High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry

AXEL DUCRET,¹ INGE VAN OOSTVEEN, JIMMY K. ENG, JOHN R. YATES III,
AND RUEDI AEBERSOLD

Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, Washington 98195

(RECEIVED August 12, 1997; ACCEPTED November 21, 1997)

Abstract

We describe an integrated workstation for the automated, high-throughput, and conclusive identification of proteins by reverse-phase chromatography electrospray ionization tandem mass spectrometry. The instrumentation consists of a refrigerated autosampler, a submicrobore reverse-phase liquid chromatograph, and an electrospray triple quadrupole mass spectrometer. For protein identification, enzymatic digests of either homogeneous polypeptides or simple protein mixtures were generated and loaded into the autosampler. Samples were sequentially injected every 32 min. Ions of eluting peptides were automatically selected by the mass spectrometer and subjected to collision-induced dissociation. Following each run, the resulting tandem mass spectra were automatically analyzed by SEQUEST, a program that correlates uninterpreted peptide fragmentation patterns with amino acid sequences contained in databases. Protein identification was established by SEQUEST_SUMMARY a program that combines the SEQUEST scores of peptides originating from the same protein and ranks the cumulative results in a short summary.

The workstation's performance was demonstrated by the unattended identification of 90 proteins from the yeast *Saccharomyces cerevisiae*, which were separated by high-resolution two-dimensional PAGE. The system was found to be very robust and identification was reliably and conclusively established for proteins if quantities exceeding 1–5 pmol were applied to the gel. The level of automation, the throughput, and the reliability of the results suggest that this system will be useful for the many projects that require the characterization of large numbers of proteins.

Keywords: electrospray ionization mass spectrometry; protein identification; proteome; *Saccharomyces cerevisiae*; sequence database; two-dimensional gel electrophoresis

The identification of proteins by their amino acid sequence has been a common and important process in diverse areas of biological research. Traditionally, proteins of a particular function were purified to homogeneity and then sequenced. The recent dramatic growth of large-scale genomic and expressed sequence tag sequence databases and their ease of accessibility through the Internet have created new challenges and opportunities for protein

identification technology. The challenges are posed by the need to assign biological function to the numerous sequences representing potentially expressed proteins. The opportunities are generated by the possibility for rapid and conclusive protein identification by correlating the information contained in their amino acid sequence with an object in a sequence database.

Any one of an array of mass spectrometric techniques has essentially replaced the less sensitive and slower chemical protein sequencing for the identification of proteins by correlation with sequence databases. The commonly used strategies can be classified into three groups. The first, peptide mass fingerprinting (reviewed in Cottrell, 1994; Patterson & Aebersold, 1995), relies on the mass measurement at high precision of peptides generated by sequence-specific enzymatic or chemical cleavage of a protein. Because a particular peptide mass can only be represented by a relatively limited number of amino acid compositions, the precise mass measurements of as few as three or four peptides derived from the same protein can be sufficiently selective to distinguish the target protein from all other sequences in a database. However, this method implicitly assumes that all the detected peptides orig-

Reprint requests to: Ruedi Aebersold, Department of Molecular Biotechnology, HSB-K343B, University of Washington, Seattle, Washington 98195; e-mail: ruedi@u.washington.edu.

¹Present address: Merck-Frosst Canada Inc., Biochemistry & Molecular Biology, P.O. Box 1005, Pointe-Claire-Dorval, Québec H9R 4P8, Canada.

Abbreviations: 2DE, two-dimensional gel electrophoresis; CID, collision-induced dissociation; ΔC_n , delta normalized cross-correlation factor; EST, expressed sequence tag; ICL, Instrument Control Language; LC-ESI-MS/MS, reverse-phase chromatography electrospray tandem mass spectrometry; MS, mass spectrometer/try; MS/MS, tandem mass spectrometer/try; PVP-40, polyvinylpyrrolidone-40; S/N, signal-to-noise; Xcorr, cross-correlation factor; CA, bovine carbonic anhydrase II; RP-HPLC, reverse-phase HPLC; DTT, dithiothreitol.

inate from a single and unmodified polypeptide, excluding de facto inhomogeneous or substantially modified protein samples. Also, peptide mass fingerprinting is not compatible with the identification of proteins by correlation with EST sequence databases. The second strategy, developed by Mann and Wilm (1994) (see also Shevchenko et al., 1996), is based on the partial manual interpretation of tandem mass spectra generated by collision-induced dissociation of selected peptide ions. The mass of the intact peptide ion combined with a short stretch of sequence within the peptide (oriented sequence tag) allows a rapid and generally unambiguous identification of a protein by the analysis of a single peptide. However, the sequence tag method is limited by the time and the considerable expertise required for the manual determination of the sequence tag and by the low potential for automation of the process. The third strategy, developed by some of us (J.R. Yates & J.K. Eng), is based on the automated correlation of uninterpreted MS/MS spectra with sequence databases using the SEQUEST software (Eng et al., 1994; Yates et al., 1995). The program correlates an experimental CID spectrum with the theoretical, computer-generated MS/MS spectra of all peptides isobaric to the parent ion that has been analyzed in the MS/MS experiment. The quality of the matches is then expressed numerically in correlation scores. This approach, as the sequence tag method, is tolerant to the analy-

sis of protein mixtures, but does not presume any specificity of the proteolytic enzyme used. In addition, because there is no requirement for manual interpretation of MS/MS spectra, the SEQUEST method can be run batchwise in a totally automated manner.

It has been suggested repeatedly that the ability to rapidly and reliably characterize proteins separated by high-resolution two-dimensional gel electrophoresis could provide a useful tool for the functional investigation of biological systems (reviewed in Anderson & Anderson, 1996). The proposed strategies for extracting system-relevant information from the 2DE protein patterns vary considerably. They include the systematic characterization and cataloging in a protein database of all the expressed proteins in a tissue or species (recently termed proteome analysis) (Wilkins et al., 1996) and the subtractive analysis of protein patterns generated from samples representing different experimental or pathological conditions followed by the identification of those proteins that differ between the samples. The different strategies share the need for the characterization of numerous proteins deemed relevant for the system under investigation.

Here we describe a protein workstation for the large-scale, automated characterization of proteins by reverse-phase chromatography electrospray tandem mass spectrometry. The system shown schematically in Figure 1 integrates in a fully automated fashion

Fig. 1. Schematic of the automated protein workstation. After separation by 2DE and enzymatic cleavage, samples were loaded to an autosampler and sequentially injected into an RP-HPLC system. Peptides were sequentially detected by their UV absorbance at 214 nm and by on-line ESI-MS/MS. MS/MS spectra were analyzed by SEQUEST and SEQUEST_SUMMARY. Results are summarized and displayed in a one- to two-page report.

the operations required for establishing the protein identity, including sample injection, peptide separation, peptide analysis by ESI-MS/MS, sequence database searching, and concise reporting of the summarized results. We report the completely automatic analysis, within a period of 48 h, of 90 yeast proteins isolated from 2DE gels, and thus demonstrate that the system offers an effective, robust, and reliable solution to the problem of large-scale identification of proteins.

Results

Performance of protein identification system

The performance of the system was assessed by the use of calibrated peptide samples. Different amounts of bovine carbonic anhydrase II ranging from 20 to 300 ng, as determined by quantitative amino acid composition analysis, were subjected to SDS-PAGE, blotted onto nitrocellulose, and digested with trypsin as described in Material and methods. The recovered peptides were separated by RP-HPLC and sequentially detected by UV absorbance and ESI-MS/MS. We recorded the number of peptides recovered, the signal-to-noise ratios in the UV absorbance and MS detection, the quality of the MS/MS spectra generated by CID, and the reliability of the results obtained from the automated protein identification software.

UV absorbance detection

In general agreement with our previous experience (Ducret et al., 1996; van Oostveen et al., 1997), the digestion protocol used was not found to interfere significantly with the UV absorbance detection of peptides. A typical result is shown in Figure 2, upper panel. After elution of the large injection peak, the peptides derived from the digestion of 200 ng CA were detected on a smooth baseline with a sensitivity limit in the high femtomole range and a typical peak width of 30–40 s. Although not essential for protein identification, the UV chromatogram provided valuable information on the quality of the samples analyzed, the level of contaminants, and proved useful for trouble-shooting the HPLC system. The use of a microbore on-column detection system and low i.d. fused silica capillaries and peek tubing minimized the loss of chromatographic resolution between the flow cell and the ion source of the mass spectrometer.

MS and MS/MS detection

The column effluent monitored by UV absorbance detection (Fig. 2, upper panel) was also analyzed by on-line ESI-MS and ESI-MS/MS (Fig. 2, lower panel). ESI-MS recorded several UV-transparent contaminants. Their presence correlated with the use of polyvinylpyrrolidone-40 to block the adsorption of the proteolytic enzyme to the nitrocellulose membrane. The extent of the contamination could be reduced, but not totally eliminated, by extensively washing the nitrocellulose membrane after PVP-40 treatment. The ESI-MS limit of detection for peptides was in the range of 500 fmol. If the ion current for a peptide exceeded a predetermined threshold, the ion of interest was selected for CID using the Instrument Control Language procedures described in Material and methods. The threshold value was empirically set at a level that minimized MS/MS analysis of background ions, yet allowed the selection of low-abundance peptide species for CID. Typically, peptide ions that were present at an S/N ratio of 10 or higher in the

Fig. 2. Chromatogram of an on-membrane tryptic digest of 200 ng (6.5 pmol) CA. Top panel: UV absorbance at 214 nm. Bottom panel: Base peak mass chromatogram. Identified, CA-derived peptides are indicated by arrows and contain the following sequences: T4, (K)QSPVNIDTK; T5, (K)AVVQDPALKPLALVYGEATSR; T5c, (K)AVVQDPALKPLALVY; T7, (K)MVNNGHSFNVYDDSDQK; T9, (R)DGPLTGTYR; T14, (K)YGD FGTAQQPDGLAVGVFLK; T15, (K)VGDANPALQK; T16, (K)VLD ALDSIK; T20, (K)EPISVSSQMLK.

MS scan were found to yield MS/MS spectra of a quality suitable for SEQUEST analysis (see below). Because in a triple quadrupole MS the quality of a MS/MS spectrum is strongly dependent on the energy applied to the peptide ion in the collision cell, we investigated the fragmentation pattern of selected peptides in the charge states $[M + H]^+$, $[M + 2H]^{2+}$, and $[M + 3H]^{3+}$ (Ducret et al., 1996) as a function of the fragmentation energy. The conditions described in Materials and methods yielded high-quality CID spectra for most of the peptides analyzed and are in good agreement with previous reports (Ducret et al., 1996; Haller et al., 1996; Stahl et al., 1996). Peptide MS/MS spectra were automatically processed by SEQUEST at the end of each HPLC analysis.

Sensitivity

The sensitivity of the system was assessed using the cross-correlation factor and the delta normalized cross-correlation factor, factors indicating the quality of the SEQUEST database search results, as well as the number of peptides unambiguously identified by SEQUEST. Results from experiments in which different amounts of CA were analyzed are summarized in Figure 3. The minimal amount of protein required for unambiguous peptide identification was approximately 2 pmol (60 ng) of protein applied to the gel. Smaller amounts did not generate MS/MS spectra of sufficient quality for successful database searches. The number of peptides identified by SEQUEST also significantly increased with increasing

Fig. 3. Dependence of Xcorr and ΔC_n values on sample amount. The Xcorr and ΔC_n values, respectively, of every CID spectrum that was unambiguously assigned to the CA protein used in this experiment are plotted versus the amount of CA protein applied to the gel. Data from one of two experiments are shown. MS/MS spectra were considered of sufficient quality for protein identification when Xcorr values exceeding 2.0 and ΔC_n values exceeding 0.1 were obtained. The number of CID spectra correlated by SEQUEST with CA is indicated by the numbers in the plot. Full line: Xcorr (♦, lower limit; ■, mean; ◆, higher limit); dashed line: ΔC_n (▼, lower limit; ◆, mean; ▲, higher limit).

protein amounts (3 peptides at 60 ng; 11 peptides at 300 ng). Because the calculated Xcorr value depends on the absolute number of ions detected during the CID experiment, low mass peptides, i.e., species for which only few fragment ions can be generated, required better-quality MS/MS spectra (and therefore higher amounts loaded onto the system) to be identified unambiguously. At the 300 ng level (10 pmol of CA applied to the gel), the quality of the MS/MS spectra was sufficiently high to generate conclusive Xcorr values, even for small peptides. At this sensitivity, the Xcorr values were between 2.11 and 6.07 for peptides between 9 and 22 amino acids in length. A second parameter, ΔC_n , was used to eliminate false positives. For peptides of similar length, which generally generate comparable numbers of fragment ions in a CID experiment, ΔC_n is indicative of the quality of the match between the experimental and a theoretical MS/MS spectrum. Eng et al. (1994) determined experimentally that the SEQUEST search was likely to identify a true positive if the difference in ΔC_n values between the best and second-best score exceeded 0.1. Search results for peptides exceeding 15 amino acids in length for which large numbers of fragment ions and high Xcorr values were generated yielded ΔC_n values of up to 0.4, indicating an excellent discrimination between true and false positives. For smaller peptides, especially when Xcorr scores were close to the chosen cut-off value (see below), ΔC_n values were often a decisive additional parameter for confident peptide identification. In our experience, peptides smaller than six amino acids were generally not useful for the identification of proteins by this method.

Automated protein identification

The SEQUEST database search results obtained from individual peptides derived from the same protein were automatically summarized by the program SEQUEST_SUMMARY, which was appended to SEQUEST. The five highest ranked matches of each database search were scored as described in Materials and methods

and the scores of peptides that originated from a common protein were summed up. Because each MS/MS spectrum represented an independent analysis, proteins that were matched by two or more first-ranked peptides (receiving scores of 20 or higher) were considered conclusively identified. Lower scores typically required manual postacquisition analysis. Using the parameters discussed above, we observed that SEQUEST correctly identified peptides for which Xcorr values above 2.0 were obtained. SEQUEST_SUMMARY was therefore set to ignore Xcorrs below that cut-off value. This eliminated ambiguous database matches from the final analysis. The signal generated by contaminating PVP-40, for which typically Xcorrs between 1.0 and 1.5 were obtained, was also eliminated from analysis by the choice of this threshold. Peptides from the same protein were usually recovered in approximately equimolar amounts after on-blot digestion. Consequently, target proteins were usually identified by several peptides. This assured a high degree of confidence in the determined protein identity and minimized time-consuming postacquisition analysis.

Automated identification of yeast proteins separated by 2DE

To assess the reliability of the automated protein characterization workstation, batches of 30–40 yeast proteins separated by 2DE were subjected to unattended analysis. Pooled protein spots from three blots were cleaved with trypsin and the acidified digests were maintained at 4 °C in the autosampler until injection. The peptides were separated by RP-HPLC and analyzed by on-line ESI-MS/MS. At the end of each run, the generated CID spectra were automatically analyzed by SEQUEST and SEQUEST_SUMMARY.

The proteins selected for analysis are indicated in Figure 4. They were divided into three classes according to the amount of protein present per spot. Spots 1–18 represented abundant proteins with quantities exceeding 1 μ g per gel. They were easily detected by Coomassie Blue staining. Spots 19–45 were clearly detected with the Silver and Amido Black stains after electroblotting onto nitrocellulose, but were only weakly visible if a similarly loaded gel was stained with Coomassie Blue. The protein amounts contained in the intermediate intensity spots 19–45 were therefore estimated in the range of 100 ng–1 μ g per gel. Faint spots, numbered 46–70, were detected with the Silver stain or faintly with the Amido Black stain and contained estimated amounts between 20–100 ng per gel.

Results are summarized in Table 1. With the exception of spot 4, which failed to be injected by the autosampler, proteins contained in spots 1–18 were reliably identified. With this amount of sample, a minimum of six peptides were characterized per protein. The analysis of spots 8, 16, and 18 identified more than one protein comigrating in the respective 2D gel spots. Most of the intermediate intensity spots 19–45 were also identified unambiguously. For these proteins, a minimum of three peptides were detected. Spots 23, 30, 33, and 35 failed to yield any detectable peptides. Most of the low-abundance spots 46–70 were also identified. Typically, a minimum of two peptides per protein were detected. As an example, the analysis of spot 59 is shown in Figure 5. An estimated 250 ng (5 pmol) of protein was present per spot. The LC-MS/MS analysis of the tryptic digest is shown in the upper panel. The MS/MS spectra that were correlated by SEQUEST with the target protein sequence are labeled by arrows. The system successfully detected peptide ions and acquired their MS/MS spectra among the contaminants introduced by the use of PVP-40 to block the nitrocellulose membrane. A partial SEQUEST_SUMMARY output is shown in the lower panel of Figure 5. The results indicate

Fig. 4. 2DE separation of total yeast cell lysate. Yeast cell lysate (200 μ g) was separated by 2DE. First dimension: linear immobilized pH gradient 4–7 (Fig. 4A) or a linear immobilized pH gradient 6–10 (Fig. 4B). Second dimension: 12% SDS-PAGE. Proteins were visualized in the gel by silver staining. For protein identification, unstained gels were electroblotted to nitrocellulose immediately after completion of electrophoresis and proteins were visualized by Amido Black or colloidal silver staining. Numbers and letters indicate the spots that were subjected to automated LC-ESI-MS/MS analysis.

Table 1. Identification of yeast proteins by automated LC-ESI-MS/MS^a

Spot #	Measured pI	Measured MW (kDa)	Database identification	Calculated pI	Calculated MW (kDa)
1	6.0	48	ENO1_YEAST	6.2	46.7
2	6.8	46	PGK_YEAST	7.1	44.6
3	9.0	30	PMGY_YEAST	8.9	27.5
4	6.7	39	No injection	—	—
5	5.7	50	ENO2_YEAST	5.7	46.8
6	5.6	48	ENO2_YEAST	5.7	46.8
7	5.7	28	TPIS_YEAST	5.8	26.6
8	5.0	16	YOL109w	5.4	12.6
			YGD7_YEAST	5.8	25.0
			YOR007c	4.7	37.2
9	5.5	28	TPIS_YEAST	5.8	26.6
10	4.9	13	HS12_YEAST	5.2	11.7
11	5.3	40	IPYR_YEAST	5.4	32.8
12	4.9	26	TSA_YEAST	5.0	21.5
13	5.5	19	SODS_YEAST	5.6	15.7
14	5.5	48	ENO2_YEAST	5.7	46.8
15	5.6	36	GBLP_YEAST	5.8	34.8
16	5.1	30	HS26_YEAST	5.3	23.8
			VATE_YEAST	5.5	26.6
17	6.3	43	ADH1_YEAST	6.3	36.7
18	6.4	39	G3P2_YEAST	6.5	35.7
			G3P3_YEAST	6.5	35.6
19	5.4	48	ENO2_YEAST	5.7	46.8
20	4.7	36	BMH1_YEAST	4.8	30.0
21	4.8	30.0	PMP_YEAST	5.1	19.1
22	4.5	33	HS71_YEAST	5.0	69.5
23	4.4	44	No data	—	—
24	4.3	36	TPH1_YEAST	4.6	23.6
25	4.5	28	HS71_YEAST	5.0	69.5
26	4.1	41	PDI_YEAST	4.3	55.3
27	4.2	28	TCTP_YEAST	4.4	18.7
28	8.1	16	HS26_YEAST	5.3	23.7
29	6.2	17	CYPH_YEAST	6.9	17.3
30	6.5	46	No data	—	—
31	8.1	39	G3P1_YEAST	8.3	35.6
32	8.6	30	PMGY_YEAST	8.9	27.5
33	8.0	11	No data	—	—
34	8.0	18	G3P2_YEAST	6.5	35.7
35	6.0	27	No data	—	—
36	6.5	9	IPB2_YEAST	6.3	8.5
37	6.6	30	ENO2_YEAST	5.7	46.8
38	9.1	16	ENO2_YEAST	5.7	46.8
39	6.6	43	ENO1_YEAST	6.2	46.7
40	9.1	15	ENO2_YEAST	5.7	46.8
41	5.7	24	YEP7_YEAST	5.7	19.0
42	4.8	18	COF1_YEAST	5.1	15.9
43	5.8	25	SODM_YEAST	6.1	23.1
44	4.7	37	BMH2_YEAST	4.8	30.9
45	5.8	30	G3P3_YEAST	6.5	35.6
46	6.8	35	No data	—	—
47	6.9	33	No data	—	—
48	9.1	25	ATP7_YEAST ^b	8.9	19.7
49	8.1	30	No data	—	—
50	8.8	21	ENO2_YEAST ^b	5.7	46.7
51	9.2	22	No data	—	—
52	5.4	24	HS26_YEAST ^b	5.3	23.7
53	5.3	26	No data	—	—
54	4.9	70	No data	—	—
55	4.8	75	No data	—	—
56	5.4	19	SODC_YEAST	5.6	15.7

(continued)

Table 1. Continued

Spot #	Measured pI	Measured MW (kDa)	Database identification	Calculated pI	Calculated MW (kDa)
57	5.7	33	PRTB_YEAST ^b GPP2_YEAST ^b	5.5 5.8	38.1 27.8
58	5.3	41	ENO2_YEAST	5.7	46.7
59	4.9	41	ADK_YEAST	5.0	36.7
60	5.8	31	DKA1_YEAST	6.1	24.4
61	4.7	38	YOR230w	5.2	48.4
62	5.4	26	TPIS_YEAST	5.8	26.6
63	4.5	40	HS78_YEAST ^b	5.1	77.2
64	7.2	15	G3P3_YEAST	6.5	35.6
65	5.8	30	ENO2_YEAST ^c	5.7	46.8
66	7.8	27	PMGY_YEAST	8.9	27.5
67	7.1	15	ALF_YEAST	5.5	39.5
68	7.8	39	G3P1_YEAST ^b	8.3	35.6
69	6.8	27	No data	—	—
70	7.9	30	ENO1_YEAST ^{bc}	6.2	46.7

^aSelected protein spots from three blots were excised and pooled, digested with trypsin, and analyzed by LC-ESI-MS/MS. The MS/MS spectra were analyzed automatically as described. The sequence database used was a selection of all the yeast protein sequences in the OWL database (release 29.1). Results were summarized by SEQUEST_SUMMARY. Spots are numbered as indicated in Figure 4.

^bOnly one peptide identified.

^cOther protein(s) were matching the identified peptide(s). The most probable entry is shown.

Partial SEQUEST_SUMMARY output of Spot 59:

scan #	Mass	Xcorr	ΔCoff	Reference	Sequence
199-203	1087.4	3.0055	0.300	ADK_YEAST	(K) ENDAILVDK
261-265	1357.3	3.6829	0.254	K2C1_HUMAN	(K) LNDLEDALQAK
441-445	1699.7	3.0725	0.495	ADK_YEAST	(K) M#AIFDELLQM#PETK
577-581	1128.4	3.4239	0.365	ADK_YEAST	(K) LVAGGAQNTAR
654-658	1479.6	3.6053	0.408	ADK_YEAST	(K) GSTYFVRFLLSSK
737-741	1699.2	3.4459	0.510	ADK_YEAST	(K) M#AIFDELLQM#PETK
765-769	2013.1	5.8912	0.563	ADK_YEAST	(K) IVDYNGAGDAFAGGFMAGLTK
779-783	2063.3	4.7486	0.487	ADK_YEAST	(R) SLVTDLGAANFFTPDHLK
793-797	1302.9	2.5986	0.087	RAD4_SCHPO	(R) SILWELNNGK
826-830	1666.5	4.3031	0.347	ADK_YEAST	(K) MAIFDELLQMPETK
946-950	2659.9	2.1392	0.063	CW41_YEAST	(K) FMTNNLEANPGLLLEYAKKIYK

1. ADK_YEAST	80	(8,0,0,0,0)
2. RAD4_SCHPO	10	(1,0,0,0,0)
3. CW41_YEAST	10	(1,0,0,0,0)
4. K2C1_HUMAN	10	(1,0,0,0,0)

Fig. 5. Automated LC-ESI-MS/MS identification of protein in spot 59. Top panel: UV absorbance at 214 nm. Bottom panel: base peak mass chromatogram. The CID spectra correlated with the sequence of the putative adenosine kinase (ADK_YEAST) are indicated by arrows. The partial SEQUEST_SUMMARY output only includes those SEQUEST analyses for which a minimal Xcorr of 2.0 was obtained. M#, methionine sulfoxide; K2C1_HUMAN, human keratin.

that spot 59 was identified as ADK_YEAST (a putative adenosine kinase) by correlating eight peptides with the ADK_YEAST sequence. One peptide was identified as derived from human keratin, a common contaminant in high-sensitivity peptide mapping experiments. From each one of the spots 48, 50, 52, 57, 63, 68, and 70, only one peptide was identified. Although the identification of these proteins is lacking the statistical significance obtained by multiple, independent analyses, a high Xcorr, a high ΔC_n , and perfect alignment of the identified peptides between predicted tryptic cleavage sites in the target protein sequence represented strong indications for the identity of the proteins contained in those spots. Spots 46, 47, 49, 51, 53, 54, 55, and 69 failed to be identified by the system. The corresponding UV traces indicated that the peptide amounts retrieved from these proteins were probably below the sensitivity limit of the system.

Characterization of simple mixtures of yeast proteins

The results shown above indicate that the system is capable of identifying individual components in simple protein mixtures. With the intent of further enhancing sample throughput, we deliberately combined protein spots from different areas of the 2D gel prior to proteolysis and analysis. The selected spots contained between 100 ng and 1 μ g of protein and are designated with labels A–H in Figure 4. The identity of these proteins was then established by the SEQUEST search results and the correlation of the observed electrophoretic mobilities with the predicted positions in the gel. Results are summarized in Table 2 and representative data obtained from spots B1/B2 are presented in Figure 6. The UV and MS base peak traces are shown in the upper panel. The MS/MS spectra that were used by the SEQUEST software to identify the proteins are labeled with arrows. The partial SEQUEST_SUMMARY output from spots B1/B2 presented in the lower panel that identified spot

Table 2. Identification of yeast protein mixtures by automated LC-ESI-MS/MS^a

Spot #	Measured pI	Measured MW (kDa)	Database identification	Calculated pI	Calculated MW (kDa)
A1	4.8	19.7	HS78_YEAST	5.1	77.3
A2	5.7	40	PYRO_YEAST	5.8	34.8
B1	5.4	34	GPP1_YEAST	6.1	30.4
B2	5.6	28	KAD1_YEAST	6.0	24.0
C1	4.4	28	HS72_YEAST	5.0	69.3
C2	5.8	19	EGD1_YEAST	6.1	17
D1	5.1	11	FKBP_YEAST	5.7	12.1
D2	5.0	11	YHO7_YEAST	5.1	12.0
E1	4.3	31	YKL7_YEAST	4.5	24
E2	5.2	10	PMGY_YEAST	8.9	27.5
F1	5.7	45	ENO2_YEAST	5.7	46.7
F2	6.9	17	YIF1_YEAST	9.3	15.9
G1	4.2	13	COX6_YEAST	4.5	12.6
G2	9.1	10	CH10_YEAST	9.0	11.4
H1	4.8	21	PMP_YEAST	5.1	19.1
H2	5.8	48	ENO1_YEAST	6.2	46.7
			ENO2_YEAST	5.7	46.7

^aPairs of selected protein spots of significantly different mobilities were excised and pooled from three blots each. Proteins were digested with trypsin and analyzed by LC-MS/MS. MS/MS spectra were analyzed automatically as described. The sequence database used was a selection of all the yeast protein sequences in the OWL database (release 29.1). Results were summarized by SEQUEST-SUMMARY. Spots are numbered as indicated in Figure 4.

B1 as glycerol-3-phosphatase 1 (GPP1_YEAST) and spot B2 as ATP-AMP transphosphorylase (KAD1_YEAST), respectively. Due to the high degree sequence homology of glycerol-3-phosphatase isoforms (92.7% sequence identity between GPP1 and GPP2), some peptides matched both isoforms (Fig. 6). All the first-ranked peptides that matched isoform 2 were also present in isoform 1, whereas several peptides were specific for isoform 1. Moreover, the second-ranked peptide in analysis #34 in Figure 6 was identified as GPP2_YEAST peptide (K)RPKYFITANDVK. This peptide differed from the best match, peptide (K)RPEYFITANDVK from GPP1_YEAST by a single Glu to Lys amino acid exchange. Although the mass of the two peptide differed by one mass unit, the corresponding MS/MS spectra were sufficiently discriminating to allow an unambiguous peptide identification ($\Delta C_n = 0.23$). The identification of spot B2 as ATP-AMP transphosphorylase (KAD1_YEAST) was based on four first-place matches and some less significant assignments in second to fourth ranked positions (analysis 9, 13, and 40 of Fig. 6). Such lower-ranked matches, although inconclusive by themselves, add to the overall reliability of an identification. As for spot 59, we also detected two peptides from human keratin (analyses 47 and 65 in Fig. 6) during the analysis of spots B1 and B2, suggesting that relatively complex protein mixtures can be deconvoluted and characterized by the LC-ESI-MS/MS system described in this manuscript.

Discussion

We describe a system for the automated, high-throughput identification of proteins by correlation of peptide CID spectra with

sequence databases. The main characteristics of the system are the sample throughput, the level of automation, and the ability to identify individual components in protein mixtures. This was achieved by the integration of a LC-MS/MS system, software for database searching with peptide CID spectra, and instrument control protocols for the coordinated operation of the system.

Using calibrated protein samples and standard sample preparation protocols, the sensitivity boundaries of the system and parameters that affect the performance were investigated. We determined that the system routinely provided unambiguous protein identities if low picomole amounts of protein were applied to the gel or if high femtomole amounts of peptides were loaded onto the HPLC column. Several factors affected the data quality obtained from the LC-ESI-MS/MS system. First, the quality of data was dependent on the amount of protein. With the digestion procedure used, the peptide yields generally correlated well with the amounts of protein in the gel. For sample amounts below the 5 pmol level, peptide detection was complicated by the presence of peaks associated with the use of PVP-40. This could be minimized by concentrating the protein on a small area of nitrocellulose because the size of such signals are proportional to the nitrocellulose surface area of the analyzed spots. Second, because the sensitivity of ESI-MS is proportional to the analyte concentration, the signal intensity of a peptide ion was dependent on the chromatographic efficiency achieved by the RP-HPLC column. For the commercially packed high-quality columns, we found that peptide amounts exceeding 500 fmol generally generated CID spectra of sufficient quality for database searching. CID spectra of lower intensity peptide ions generally failed to be correctly correlated with a database by

Partial SEQUEST-SUMMARY report for spots B1/B2 :

#	Scan #	Mass	Xcorr	ΔCoff	Reference	Sequence
1	184-188	949.2	3.0038	0.036	KAD1_YEAST	(R) SDONADALK
9	269-273	1115.1	2.7207	-0.06	KAD1_YEAST	(R) CTQAPNLQER
10	316-320	1516.2	3.7922	0.270	GPP1_YEAST	(R) NGLGFFPINEQPSK
13	351-355	546.7	2.1318	-0.05	KAD1_YEAST	(R) SQIAK
17	398-402	1385.5	4.9350	0.312	GPP1_YEAST	(K) YGHSIEVPGAVK
23	488-492	2419.9	2.3638	0.088	KAD1_YEAST	(K) IFNPPKEDN#KDDVTGEALVQR
30	541-545	2404.5	3.9306	0.365	KAD1_YEAST	(K) IFNPPKEDN#KDDVTGEALVQR
34	555-559	1453.9	5.3107	0.231	GPP1_YEAST	(K) VVVFEDADAGIAAGK
36	586-590	1445.6	4.7024	0.314	GPP1_YEAST	(K) VVVFEDADAGIAAGK
40	614-618	1056.0	2.5030	-0.07	KAD1_YEAST	(R) M#VLGPPVAGK
47	664-668	1475.8	4.3654	0.207	K2C1_HUMAN	(K) WELLLQVDTSTR
51	678-682	1294.0	3.4244	0.074	GPP1_YEAST	(K) QGKPHPEPYLK
55	720-724	2440.4	5.6209	0.402	GPP1_YEAST	(K) PAPDFADEEYVKNLEGEIPEK
58	741-745	1695.8	4.2165	0.238	GPP1_YEAST	(K) IRRPEYTTANDVK
65	796-800	1303.6	4.1395	0.126	K2C2_HUMAN	(R) SLDLSIIAEVGA
67	856-860	2412.6	4.9364	0.347	KAD1_YEAST	(K) TGIWAGVDASQPPATVWADILNK
68	863-867	1553.6	3.5368	0.404	GPP1_YEAST	(K) IVGLIATTFDLDFLK
69	896-900	2598.8	5.0891	0.429	GPP1_YEAST	(R) VGEYNAETDEVELIFDDYLYAK

1.	GPP1_YEAST	90	(9,0,0,0,0)	{ 10 17 32 36 49 55 58 68 69 }
2.	KAD1_YEAST	58	(4,1,1,1,0)	{ 1 23 30 67 }
3.	GPP2_YEAST	28	(2,1,0,0,0)	{ 36 49 }

Fig. 6. Automated LC-ESI-MS/MS identification of proteins in the pooled spots B1/B2. Top panel: UV absorbance at 214 nm. Bottom panel: base peak mass chromatogram. The CID spectra correlated with the sequence of either glycerol-3-phosphatase 1 (GPP1_YEAST) or ATP-AMP transphosphorylase (KAD1_YEAST) are labeled with arrows and are numbered. Peak numbers indicate the corresponding lines in the SEQUEST_SUMMARY output. The partial output only includes those SEQUEST analyses for which a minimal Xcorr of 2.0 was obtained. The analyses for which the GPP1_YEAST or KAD1_YEAST reference is labeled in bold represent first position matches in the database search with the respective CID spectrum. Analyses 9, 13, and 40, which identified peptides from the ATP-AMP transphosphorylase ranked in third, second, and fourth position, respectively (as indicated by the negative ΔC_n coefficient), were included manually in the summary output. M#, methionine sulfoxide; K2C1_HUMAN and K2C2_HUMAN, human keratin.

SEQUEST. Third, the ICL procedure for the automated acquisition of MS/MS spectra contained a user-defined threshold. If the MS signal exceeded the threshold, the instrument switched to product ion scanning mode. The value was empirically set so that, from the number of peptide ions detected, a good-quality CID spectrum could be expected. Typically, a concentration of 100–300 fmol/ μ L (as determined by flow injection experiments) was required to generate CID spectra of sufficient quality for database searching. To eliminate low-quality CID spectra that were observed occasionally, even if the peptide ion intensity exceeded the threshold, SEQUEST_SUMMARY was set to consider only analyses for which Xcorr above 2.0 were obtained. The byproduct peaks that correlated with the use of PVP-40 generated Xcorr between 1.0 and 1.5 and were therefore eliminated from the final scoring of the experiment. Although the stringency of the parameters enhanced reliability and robustness of the scoring system, proteins below the picomole range were generally not identified. Furthermore, under these conditions, MS/MS spectra of peptides shorter than eight amino acids were frequently ignored by SEQUEST_SUMMARY due to their low Xcorr. The system was therefore configured for a high level of automation and robustness. The identification of smaller amounts of proteins, although possible with the same basic set up, required relaxation of the search parameters and the evaluation of borderline data by the operator.

The reliability and robustness of the LC-ESI-MS/MS protein workstation was demonstrated by the automated identification of a large number of yeast proteins. The recent completion of the yeast genome sequence (Goffeau et al., 1996) assured that essentially every protein sequence was represented in the database searched. High-resolution 2DE represented a convenient and powerful method for the separation of the proteins contained in the total yeast cell lysate and the methods for the mass spectral analysis of gel-separated proteins are well established (reviewed in Patterson & Aebersold, 1995). Batches of 40 samples each were analyzed in unattended runs of approximately 24 h duration. With the exception of two failed injections from the autosampler, 90 samples were processed reliably (Table 3). The failed injections were due to particles in the vial, possibly introduced when the needle of the autosampler punctured the septum. Using the described protocols, we did not observe a degradation of performance, neither of the

Table 3. Summary of large-scale analysis of yeast proteins by automated LC-ESI-MS/MS^a

Spot intensity (estimated)	Total spots analyzed	Total proteins identified
>1 μ g	17 of 18 ^b	21 proteins 2 samples: mixture of 2 proteins 1 sample: mixture of 3 proteins
100 ng–1 μ g	27 of 27	21 proteins
<100 ng	25 of 25	18 proteins
Pooled (2 spots per analysis)	9 of 10 ^b	19 proteins
100 ng–1 μ g		1 sample: mixture of 3 proteins

^aNinety digests were analyzed in an unattended manner by the LC-MS/MS protein workstation in two batches. Protein spots are indicated in Figure 4 and results are detailed in Tables 1 and 2.

^bSample was not injected.

HPLC column nor the MS instrument over time. In particular, the repeated injection of 20 μ L of 100 mM NH_4CO_3 did not interfere with the ionization source. Data analysis was performed on the same computer workstation that also acquired the MS/MS data and controlled the instrument. After completion of each run, SEQUEST was automatically launched in the background as a post-analysis application. The time required to search the complete yeast database using the conditions described was between 45 and 75 s. This search time could be reduced by two- to threefold if SEQUEST was instructed to only search for peptides that conformed to tryptic protein cleavage and not to consider methionine sulfoxide, a frequent artifactual protein modification. Because typically 10–15% of the analyzed peptides did not conform to tryptic cleavage and/or contained an oxidized methionine, the slower search time was well justified by the improved identification score. Data processing was usually complete within the 32-min run time of the next sample. For LC-MS/MS runs of abundant proteins, for which up to 80 CID spectra were generated, two batches of SEQUEST analyses were run in parallel without major apparent slowdown of the computer workstation. The results of both batches were collated in the SEQUEST_SUMMARY output.

Yeast proteins were identified by the combined results from database searches from individual peptides that matched to the same sequence in the database. This was accomplished by the program SEQUEST_SUMMARY. The program assigned decreasing scores from 10 for the best match to 2 for the fifth-best match and summed up the individual scores from peptides that matched the same protein sequence. A score of 20, i.e., the identification of at least two first-ranked peptides originating from the same protein, was considered statistically significant. In some instances, scores of 10, i.e., the identification of only one first-ranked peptide, were obtained for proteins. This was usually the case for faint spots containing less than 100 ng of protein. LC-MS/MS runs of such proteins were characterized by a small number of CID spectra, of which only a minority was found to contain the typical signature of peptide CID spectra. In these situations, the user determined the protein identity based on the available MS/MS data, as well as auxiliary information such as a match of the calculated pI and molecular mass with the respective values derived experimentally from the 2D gel.

In general, the calculated and experimentally observed pI and molecular mass values of the identified proteins were in good agreement. As also noticed by Shevchenko et al. (1996), the pI and mobility of some proteins deviated significantly from the expected values. In this study, this applied to heat shock proteins (spots 16, 22, 25, 28, and 63) and other abundant proteins, such as enolase (spots 38 and 40). Some of these aberrant mobilities may be due to artifactual proteolysis during cell lysis and sample preparation. It cannot be ruled out, however, that aberrant mobilities reflect specific protein processing events that might be important for protein function. Artifactual and intended proteolysis can be distinguished by comparing protein 2D patterns of lysates prepared under different conditions. Some proteins were also found at electrophoretic mobilities that suggested higher masses than those predicted from the amino acid sequences (spots 21, 24, and 27). Such a pattern is incompatible with artifactual proteolysis and might suggest extensive posttranslational modification. However, because electrophoretic mobility in SDS gels does not always correlate with the molecular mass of the polypeptide, a slower than expected mobility cannot necessarily be attributed to protein modifications. We also observed protein spots of similar electrophoretic mobility, but

different pI, that were the products of the same gene (spots 5, 6, 14, and 19; spots 7 and 9; spots 3 and 32; spots 1 and 39). Such spots were likely to represent differentially phosphorylated and/or glycosylated forms of the same polypeptide, although other modifications and short N- or C-terminal truncations cannot be ruled out. With moderate frequency, we detected the presence of two or more proteins co-migrating to a single spot. These were either unrelated proteins (spots 8, 16, and 57) or isoforms of a particular protein (spot 18: phosphoglycerate kinase isoforms 2 and 3).

The capacity to identify co-migrating proteins is limited by the limit of detection of the system. It is therefore likely that minor proteins co-migrating with identified proteins went undetected if their abundance fell below the limit of detection of the system. Protein identification techniques with higher sensitivity of detection have been developed that are suitable for characterizing co-migrating proteins of low abundance (Figeys et al., 1996a; Wilm et al., 1996). Currently, these systems are not compatible with high-throughput protein analysis.

The observed relatively high frequency of aberrant migration and co-migration of proteins in 2D gels suggests that, for large-scale protein identification projects, methods that use CID spectra of individual peptides for protein identification are preferred, probably even essential. The MS/MS-based approach eliminates ambiguities due to co-migrating proteins inherent in peptide mass mapping approaches (reviewed in Patterson & Aebersold, 1995) and the need for secondary criteria, such as protein molecular mass, pI, or N-terminal sequence tags. Such secondary criteria are used routinely to reduce the likelihood of false positive identifications, particularly if the identification is primarily based on amino acid composition analysis (Wilkins et al., 1996).

The inherent capability of SEQUEST to identify individual components in protein mixtures was used to further reduce the analysis time per protein by combining two spots from different areas of the gel prior to analysis. Similar to the 70 digests analyzed individually, some of the proteins identified in such deliberate mixtures represented degradation products (spots A1, C1, E2) or posttranslationally modified polypeptides (spot D1). The conclusive identification of proteins in mixtures was sometimes difficult if the observed and calculated mobilities differed significantly. An example of this situation are spots D1/D2. The spots were clearly separated by 0.6 pH units on the gel. However, according to their calculated mobilities, the two proteins should be almost co-migrating in the gel. It is therefore important for the success of this strategy to combine protein spots that differ significantly by both their pI and molecular mass (e.g., spots F1/F2 or G1/G2). The results from this large-scale identification of yeast proteins also indicate that the number of protein species observed by 2D gel electrophoresis and, by implication, by other high-resolution analytical techniques does not necessarily reflect the number of expressed genes.

The application of the system as described in this manuscript is limited to the identification of proteins for which a sequence has been deposited in a database. For many species, human and mouse in particular, EST sequence databases are growing more rapidly than genomic or full-length cDNA databases. The described system is compatible with the automated correlation of peptide CID spectra with specific ESTs. In most cases, protein identification through EST databases relies on a single match. According to the scoring criteria described above, a protein identification by means of a single match to an EST has to be considered tentative. To make such identifications conclusive, it is essential to carefully review the data obtained by the auto-

mated system for accuracy and, if possible, to invoke confirmatory secondary evidence.

Conclusion

In this manuscript, we demonstrate that peptide mapping by LC-ESI-MS/MS is a robust, sensitive, and conclusive method for the identification of proteins. With suitable software support, large-scale protein identification projects can be performed routinely and reliably in an automated manner.

Although presently less sensitive than the solid-phase extraction capillary zone electrophoresis MS/MS system we recently described (Figeys et al., 1996a, 1996b) or the nanospray MS/MS method developed by Mann and co-workers (Shevchenko et al., 1996; Wilm et al., 1996), the level of system automation makes this approach the method of choice for large-scale protein identification projects. This was demonstrated recently by a partial proteome analysis of *Haemophilus influenzae* (Link et al., 1997). Through implementation of capillary LC-MS technology pioneered by Simpson and co-workers (Moritz & Simpson, 1992) and Lee and co-workers (Davis et al., 1995) into the automated system, we anticipate significant sensitivity increases for automated protein identification. A first design for an automated protein workstation based on 100 μ m i.d. capillary RP column coupled to micro-electrospray MS/MS was presented at the ABRF '97 meeting in Baltimore (Ducret et al., 1997).

We anticipate that methods for the automated identification of proteins will grow in importance as more genome sequencing projects are being completed. The strategy described in this paper will contribute significantly to the current efforts aimed at the systematic identification of proteins, in proteome-type projects, as well as in the targeted investigation of biochemical processes.

Materials and methods

Chemicals

Methanol, acetonitrile, glutaraldehyde, formaldehyde, ethanol, and acetone were purchased from J.T. Baker (Phillipsburg, New Jersey). Water (18 M Ω) was purified by a MilliQ water purification device from Millipore (Burlington, Massachusetts). Acrylamide, DTT, urea, and glycine were from Millipore, and carrier ampholytes were purchased from Pharmacia Biotech (Piscataway, New Jersey). Molecular weight and isoelectric focusing standards were from Bio-Rad (Hercules, California) and nitrocellulose membranes (0.45 μ m) were obtained from Schleicher & Schuell (Keene, New Hampshire). Trifluoroacetic acid was purchased from Applied Biosystems (Foster City, California). DNase, RNase carbonic anhydrase II, leupeptin, and aprotinin were obtained from Sigma (St. Louis, Missouri). Modified porcine trypsin (sequence grade) was from Promega (Madison, Wisconsin). All other chemicals were obtained from Sigma and were of the highest grade available.

2DE (IEF/SDS-PAGE) separation of yeast cell extract

Cell extract preparation

Yeast cell extract was essentially prepared according to a protocol described by Jazwinski (1990). Briefly, yeast cells (*Saccharomyces cerevisiae* strain AAY1048) grown to saturation in 1 L of

YPD medium at 30°C were pelleted by low-spin centrifugation, rinsed with 1 L of water, and lysed in 10 mL of lysis buffer (50 mM Tris, pH 8.0, 5 mM Na₂EDTA, 1 mM Pefabloc, 50 mM NaF, 10 mM NaVO₄, 10 μ g/mL leupeptin, 10 μ g/mL aprotinin). An equal volume of glass beads (425–600 μ m, acid washed) was added and the suspension was vortexed five times for 1 min with intermediate cooling on ice. β -Mercaptoethanol (2%, v/v) and SDS (0.6%, w/v) were then added and the cells were lysed at 100°C for 5 min. After treatment with DNase/RNase and removal of cell debris by centrifugation, the supernatant was subjected to precipitation in acetone:H₂O (80:20, v/v) for 3 h at –20°C. Proteins were pelleted by low-spin centrifugation and resuspended in 4 mL of sample buffer [10 M urea, 4% (w/v) CHAPS, 65 mM DTT, 1% (v/v) Ampholine 3–10]. Insoluble material was removed by centrifugation and the cell extract (estimated concentration: 10 mg/mL) was aliquoted and kept frozen at –80°C until use.

2DE

First dimension. Isoelectric focusing was performed in a Pharmacia Multiphor unit at 20°C according to the manufacturer's specifications. Twenty microliters of cell extract were mixed with 80 μ L of sample buffer [10 M urea, 4% (w/v) CHAPS, 65 mM DTT, containing either 1% (v/v) Ampholine 3–10 or 0.5% (v/v) Ampholine 6–8 and 0.5% (v/v) Ampholine 8–10] and the resulting mixture was loaded with the sample cup onto the neutral end of an 18-cm Pharmacia DryStrip 4–7L or 6–10L. The strips were equilibrated overnight in 8 M urea, 10 mM DTT, 2% (w/v) CHAPS, containing either 2% (v/v) Ampholine 3–10 or 1% (v/v) Ampholine 6–8 and 1% (v/v) Ampholine 8–10. Proteins were focused under the following conditions: 1 h at 150 V constant; 3 h at 500 V constant; linear voltage gradient from 500 to 3,500 V over 5 h; 10 h at 3,500 V constant; 2 h at 5,000 V constant (total: 61.6 kVh). All electrode buffers were water, with the exception of the cathode buffer of the 6–10 gradient, which was 10 mM DTT (Görg et al., 1995).

Second dimension. SDS-PAGE was performed in a Millipore Investigator second dimension unit according to the manufacturer's specifications with the following modifications. The first dimension strips were removed from the Multiphor unit and immediately equilibrated for 5 min in the equilibration buffer [0.3 M Tris/HCl, pH 8.6, 4% (w/v) SDS, 10 mM DTT, 0.005% (w/v) Bromophenol blue]. The strips were then gently squeezed into a 2-cm concentrating gel (5% T, 2.67% C, 125 mM Tris/HCl, pH 8.8, 0.1% SDS) that had been layered onto a separating gel [20 cm $\times$ 20 cm, 15% T, 2.67% C, 0.375 M Tris/HCl, pH 8.8, 0.1% (w/v) SDS, 0.15 M sucrose]. Two sets of five gels each were run simultaneously at 16 W constant per gel for 5 h at 20°C with a 2 $\times$ Tris/glycine running buffer [384 mM glycine, 50 mM Tris-base, 0.2% (w/v) SDS].

Staining and blotting procedures

SDS gels were immediately stained with Coomassie Blue R250 or silver (Rabilloud et al., 1994) following electrophoresis. Alternatively, 15 $\times$ 15-cm squares were excised from gels and electroblotted against nitrocellulose overnight at 15 V following standard procedures. Proteins on membranes were stained with Amido Black.

In situ digestion of electroblotted proteins

Proteins blotted on nitrocellulose were digested in situ following a standard protocol (Aebersold et al., 1987; Ducret et al., 1996).

Briefly, nitrocellulose membranes were blocked for 1 h in 0.5% (w/v) PVP-40 dissolved in 100 mM acetic acid at room temperature before being washed extensively (5×1 h) in water. Spots with identical electrophoretic mobilities were pooled from three membranes, cut in pieces of approximately 1 mm^2 , and sampled in a 250- μL tube. Digestions were performed in 20 μL digestion buffer [100 mM NH_4HCO_3 , pH 7.8, 10% (v/v) acetonitrile] containing 1 μg trypsin for 3 h at 37°C . The supernatants were collected and the membrane pieces were rinsed with 10 μL of digestion buffer. The tubes were spun briefly and the supernatants of both extractions were pooled. Digests were either analyzed immediately or kept frozen at -20°C until use.

Instrumentation

The instrumentation used for the analysis of the yeast digests is represented schematically in Figure 1. Specific details about the instrument configuration are discussed in the following sections.

RP-HPLC separation of yeast digests

Up to 40 samples were loaded onto a refrigerated autosampler (Alcott model 738R, Norcross, Georgia) and injected sequentially onto a microbore HPLC system (Ultrafast Microprotein Analyzer, Michrom BioResources Inc, Auburn, California) equipped with a 100- $\text{\AA}$, 5- μm Reliasil C₁₈ BDX column ($0.5 \times 150 \text{ mm}$) purchased from the same supplier. Eighty percent (20 μL) of each sample was injected automatically. The time period from injection to injection was 32 min, and the following program was used [buffer A: 6% (v/v) acetonitrile, 0.025% (v/v) trifluoroacetic acid; buffer B: 80% (v/v) acetonitrile, 0.024% trifluoroacetic acid]: 0–3 min, 0–20% B at 15 $\mu\text{L}/\text{min}$; 3–21 min, 20–50% B at 15 $\mu\text{L}/\text{min}$; 21–21.5 min, 50–100% B from 15 to 25 $\mu\text{L}/\text{min}$; 21.5–24.5 min, 100% B at 25 $\mu\text{L}/\text{min}$; 24.5–25 min, 100–0% B at 25 $\mu\text{L}/\text{min}$; 25–30.5 min, 0% B at 25 $\mu\text{L}/\text{min}$; 30.5–31 min, 0% B from 25 to 15 $\mu\text{L}/\text{min}$; 31–32 min, 0% B at 15 $\mu\text{L}/\text{min}$. The autosampler lines were washed between injections with buffer A at approximately 5 $\mu\text{L}/\text{min}$ using a Flow Delivery Module (Michrom BioResource, Inc.) to reduce memory effects. Peptide elution was monitored by UV absorbance at 214 nm.

On-line analysis of peptides by ESI-MS/MS

Instrument configuration. The HPLC column outlet was connected directly to a Finnigan MAT TSQ 7000 triple quadrupole mass spectrometer (San Jose, California) equipped with the standard electrospray ionization source via a 75- μm i.d. fused silica capillary. The electrospray needle was set at a voltage of 5.0 kV and the sheath gas pressure (nitrogen, 99.997%) was 75 psi. The temperature of the heated capillary was 200°C . The second quadrupole was pressurized at 3.5–4.0 mTorr with Argon to allow CID of peptides on-the-fly. The electron multiplier was set to obtain approximately $3\text{--}4 \times 10^5$ counts total ion current (resulting in approximately 2,000–3,000 counts average background noise level) when scanning the mass range 400–2,000 at 1.6 s/scan with quadrupole 3 at starting conditions. The MS was calibrated and optimized using the conditions described above with singly charged synthetic peptides of masses 130–1,500, which were flow-injected at a flow rate of 5 $\mu\text{L}/\text{min}$. Ion transmission was maximized by setting quadrupole 1 and 3 at a resolution of 300 and 500, respectively, at mass 1,000. The usual mass tolerance for this setting was ± 1.0 mass unit.

Acquisition of mass spectra. Mass spectra were recorded by scanning the third quadrupole over the mass range 400–2,000 at a rate of 1,000 mass units per second throughout the HPLC gradient (0–26 min after injection). Under these conditions, undigested trypsin and hydrophobic, nonpeptidic material did not elute during the gradient, but were eluted during the column wash cycle.

Automatic acquisition of MS/MS spectra. Using the built-in ICL capabilities of the instrument, we wrote a procedure to allow the automatic MS/MS analysis of eluting peptides (Fig. 7). When the current for a particular peptide ion reached a preset threshold of 70,000 counts, the parent was selected by the first quadrupole and was subjected to low-energy CID in quadrupole 2 over six scan periods. The collision energy factor, *coff*, was empirically set according to the relationship $\text{coff (in eV)} = -(16.44 \times 10^{\text{mass-of-peptide} \times 3.269 \times 10^{-4}})$ and was slightly increased over the scans to enhance the production of low mass fragment ions, which are important for the interpretation of peptide MS/MS spectra. Resulting MS/MS spectra were recorded by scanning quadrupole 3 between the mass ranges 50 and twice the mass of

Fig. 7. Schematic representation of strategy used for automated protein identification. CID spectra generated by the LC-ESI-MS/MS analysis of a protein digest are processed individually by SEQUEST and the database search results are recorded. After this initial analysis, SEQUEST-SUMMARY ranks the five best matches obtained for each MS/MS spectrum by assigning a score of 10 for a first-ranked match, 8 for a second-ranked match, 6 for a third-ranked match, 4 for a fourth-ranked match, and 2 for a fifth-ranked match, provided the respective Xcorr exceed 2.0. The scores of peptides originating from a common protein are added. Total score and protein identity(ies) are displayed in a summary report for each LC-MS/MS run.

the parent ion + 20 at a scan rate of 1,000 mass units per second before returning the MS in full scan mode. The procedure prevented two consecutive MS/MS spectrum analysis of the same parent ion and thus permitted the analysis of co-eluting peptides. A version of the ICL procedure is available on the Internet at the address http://weber.u.washington.edu/~mbt/groups/aebersold/methods/icl_proc.html.

Identification of proteins by SEQUEST

MS/MS spectra were analyzed with the program SEQUEST, which allows the correlation of experimental data with theoretical spectra generated from known protein sequences (Eng et al., 1994; Yates et al., 1995). The program was run on the same local workstation, which controlled the MS and data acquisition. Typically, the program was launched in the background when data acquisition was terminated. Automation was achieved by a small UNIX script file that was run as a postacquisition application from the Finnigan analysis window. The script contained the same commands that would otherwise be typed manually by the user to process the MS/MS spectra of the current acquisition file (conversion of the Finnigan MS/MS data to text files, launching of the SEQUEST and SEQUEST_SUMMARY programs and printing of the resulting report file), except that they were executed as a batch in the background. Rapid processing was achieved by searching a small yeast protein database (8,877 entries) that was extracted from the nonredundant protein database OWL version 29.1 (Bleasby et al., 1994).

The automatic identification of proteins by SEQUEST is described schematically in Figure 8. Briefly, SEQUEST selected from the protein database any peptides (i.e., SEQUEST was not specified that peptides were generated by a tryptic digest) matching to within two mass units of the mass of the precursor ions. In addition, SEQUEST dynamically searched the database for methionine sulfoxides (+16.0) to detect this common oxidation product. A preliminary score was obtained by comparing numerically the presence of the peptides fragment ions in the experimental spectrum. Theoretical CID spectra of the 500 best preliminary-scoring peptides were then generated and were cross-correlated against the experimental MS/MS spectrum. For tryptic cleavage, a true positive was generally characterized by a peptide bearing a C-terminal Lys or Arg residue, a high preliminary score, an Xcorr greater than 2.0, and a ΔC_n greater than 0.1 compared with the second-best ranked peptide (Eng et al., 1994). Because of the high number of CID analyses that were generated in a single LC-MS/MS run (typically around 30–50), a second program, SEQUEST_SUMMARY, was appended to SEQUEST for rapid viewing of the results. The program gave a score of 10 for every first-ranked match, a score of 8 for every second-ranked match, a score of 6 for third-ranked match, a score of 4 for fourth-ranked match, and a score of 2 for a fifth-ranked match, provided their Xcorrs were greater than 2.0. The scores of peptides originating from a common protein were then summed and the program listed the target(s) that were identified in the LC-MS/MS run. Because each CID analysis gave rise to an independent identification by SEQUEST, a score of 20 or higher (that is, the identification of at least two first-ranked peptides originating from the same protein) was considered as statistically relevant and identified a target protein. However, in some few instances, a score of 10 (i.e., the identification of one first-ranked peptide) was used to identify a protein if the sample applied to the HPLC was known to be close to the limit of

Fig. 8. Schematic representation of the ICL protocol for automated acquisition of MS/MS spectra. The main procedure, AutoCID, is launched as the acquisition is started by the injection of a sample by the autosampler. Each scan is screened for an ion of intensity greater than a user-given threshold. m/z values analyzed previously in the same LC-MS/MS run are excluded from further analysis. If an ion meets these conditions, the procedure GetCID is launched. The ion of interest is selected by the first quadrupole and fragmented with increasing energy in subsequent scans in the collision cell. The MS/MS spectra are acquired by scanning the third quadrupole between m/z 50 and twice the apparent mass of the parent ion plus 20 over a period of six scans. The MS is then returned in full scan mode and the main procedure AutoCID screens for the next ion of interest. The ICL procedures are automatically completed upon termination of the acquisition.

detection of the system and if secondary criteria, such as protein pI and molecular mass, were available. A commercial version of the SEQUEST and SEQUEST_SUMMARY programs can be obtained from Finnigan MAT (San Jose, California).

Acknowledgments

We are grateful to Dr. Bengt Bjellqvist (Pharmacia Biotech, Sweden) for the immobilized pH gradient 6–10L used in this study. This work was supported by the Science and Technology Center at the Department of Molecular Biotechnology, the University of Washington, Seattle, Washington.

References

- Aebersold RH, Leavitt J, Saavedra RA, Hood LE, Kent SBH. 1987. Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proc Natl Acad Sci USA* 84:6970–6974.
- Anderson NG, Anderson NL. 1996. Twenty years of two-dimensional electrophoresis: Past, present and future. *Electrophoresis* 17:443–453.
- Bleasby AJ, Akrigg D, Attwood TK. 1994. OWL—A non-redundant composite protein sequence database. *Nucleic Acids Res* 22:3574–3577.
- Cottrell JS. 1994. Protein identification by peptide mass fingerprinting. *Peptide Res* 7:115–124.

- Davis MT, Stahl DC, Hefta SA, Lee TD. 1995. A microscale electrospray interface for on-line, capillary complex peptide mixtures. *Anal Chem* 67:4549–4556.
- Ducret A, Foyn Brunn C, Bures EJ, Marhaug G, Husby G, Aebersold R. 1996. Characterization of human serum amyloid A protein isoforms separated by two-dimensional electrophoresis by liquid chromatography/electrospray ionization tandem mass spectrometry. *Electrophoresis* 17:866–876.
- Ducret A, Gu M, Haynes PA, Yates JR III, Aebersold R. 1997. Simple design for a capillary liquid chromatography–microelectrospray–tandem mass spectrometric system for peptide mapping at the low femtomole sensitivity range. *Proceedings of the ABRF'97 international symposium: Technique at the genome/proteome interface*:69.
- Eng J, McCormack AL, Yates JR III. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Mass Spectrom* 5:976–989.
- Figeys D, Ducret A, Yates JR, Aebersold R. 1996a. Protein identification by solid phase microextraction–capillary zone electrophoresis–microelectrospray–tandem mass spectrometry. *Nature Biotech* 14:1579–1583.
- Figeys D, van Oostveen I, Ducret A, Aebersold R. 1996b. Protein identification by capillary zone electrophoresis/microelectrospray ionization–tandem mass spectrometry at the subfemtomole level. *Anal Chem* 68:1822–1828.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Lacq L, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SB. 1996. Life with 6000 genes. *Science* 274:546–567.
- Görg A, Boguth G, Obermaier C, Posch A, Weiss W. 1995. Two-dimensional polyacrylamide gel electrophoresis with immobilized pH gradients in the first dimension (IPG-Dalt): The state of the art and the controversy of vertical versus horizontal systems. *Electrophoresis* 16:1079–1086.
- Haller I, Mirza UA, Chait BT. 1996. Collision induced decomposition of peptides. Choice of collision parameters. *J Am Mass Spectrom* 7:677–681.
- Jazwinski M. 1990. Preparation of extracts from yeast. *Methods Enzymol* 182:154–174.
- Link AJ, Hays LG, Carmack EB, Yates JR III. 1997. Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* 18:1314–1334.
- Mann M, Wilm M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 66:4390–4399.
- Moritz RL, Simpson RJ. 1992. Application of capillary reversed-phase high-performance liquid chromatography to high-sensitivity protein sequence analysis. *J Chromatogr* 599:119–130.
- Patterson SD, Aebersold R. 1995. Mass spectrometric approaches for the identification of gel-separated protein. *Electrophoresis* 16:1791–1814.
- Rabilloud T, Vuillard L, Gilly C, Lawrence JJ. 1994. Silver-staining of proteins in polyacrylamide gels: A general overview. *Cell Mol Biol* 40:57–75.
- Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortenson P, Shevchenko A, Boucherie H, Mann M. 1996. Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci USA* 93:14440–14445.
- Stahl DC, Swiderek KM, Davis MT, Lee TD. 1996. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J Am Mass Spectrom* 7:532–540.
- Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphrey-Smith I, Hochstrasser DF, Williams KL. 1996. Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotech Gen Eng Reviews* 13:19–50.
- Wilm M, Shevchenko A, Houthaeve T, Breit S, Schweigerer L, Fotsis T, Mann M. 1996. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* 379:466–469.
- Yates JR III, Eng JK, McCormack AL, Schieltz D. 1995. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 67:1426–1436.