl ether, hydrochloric acid (min. 37%), and HPLC grade acetonitrile were purchased from Riedel-de Haën AG (Seelze, Germany). Sequencing-grade modified trypsin was purchased from Promega (Zürich, Switzerland). DNase was purchased from Boehringer (Mannheim, Germany). PTFE membranes with C-18 embedded material (Empore, 3M) were from Varian (Zug, Switzerland). Peptide C21W (WFRGLNRIQTQIRVVNAFRSS) was synthesized manually using Fmoc chemistry.

Synthesis of Thioacetylthioglycollic Acid (TATG). The synthesis was based on the description of Mross and Doolittle.[21] Mercaptoacetic acid (46 g, 0.50 mole) and acetonitrile (22.5 g, 0.55 mole) were mixed in a reaction flask and cooled on ice. The mixture was overlaid with 2−3 cm petroleum ether and HCl gas was bubbled for 1 h through the lower phase until a white solid appeared (carboxymethyl-thioimidate). CAUTION, HCl gas is extremely agressive and all handling should be carried out in an appropriate fume cupboard. The solvent was then evaporated under vacuum with a rotary evaporator and 250 mL of dry pyridine were added. H_2S was bubbled through for 4 h until the reaction was completed. 100 mL ice water was added and stirred to dissolve the ammonium chloride. The liquid was then poured over a mixture of 300 mL concentrated hydrochloric acid, 100 mL water and 300 g ice. The product was extracted with 300 mL diethyl ether. The aqueous phase was extracted twice more with 150 mL diethyl ether. The combined diethyl ether layers were then washed with 3 M hydrochloric acid, following which the diethyl ether was dried with Na_2SO_4. The diethyl ether was removed by rotary evaporation, yielding about 20 g of oil. The TATG was crystallized from warm carbon tetrachloride, the yield was 5 g of yellow solid. ^{1}H NMR ($CDCl_3$, 300 MHz): $\delta = 2.89$ (s, CH_3), 4.13 (s, CH_2) ppm. ^{13}C NMR ($CDCl_3$, 75 MHz, proton decoupled): $\delta = 38.57$ (CH_3), 38.86 (CH_2), 173.9 (COOH), 230.8 (C=S) ppm. ESI-MS: 45 (6.88), 59 (80.9), 76 (4.86), 91 (7.51), 117 (4.83), 150 (100) M^+.

Ribosome Isolation. *Escherichia coli* MC4100 F$^-$ *araD*139 Δ(*argF-lac*) U169 *rspL*150 *relA*1 *deoC*l *ptsF*25 *rpsR flbB*5301 was obtained from the Cold Spring Harbor Laboratory collection.[22] Bacteria were cultivated in a sulfur-free, synthetic glucose-salts medium, with the addition of 500 μM inorganic sulfate as described before.[23] The cultures were grown aerobically on a rotary shaker (180 rpm) at 37 °C, and growth was monitored spectrophotometrically at 600 nm. Cells were harvested in the mid-exponential phase ($A_{600} = 1$) by centrifugation at 9000 rpm in a Sorval GSA rotor for 10 min and washed twice with 50 mM Tris-HCl, pH 7.0. The cells (10 g of wet cells) were resuspended in 70 mL of extraction buffer (20 mM Tris-HCl, pH 7.4, 40 mM NH_4Cl, 10 mM $MgCl_2$, and 7 mM β-mercaptoethanol). One tablet of a cocktail of protease inhibitors (Boehringer, Mannheim, Germany) was added, and the cells were ruptured by two passes through a chilled French pressure cell at 20 000 psi. Ten microliters of DNase (10 μg/μL) was added to the ruptured cells and incubated at 37 °C for 40 min. Cell debris was removed by centrifugation at 16 000 rpm for 30 min at 4 °C in a Sorvall SS-34 rotor. Ribosomes were isolated as previously described.[24] The supernatant was centrifuged at 45 000 rpm for 2 h at 4 °C in a Beckman Ti45 rotor. The pellet contains the crude ribosomes and was resuspended in 20 mL of high-salt buffer (20 mM Tris-HCl, pH 7.4, 400 mM NH_4-Cl, 10 mM $MgCl_2$, and 7 mM β-mercaptoethanol). Then, 5 mL of the crude ribosomal solution was layered onto a 7-mL 17.5% sucrose cushion (in high-salt buffer) and centrifuged at 45 000 rpm in a Beckman Ti45 rotor for 3 h at 4 °C. The pellet (containing 70S ribosomes) was resuspended in 10 mL of low-Mg^{2+} buffer (20 mM Tris-HCl, pH 7.4, 40 mM NH_4Cl, 0.3 mM $MgCl_2$, and 7 mM β-mercaptoethanol) layered onto four continuous 10−30% sucrose gradients, and centrifuged in a Beckman Ti28 swing-out rotor at 18 000 rpm for 14 h at 4 °C. The fractions containing

(15) Aimoto, S.; Takao, T.; Shimonishi, Y.; Hara, S.; Takeda, T.; Takeda, Y.; Miwatani, T. *Eur. J. Biochem.* **1982**, *129*, 257−63.

(16) Chait, B. T.; Wang, R.; Beavis, R. C.; Kent, S. B. *Science (Washington, D.C.)* **1993**, *262*, 89−92.

(17) Bartlet-Jones, M.; Jeffery, W. A.; Hansen, H. F.; Pappin, D. J. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 737−42.

(18) Gu, Q. M.; Prestwich, G. D. *J. Pept. Res.* **1997**, *49*, 484−91.

(19) Korostensky, C.; Staudenmann, W.; Dainese, P.; Hoving, S.; Gonnet, G.; James, P. *Electrophoresis* **1998**, *19*, 1933−40.

(20) Patterson, D. H.; Tarr, G. E.; Regnier, F. E.; Martin, S. A. *Anal. Chem.* **1995**, *67*, 3971−8.

(21) Mross, G. A.; Doolittle, R. F. In *Advanced methods in protein sequence determination;* Needleman, S. B., Ed.; Springer-Verlag: Berlin, 1977.

(22) Silhavy, T. J.; Berman, M. L.; Enquist, L. W. *Experiments with Gene Fusions*; Cold Spring Harbor Laboratory: New York, 1984.

(23) Kertesz, M. A.; Leisinger, T.; Cook, A. M. *J. Bacteriol.* **1993**, *175*, 1187−90.

(24) Traub, P.; Mizushima, S.; Lowry, C. V.; Nomura, M. *Methods Enzymol.* **1979**, *49*, 391−407.

protein were pooled (60 mL total) and dialyzed overnight against 5 L of low-Mg^{2+} buffer. In the presence of a low Mg^{2+} concentration, the two subunits dissociate. The dialyzed solution was centrifuged at 45 000 rpm for 4 h at 4 °C and the ribosomes were resuspended in 7.5 mL of low-Mg^{2+} buffer and stored at −20 °C. Protein was determined, according to a modified method of Lowry,[25] as 7.5 mg/mL.

Separation of Ribosomal Subunits. Before separating the ribosomal proteins by reversed-phase HPLC, the rRNA was extracted. $MgCl_2$ and acetic acid were added to the ribosomes to a final concentration of 67 mM and 67%, respectively, and then the entire solution was incubated on ice for 1 h. The mixture was centrifuged at 10 000 rpm for 10 min at 4 °C in a Sorvall SS-34 rotor. The procedure was repeated once more with the pellet.[26] The supernatants were combined and concentrated in the speed-vac. The proteins were redissolved in 3% acetic acid, centrifuged at 10 000 rpm for 20 min at room temperature in an Eppendorf centrifuge to remove any undissolved particles, and injected onto a preparative reversed-phase HPLC system (L-6220 Intelligent Pump, L-4250 UV−vis Detector, Merck-Hitachi AG, Darmstadt, Germany). A gradient of 10−25% B in 30 min, 25−35% B in 40 min, 35−36% B in 30 min, 36−40% B in 40 min, 40−55% B in 60 min, and 55−90% B in 40 min was run at 2 mL/min (A = 0.1% TFA, B = 80% acetonitrile/0.08% TFA), using a C_{18} preparative column (250 × 21 mm, Nucleosil 100−12 μm, Macherey-Nagel AG, Oensingen, Switzerland). The absorbance was measured at 220 nm. Fractions of 2 mL were collected.[27,28]

Protein Digestion. Fractions containing protein from the preparative reversed-phase HPLC were dried in a speed-vac and redissolved in 100 μL of water. For tryptic digestion, 20 μL of protein was taken (approximately 120 μg of protein), and porcine trypsin (sequencing grade) was added at 2% (w/w) with respect to the ribosomal protein in a buffer of 100 mM NH_4HCO_3, pH 8.0. Digestion was performed for 48 h at 37 °C in an Eppendorf shaker and stopped by addition of 5 μL of formic acid. The solutions were dried in the speed-vac to remove the volatile buffer, resuspended in 100 μL of water, and dried again, and finally, the peptides were redissolved in 50 μL of water and stored at −20 °C.

Peptide Degradation. The peptide degradation was carried out with a prototype device which was designed to deliver reagents in the gas phase (*N*-methylpiperidine (NMP) or *N*-ethylmorpholine (NEM) and trifluoroacetic acid (TFA)) at low pressure to 10 samples, in parallel, in a thermostatically controlled block. Standard peptides or tryptic digests (usually 1−20 pmol in 0.1% TFA) were loaded slowly (2−5 μL/min) onto PTFE membranes with embedded C_{18} reversed-phase material in a multiple sample loader and washed with 50 μL of 0.1% TFA. Nitrogen gas at 0.5 bar pressure was used to deliver either gas or liquid reagents. The ladder generating steps were performed at 40 °C. The degradation steps were carried out as follows: (a) purge with N_2 for 10 min at 1.0 bar, (b) NMP base (gas) delivery for 3 min at 0.5 bar, (c) load 2 μL of 50 mM TATG in 1% NEM (cycle 1 and 2) or 2 μL of 100 mM TATG in 2% NEM (cycle 3 and 4) and incubate 5 min, (d) purge with N_2 for 5 min at 1 bar, (e) TFA gas delivery for 3 min at 0.5 bar, (f) purge with N_2 for 10 min at 0.5 bar, and (g) wash with 50 μL of water. Then, the cycle was usually repeated (steps a−g) three times. Finally, the peptides were slowly eluted with 50 μL of 60% acetonitrile/0.1% TFA.

MALDI Mass Spectrometry. The masses of the intact proteins from the HPLC separation as well as those of the tryptic peptides and the degraded peptides were determined by MALDI-MS. Approximately 1 pmol in 0.5 μL was added to the same amount of a saturated matrix solution (15 mg of α-cyano-4-hydroxy-cinnamic acid in 50% acetonitrile and 1.25% TFA in water) and allowed to dry at ambient temperature. Spectra were recorded using a Voyager Elite MALDI-TOF mass spectrometer (Perseptive Biosystems, Framingham, MA). Samples were analyzed in delayed extraction reflector mode using an accelerating voltage of 20 kV, a pulse delay time of 150 ns, a grid voltage of 60%, and a guide wire voltage of 0.05%. Spectra were accumulated for 32 or 64 laser shots.

MS/MS Sequencing of Ribosomal Proteins. Two microliters of each tryptic digest (~5 pmol) from the ribosomal proteins was desalted on the C_{18} membranes and redissolved in 50% methanol and 1% acetic acid before being introduced into the mass spectrometer at a flow rate of 0.2 μL/min with a syringe pump. MS/MS sequencing was performed on a Finnigan MAT LCQ ion trap mass spectrometer (San Jose, CA). The peaks of interest were selected with a mass window of ±3 amu, and fragmentation of each peptide was established using a relative collision energy of 35−60 for MH^+ ions and 20−30 for MH^{2+} ions. On average, 20 MS/MS spectra were measured within 15 min for each protein digest.

N-terminal Sequencing of Ribosomal Proteins. Intact ribosomal proteins (0.5−1 μL) from the preparative HPLC separation were spotted onto methanol-wetted PVDF membrane. N-terminal sequence analysis was directly performed on a Hewlett-Packard G1000A protein analyzer, equipped with four cartridges. Released PTH amino acids were analyzed on a Hewlett-Packard HPLC series 1100 (Palo Alto, CA). Six sequence cycles were performed according to the standard protocols provided by the manufacturer.

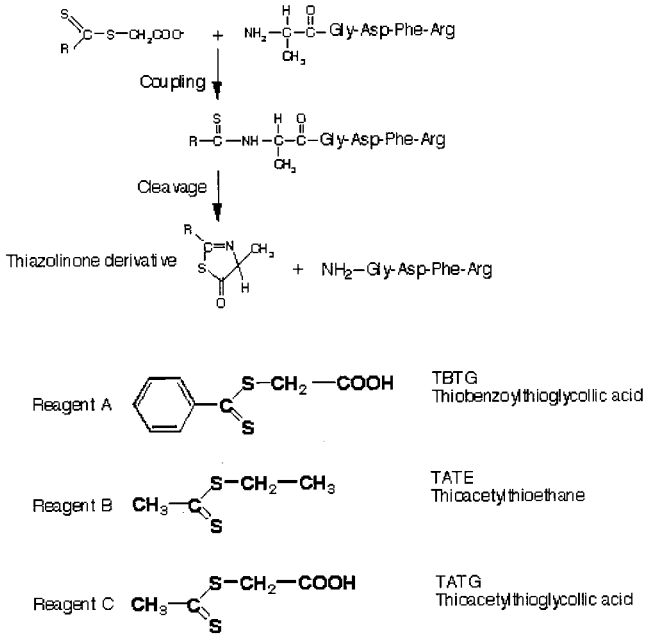
Database Searching. N-terminal sequences were searched against the Swissprot and trEMBL databases using the FASTA program. Peptide mass fingerprinting searches were carried out using either the MassSearch or PeptideSearch algorithms.[8,9] Database searching using the chemical tag data was carried out with MassDynSearch. The program searches the SwissProt and trEMBL by peptide mass after N− or C-terminal degradation (enzymatic or chemical). The algorithm was developed in collaboration with the group of Professor Gaston Gonnet (Computation Biology Research Group, ETH Zürich) and a detailed description of the algorithm and its application to exopeptidase-based ladder searching has been published.[19] Information for obtaining the code for MassDynSearch is available at the web site http://cbrg.inf.ethz.ch. MS/MS searches were carried out using the Sequest program and the nonredundant database.[14]

RESULTS AND DISCUSSION

Ladder Sequencing Chemistry. Peptide sequencing can be carried out by generating a series of degradation fragments so

(25) Markwell, M. A.; Haas, S. M.; Bieber, L. L.; Tolbert, N. E. *Anal. Biochem.* **1978**, *87*, 206−10.
(26) Hardy, S. J.; Kurland, C. G.; Voynow, P.; Mora, G. *Biochemistry* **1969**, *8*, 2897−905.
(27) Kamp, R. M.; Wittmann-Liebold, B. *FEBS Lett.* **1984**, *167*, 59−63.
(28) Kamp, R. M.; Bosserhoff, A.; Kamp, D.; Wittmann-Liebold, B. *J. Chromatogr.* **1984**, *317*, 181−92.

Scheme 1. General Outline of the Degradation Chemistry and Reagents Used



that the sequence can be read out by sequentially measuring the mass differences between the products starting from the parent mass. The ladder can be generated by either chemical or enzymatic means and was first shown in the early 1980s[15] using carboxypeptidase digestion of a small enterotoxin. The development of MALDI mass spectrometers made this method very attractive, and the method was rediscovered. Single purified peptides were digested with either amino- or carboxypeptidases to generate N- or C-terminal sequence ladders which were then analyzed by MALDI-MS[20] or by a modified Edman degradation to generate ladders.[16,17,18] Unfortunately, the exopeptidase approach is limited to single peptides and the outcome depends strongly on the sequence specificity of the exopeptidase being used. The length of the ladder and especially its starting point is very variable and hence is not practical for use with peptide mixtures. Chemical methods are to be preferred since the number of amino acids removed can be easily controlled, greatly simplifying spectral interpretation. However the Edman type isothiocyanate degradation is not easy to carry out with low picomole amounts of material since the peptides must either be immobilized on a support, which must be washed after the coupling step to remove excess reagent,[16] or the peptides must be aliquoted and added back after each cycle when using the volatile reagents.[17]

We therefore explored the use of alkyl alkoxydithioformates for peptide degradation as had been first demonstrated by Kenner and Khorana in 1952.[29] We tested three reagents for their suitability as ladder-generating reagents (see Scheme 1); reagents A and B are commercially available, and C was synthesized as described by Mross and Doolittle.[21] The reagents were coupled in the presence of either *N*-ethylmorpholine (NEM) or *N*-methylpiperidine (NMP). Other bases, such as *N*,*N*-dimethylamino(pyridine), trimethylamine, and pyridine, were found to be effective but were not used due to their unpleasant smell. The

(29) Kenner, G. W.; Khorana, H. G. *J. Chem. Soc.* **1952**, 2076−81.

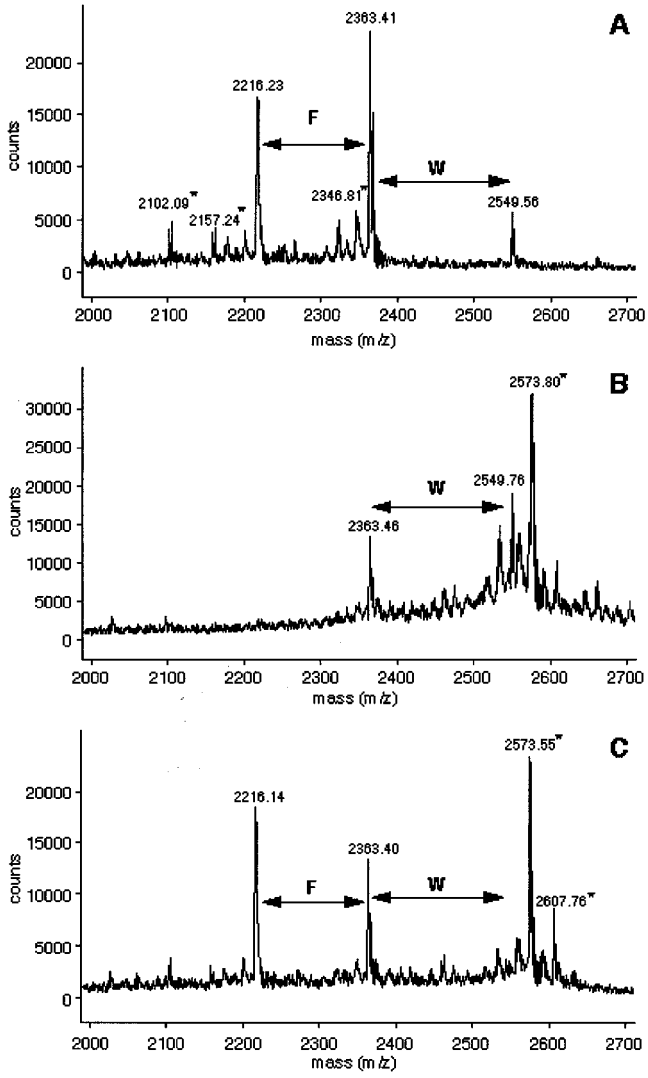


Figure 1. Comparison of the degradation reagents. (A) MALDI-TOF spectrum of the peptide C21W after two rounds of degradation with thiobenzoylthioglycollic acid. * indicates the peak is a byproduct. (B) As above but using the reagent thioacetylthioethane. (C) As above but using the reagent thioacetylthioglycollic acid.

coupling of all reagents was found to occur rapidly and completely at temperatures between 40 and 80 °C and reaction times of 5−30 min. Higher temperatures led to a number of side reactions. One reaction which occurs if air is not rigorously excluded is the S → O substitution in the thioacetyl-peptide intermediate after the coupling reaction which leads to blocking of the degradation since the acetyl-peptide cannot be cleaved. The thioacetyl-peptide derivative is fairly stable; however, if water is present during the cleavage step with the TFA, the back reaction, cleavage to thioacetate, competes with the release of the thioazolinone amino acid. This is the basis of the generation of the ladder sequence. Figure 1 shows the mass spectra obtained from the peptide C21W immobilized on reversed-phase material after two cycles of degradation with each of the three reagents.

Reagent A, thiobenzoylthioglycollic acid (TBTG), reacted well and gave the two expected degradation products (−W and −WF); however, there were also a number of side products whose formation could not be avoided and that we could not remove using different conditions (Figure 1A). The main problem with

reagent A is that it causes considerable loss of more hydrophilic peptides from the reversed-phase support, probably by a bulk elution effect. Reagent B, thioacetylthioethane (TATE), allows only the removal of the N-terminal amino acid, after which no further cleavage is seen to occur (Figure 1B). This was found to be due to the buildup of hydrophobic salt (base−TFA) which buffers subsequent base additions, lowering the coupling efficiency and peptide loss. The third reagent C, thioacetylthioglycollic acid (TATG), is the most hydrophilic of the series and gives an excellent ladder with very few side reactions (Figure 1C). A considerable amount of Na^+ adduct was found as well as a significant amount of uncleaved product, especially at large sterically hindered residues. This was subsequently corrected by adding a longer cleavage step after the final coupling reaction and including a final water wash of the peptides on the membrane. A series of experiments using standard peptides were carried out to determine the sensitivity level of the sequencing method. The chemistry described in the Experimental Section worked well without modification from the mid femtomole (100 femtomole of C21W gave a clear signal after three cycles) to the low nanomole range. The limiting factors are sample handling and the final elution volume from the support.

Construction of a Sequencer Prototype. A prototype sequencer was constructed to allow multiple samples to be analyzed and to ensure reproducibility of reagent delivery (Figure 2). Five reagent/solvent bottles are connected to two 6-way rotary valves the sixth position is used to flush the lines between deliveries. Nitrogen is delivered from a pressure-reducing valve via the first rotary valve to a selected bottle to pressurize it before the second rotary valve is opened to allow the chemical to be delivered to the sample block. It is important that all the components are either PEEK or made of glass since TFA is very aggressive and will corrode most materials. The entire bottle assembly is contained with an explosion/TFA proof chamber for safety reasons. One-way valves are installed on all the lines to prevent TFA seeping back to the rotary and pressure-reducing valves. The reagent flow is directed from the bottles to either waste or to a stainless steel block. The block consists of three metal plates, each with a heating element inside, which sandwich a Teflon membrane containing C-18 reversed-phase particles (Empore). In the version shown in Figure 2, 10 positions have been drilled in the bottom two plates which serve as sample-loading and reaction chambers. The protein digests are acidified and then loaded into the funnel-shaped vessels. The top plate is then attached, and the digests are slowly pushed by gas pressure through the Empore membrane. The membrane is then washed with 0.1% TFA to remove any buffers prior to starting the degradation steps. The block temperature is maintained at a constant value that is preset 15 min before starting the degradation.

Protein Identification by Chemical Tag Searching. We chose the ribosome complex from *Escherichia coli* as a test system to investigate the effectiveness of the degradation procedure on unseparated protein digests to generate sequence tags for protein identification by database searching. The *E. coli* ribosome consists of 57 proteins arranged in two subunits, the 30 S made up of proteins S1−21 and the 50S proteins L1−36. The complex was isolated, and the rRNA was removed before a rough separation of the subunits was carried out by reversed-phase chromatography (Figure 3). Forty-seven fractions were obtained. Each fraction was analyzed on an automated Edman sequencer to obtain the N-terminal sequence(s) of the protein(s) present, and a MALDI spectrum of the intact proteins was measured. An aliquot of each HPLC fraction was then digested with trypsin and split into two for either MS/MS analysis on an electrospray ion trap mass spectrometer or for chemical tagging in the prototype instrument described above. Figure 4 shows the results obtained with fraction 36. Since several proteins were observed to be present in the fraction by MALDI-MS (Figure 4A), the mixture of sequences obtained from N-terminal sequencing was impossible to interpret. There were not many peptides present, and the digest was obtained from a mixture of proteins, so no clear identification could be obtained by protein mass fingerprinting (Figure 4B). The partial degradation approach gave three clear tag sequences (Figure 4C) which matched the *E. coli* 50S subunit ribosomal proteins L1 and L11 as well as the 30S subunit protein S7. This agreed well with the masses obtained by MALDI analysis of the intact proteins as well as the results obtained by MS/MS analysis of the digest mixture (Figure 4D) and database searching using Sequest.

The results are summarized in Table 1. The forty-seven HPLC fractions each contained between 1 and 7 proteins, giving a total of 106 proteins (from 62 different genes), since many proteins occurred in more than one fraction. All of the proteins S1−21 in the 30S subunit were found to be present as well as all of the 50S proteins, L1−36, except for L8, L9, and L21. However, L8 does not really exist—it is a complex of L10, L7/12, all of which were identified. Five ribosome-associated proteins were identified and confirmed by manual interpretation of the MS/MS data (Fraction 15, P52098 yaeO, a hypothetical open reading frame product; Fraction 38, P21507, RNA helicase and P07012, peptide chain release factor 2; Fraction 39, P07019, GTP binding export factor; and Fraction 43, P52084, another hypothetical open reading frame product). The 'Gold Standard' for the comparison of methods were the twenty or so MS/MS spectra obtained per fraction which were used to search the databases automatically using the Sequest program. Proteins identified by less than 5 independent MS/MS spectra searches were confirmed by manual sequence interpretation of the data.

The relative effectiveness versus the time consumption of each method is highlighted below. The most effective method is clearly the use of MS/MS data in conjunction with a database searching program such as Sequest. One hundred and six proteins could be identified in a total sample preparation/data accumulation/database searching time of 30 h. Edman degradation was the least effective with only 30 proteins identified in 450 h (cost-effectiveness not even considered). The most interesting comparison is between peptide mass fingerprinting and chemical tagging. The partial degradation method using unseparated protein digests could identify 50 proteins in a first-pass search (using only 1−3 very clear sequence tags and the peptide masses) and a further 30 in a second-pass search using the masses left over from the first search (after removing those which matched the highest scoring protein) with a total data accumulation and search time of 5 h. The sample loading and degradation is very rapid and can be carried out on a MALDI target-sized block with 50 positions in less than 2 h. The final wash on the support desalts the samples, and they are simultaneously eluted, using matrix in 60% acetoni-

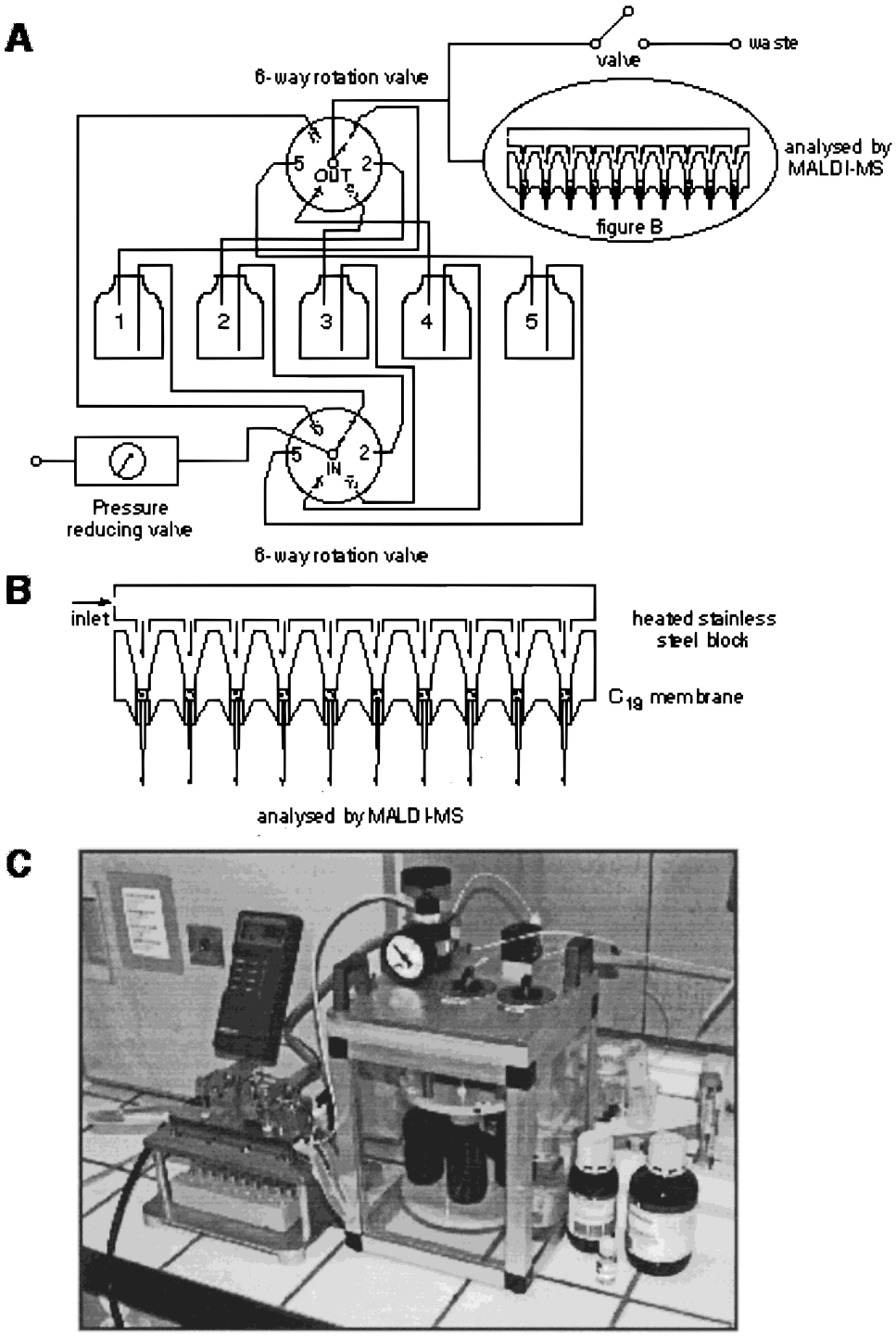


Figure 2. Prototype reagent delivery device and reaction chamber. (A) The general layout of the tubing and valves is indicated. (B) The sample holder and heating elements are shown (indicated in A). (C) Photograph of the actual prototype.

trile, 0.1% TFA, directly onto the MALDI target in a minute. Peptide mass fingerprinting (mass accuracy first pass, 0.1 mass units; second pass, 0.04 units) identified only 30 proteins in 3 h.

Chemical tagging is thus a viable alternative to simple peptide mass fingerprinting. It is formally analogous to the peptide sequence tag approach advocated by Matthias Mann in which the peptide sequence tag is derived from manual inspection of an MS/MS spectrum.[13] The main advantages of the "Chem-Tag" approach is that it can be used with nonsequence specific digestions such as elastase or partial acid hydrolysis since the sequence tag is always N-C-terminal and the length is predefined. The database search data can include the sequence specificity of the digestor and hence the database search has all the advantages of a peptide fingerprint search with the added advantage of partial sequence information. The confidence level of such a search containing partial sequence information is on average about 1000-fold higher than that based on mass alone (given the same mass accuracy). A related approach using conventional Edman chemistry-derived N-terminal sequence information from intact proteins and mass-spectral data has been suggested,[30] but this is much slower, relies

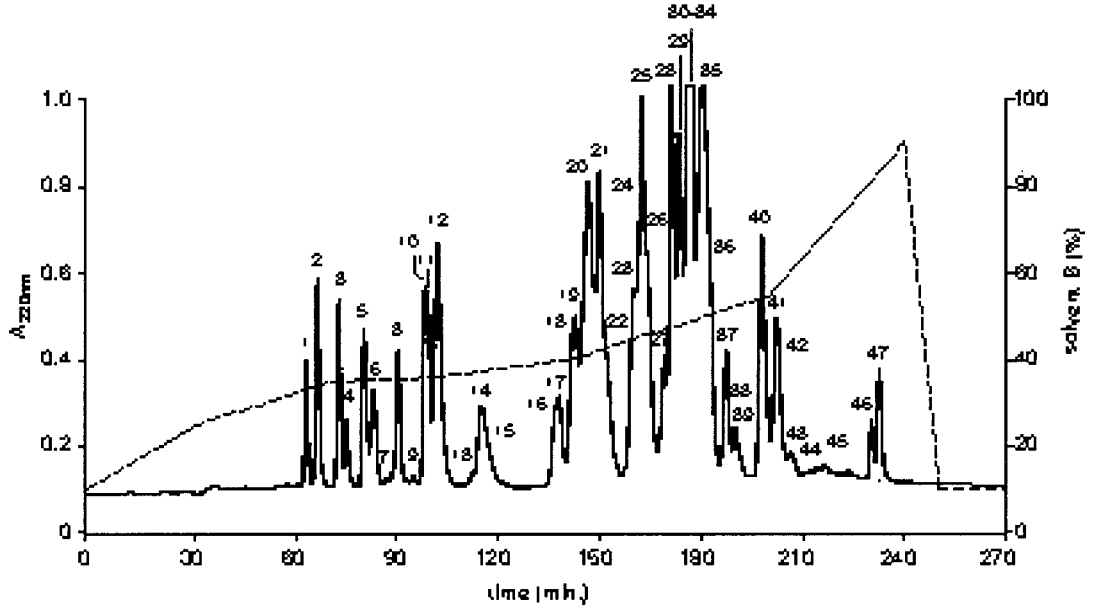


Figure 3. HPLC chromatogram of *E. coli* ribosomal subunits. The isolated ribosomes were injected onto a preparative reversed-phase HPLC and 2-mL fractions were collected each minute.

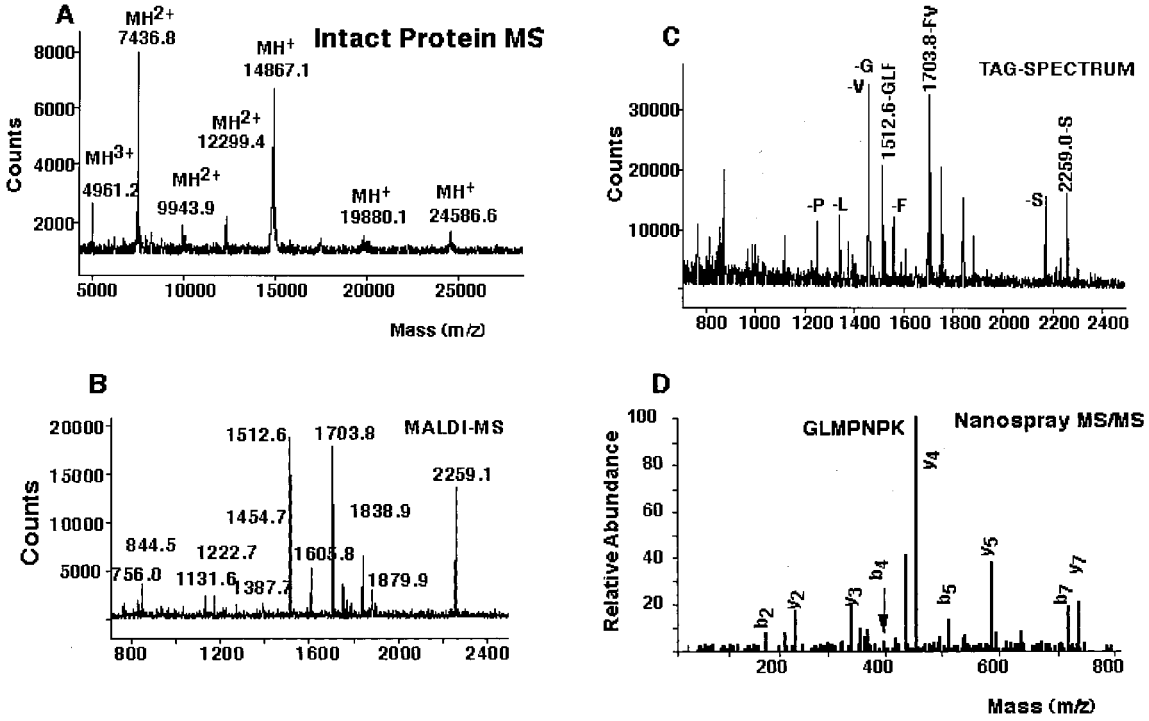


Figure 4. Mass Spectrometric analysis of fraction 36. (A) MALDI-TOF spectrum of the fraction before digestion. (B) MALDI-TOF spectrum of the tryptic digest. (C) MALDI-TOF spectrum of the tryptic digest after ladder sequencing. (D) ESI-MS/MS spectrum of a peptide at *m/z* 380.5 (2+ ion).

on only a single sequence rather than multiple short sequences, and is much less stringent.

CONCLUSIONS

We have demonstrated that the use of multiple short chemically generated sequence tags is much more effective in terms of protein identification as well as confidence level than peptide mass fingerprinting alone. Data extraction using chemical tag is much simpler than that from MS/MS data and is much less time consuming in terms of data accumulation and database searching (due to the predefined directionality and starting point). The main advantage of Chem-Tag searching is that it can be directly applied to a DNA sequence (especially EST libraries) with much more confidence than simple mass fingerprinting and much less effort than MS/MS peptide-tag searching. The ribosomal analysis emphasizes this advantage in the case of small proteins that

(30) Wilkins, M. R.; Ou, K.; Appel, R. D.; Sanchez, J. C.; Yan, J. X.; Golaz, O.; Farnsworth, V.; Cartier, P.; Hochstrasser, D. F.; Williams, K. L.; Gooley, A. A. *Biochem. Biophys. Res. Commun.* **1996**, *221*, 609−13.

Table 1. Identification of Proteins in the HPLC Fractions

fraction no.[a]	protein ID[b]	mass calc. (av.)[c]/obs. (av.)[d]	N-terminal protein sequence DB[e]/Exp[f]	N-terminal peptide tag (MH^+ mono)[g]	protein identified by Mass-DynSearch[h]	protein identified by Mass-Search[i]	protein identified by FASTA[j]
1	50S subunit L34	5380/5384	MKRTFQ/MKRTFQ (ragged)	919.6 T	50S L34		50S L34
2	50S subunit L32	6446/6320 (−M)	MAVQQN/MAVQQN	1232.6 HH	50S L32		50S L32
3	50S subunit L33	6371/6259 (−M + CH_3)	MAKGIR/AKGIRE	1568.5 LVS	50S L33		50S L33
4	50S subunit L36	4364/4370	MKVRAS/MKVRAS	none			50S L36
5	50S subunit L27	9124/8995 (−M)	MAHKKA/AHKKAG	1404.6 FGGE 1300.5 V	50S L27		50S L27
6	50S subunit L14	13541/13681 (conflict in sequence)	unreadable	1873.5 IG (chymo)	30S S12		
	30S subunit S12	13606/13681 (−M)					
7	50S subunit L31	7871/7872	MKKDIHPKYE/MKKDI	none			50S L31
	50S subunit L27	9124/8998 (−M)	MAHKKA/no sequence				
	30S subunit S17	9704/not detected	MTDKIR/no sequence				
8	50S subunit L31	7871/7874	MKKDIHPKYE/MKKDI	none			50S L31
	50S subunit L7/12	12295/not detected	MSITKD/no sequence				
9	30S subunit S21	8500/8371 (−M)	MPVIKV/ragged N terminus	1189.5 E	30S S21		
10	50S subunit L24	11316/11184 (−M)	MAAKIR/no sequence	1804.1 EAA	50 S L24	50 S L24	
	50S subunit L34	5380/not detected	MKRTFQPSVLK/ FQPSVL				50S L34
11	50S subunit L24	11316/11184 (−M)	MAAKIR/no sequence	none	50 S L24	50 S L24	
	30S subunit S21	8500/8371 (−M)	MPVIKV/no sequence				
	50S subunit L35	7289/7161 (−M)	MPKIKTVRGAA/PKIKT				50S L35
12	30S subunit S14	11580/11449 (−M)	MAKQSM/AKQSMK	1085.6 G	50S L28		30S S14
	50S subunit L28	9006/8876 (−M)	MSRVCQ/SRVCQV				50S L28
13	30S subunit S11	13844/13880 (−M + CH_3)	MAKAPI/????	none			30S S11
14	30S subunit S11	13844/13805 (−M)	unreadable	1280.6 KG			
	30S subunit S19	10430/10446			30S S19	30S S19	
15	30S subunit S11	13844/13729 (−M)	no sequence	2478.0 pEVS	30S S11		
	ACP52098, yaeO	9698/not detected	MSMNDT/MNDTYQ				ACP52098
16	30S subunit S20	9684/9556 (−M)	MANIKS/ANIKSA	1319.8 AF	30S S20		30S S20
	50S subunit L34	5380/not detected	MKRTFQPSVLKRNR/VLKRNR				50S L34
17	30S subunit S20	9684/9555 (−M)	unreadable	1240.7 YLSL			
	30S subunit S18	8986/8899 (−M + Acetyl)			30S S18		
	50S subunit L24	11316/not detected					
18	50S subunit L25	10693/10694	unreadable	1080.7 M	50S L25	50S L25	
	30S subunit S15	10268/10139 (−M)		1027.7 LQ			
19	50S subunit L2	29860/29712 (−M)	unreadable	2632.2 DAND (chymo)	30S S15		
	30S subunit S15	10268/10137 (−M)		922.3 Y			
20	50S subunit L19	13133/12999 (−M)	unreadable	1315.9 (QA	50S L19	50S L2	
	50S subunit L2	29860/not detected			50S L2		
	50S subunit L14	13541/13537					
21	50S subunit L2	29860/29678 (−M)	unreadable	1202.8 H ($+Na^+$)	50S L2	50S L2	
	30S subunit S5	17603/not detected					
	50S subunit L30	6541/6412 (−M)			50S L30		
22	50S subunit L30	6541/6413 (−M)	MAKTIK/AKTIKI	913.6 AT	50S L30	50S L30	50S L30
	50S subunit L2	29860/29707 (−M)	MAVVKC/no sequence		50S L2		
23	50S subunit L3	22243/22245	MIGLVG/MIGLVG	2187.7 IFT	50S L3	50S L3	50S L3
	30S subunit S17	9704/9574 (−M)	MTDKIR/TDKIRT				30S S17
24	50S subunit L13	16018/16017	MKTFTA/MKTFTA	2187.5 IFT	50S L13	50S L13	50S L13
	50S subunit L3	22243/22252	MIGLVG/no sequence		50S L3		
25	50S subunit L13	16018/16019	MKTFTA/MKTFTA	2233.4 AE	50S L13	50S L13	50S L13
	50S subunit L17	14364/14365	MRHRKS/no sequence	2104.2 VY (C-term)	50S L17	50S L17	
26	50S subunit L17	14364/14363	MRHRKS/MRHRKS	1634.4 AGDN	50S L17	50S L17	50S L17
27	50S subunit L18	12769/12769	MDKKSA/MDKKSA	951.5 SG	50S L18	50S L18	50S L18
28	30S subunit S4	23469/23240 (conflict in sequence)	unreadable	1456.6 GNT	30S S4	30S S4	
	50S subunit L22	12226/12226					
	30S subunit S3	25983/25828 (−M)					
	50S subunit L18	12769/12767					
29	30S subunit S13	13099/12965 (−M)	unreadable	1213.8 TS (C-term)			
	30S subunit S10	11735/11733					
	50S subunit L22	12226/12224			50S L22	50S L22	
30	50S subunit L22	12226/12223	unreadable	1521.7 L (C-term)			
	30S subunit S13	13099/12967 (−M)					
	50S subunit L23	11199/11197					
	50S subunit L16	15281/15312 (+CH_3)					
	30S subunit S8	14126/13990 (−M)					
	30S subunit S1	61158/61508					
	30S subunit S21	8500/not detected			30S S21		
31	50S subunit L22	12226/not detected	unreadable	1521.7 LA (C-term)	50S L22	50S L22	
	30S subunit S16	9190/9187					
	50S subunit L23	11199/11193					
	30S subunit S21	8500/not detected			30S S21		
	30S subunit S8	14126/13988 (−M)					
32	50S subunit L22	12226/not detected	unreadable	1685.0 KVE	50S L22	50S L22	
	50S subunit L29	7273/7276					
	50S subunit L16	15281/15319 (+CH_3)					
	50S subunit L20	13496/not detected					
	30S subunit S19	10430/not detected			30S S19		
33	50S subunit L29	7273/7276	MKAKEL/MKAKEL	1643.2 SV	50S L29	50S L19	50S L19
	50S subunit L15	14980/14974	MRLNTL/no sequence		50S L15		
34	50S subunit L15	14980/14974	unreadable	1989.9 LNT	50S L15	50S L15	
	30S subunit S6	15703/15763					
	30S subunit S9	14856/14974 (conflict in sequence)			30S S9	30S S9	
	50S subunit L6	18903/not detected					
	30S subunit S5	17603/17507 (−M + Acetyl)					

Table 1 (Continued)

fraction no.[a]	protein ID[b]	mass calc. (av.)[c]/obs. (av.)[d]	N-terminal protein sequence DB[e]/Exp[f]	N-terminal peptide tag (MH^+ mono)[g]	protein identified by Mass-DynSearch[h]	protein identified by Mass-Search[i]	protein identified by FASTA[j]
35	50S subunit L1	24729/24575 (−M)	unreadable	1704.0 FV	50S L1	50S L1	
	50S subunit L11	14875/14855 (−M + CH_3)					
	30S subunit S5	17603/17486 (−M + Acetyl)					
36	50S subunit L1	24729/24586 (−M)	unreadable	1704.0 FV	50S L1		
	50S subunit L11	14875/14867 (−M + CH_3)		1512.6 GLP	50S L11		
	30S subunit S7	20019/19880 (−M)		2259.0 S	30S S7		
37	50S subunit L5	20301/20162 (−M)	MAKLHD/AKLHDY	1267.6 ALLA	50S L5	50S L5	50S L5
38	30S subunit S2	26743/not detected	MATVSM/ATVSM	none	30S S2	30S S2	30S S2
	P21507, RNA helicase	49914/not detected					
	50S subunit L15	14980/not detected					
	P07012, peptide chain release factor 2	41250/not detected					
39	30S subunit S2	26743/not detected	MATVSM/ATVSM	none		30S S2	30S S2
	P07019, GTP binding export factor	49787/not detected					
40	50S subunit L20	13496/13359 (−M)	MARVKR/ARVKRG	1104.5 IL	50S L20		50S L20
	30S subunit S2	26743/26712	MATVSM/no sequence				
41	50S subunit L10	17711/17569 (−M)	unreadable	1616.9 AA (chymo)	50S L10	50S L10	
	50S subunit L4	22086/22079		1446.7 LAT	50S L4	50S L4	
42	50S subunit L10	17711/17571 (−M)	unreadable	2561.6 DAF	50S L10	50S L10	
	50S subunit L4	22086/22061		1688.8 AAA	50S L4	50S L4	
43	50S subunit L10	17711/17574 (−M)	unreadable	none		50S L10	
	EC2215, AC unassigned	11305/11171					
44	50S subunit L10	17711/17574 (−M)	unreadable	none			
45	50S subunit L7/L12	12295/12297 (−M)	MSITKD/SITKD	none			50S L7/L12
46	50S subunit L7/L12	12295/12158 (−M)	MSITKD/SITKD	2015.2 FG 964.4 T	50S L7/L12	50S L7/L12	50S L7/L12
47	50S subunit L7/L12	12295/12199 (−M)	MSITKD/SITKD	2014.9 FGV 1244.5 A	50S L7/L12		50S L7/L12

[a] The fraction number corresponding to the peaks in the HPLC chromatogram shown in Figure 3. [b] Protein ID gives the protein name (from *Escherichia coli*) or if nonribosomal, the SwissProt accession number (AC). All proteins were identified by MS/MS using the Sequest program and confirmed by manual interpretation of the spectra. [c] The average intact protein mass was calculated from the database entry. [d] The experimentally determined average intact protein mass, given as average MH^+. The annotation (−M) indicates that the N-terminal initiator, methionine, is cleaved off. Acetyl and CH_3 indicate that a protein is posttranslationally modified by acetylation or methylation and conflict means that there are sequence differences between various entries for the same protein. [e] Representation of the N termini of the protein according to the database. [f] The experimentally determined N-termini of the protein (by Edman degradation). When no sequence was found or the sequence was unreadable, this is stated. [g] The experimentally obtained monoisotopic masses of the peptides with the partial N-terminal sequence tag obtained after digestion with trypsin and TATG degradation. Peptides generated by chymotryptic activity are indicated by (chymo), those from the C-terminal by (C-term), and pE indicates that the N-terminal amino acid was pyroglutamate. [h] Protein identification using the N-terminal chemical tag by searching SwissProt and trEMBL with the program MassDynSearch (Korostensky et al., 1998). Note that this search includes the masses of nontagged peptides as well. [i] Protein identification from the data of the tryptic peptide masses by searching SwissProt and trEMBL using the program MassSearch (James et al., 1993). [j] Protein identification from the data obtained from Edman N-terminal degradation and searching SwissProt and trEMBL using the program FASTA/TFASTA (Pearson and Lipman, 1988).

generate only a few peptide masses; this is of especial concern when one considers that the average length of a translated EST sequence is less than 300 amino acids. Although the procedure is effective at cleaving all amino acids, the spectra of digests often show only three or four tags. This is due to suppression effects which have been demonstrated to occur with MALDI-MS.[31] The removal of amino acids changes the relative ionization efficiency, giving rise to altered relative intensities. This is clearly seen when one compares the spectra obtained, from identical samples, by MALDI-MS and by ESI-MS: the main peaks are usually observed in both techniques but the tag sequences found are often completely different. We are currently investigating the use of degradation reagents that carry a fixed positive charge which we hope should reduce the suppression effects.

(31) Kratzer, R.; Eckerskorn, C.; Karas, M.; Lottspeich, F. *Electrophoresis* **1998**, *19*, 1910−9.

ACKNOWLEDGMENT

Funding was provided by the Swiss National Foundation for Scientific Research.

Received for review October 13, 1999. Accepted December 13, 1999.

AC9911847