ELSEVIER

Journal of Chromatography B, 745 (2000) 197–210

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Strategy for qualitative and quantitative analysis in proteomics based on signature peptides

Junyan Ji, Asish Chakraborty, Ming Geng, Xiang Zhang, Ahmad Amini, Minou Bina, Fred Regnier*

Department of Chemistry, Purdue University, Lafayette, IN 47907, USA

Abstract

This paper describes a new analytical strategy for identifying proteins in concentration flux based on isotopic labeling peptides in tryptic digests. Primary amino groups in peptides from control and experimental samples were derivatized with acetate and trideuteroacetate, respectively. After mixing samples thus labeled from these two sources, the relative concentration of peptides was determined by isotope ratio analysis with MALDI and ESI mass spectrometry. More than a 100-fold difference in relative concentration could be detected. Simplification of complex tryptic digests prior to mass spectral analysis was achieved by selection of histidine-containing peptides with immobilized metal affinity sorbents or of glycopeptides by lectin columns. Because most of these peptides have sequences that are unique to a single protein, they are a signature of the protein from which they were derived; providing a facile route to protein analysis.

Keywords: Proteomics; Signature peptides; Proteins

1. Introduction

Molecular biology and molecular medicine have as their focus the explanation of biological phenomena in terms of molecular structure. This has led to the enormous effort to identify all the molecular elements of biological systems and the mechanism by which they function. The Human Genome Project [1] and proteomics [2] are both examples of efforts to define the molecular elements of biological systems and understand how they interact. Within the next 5–10 years it is likely that the world will identify most of the ''molecular players'' in humans, domestic plants and animals, some pathogens, and many common microorganisms. Yet with all this, we will still not understand the dynamics of how living things respond to stimuli. Critical elements of homeostasis and how systems succumb to diseases will evade us in most cases, unless we also know how biological systems are regulated. A major component of understanding cellular regulation is in being able to identify proteins in flux in the complex milieu of the proteome.

Two approaches for examining cellular dynamics have been suggested. One is to follow genetic expression through the synthesis of specific mRNA species [3,4]. This approach assumes that most changes of protein concentration are the result of de novo protein synthesis. The second is by identifying changes in specific proteins themselves. It has been found that the correlation between the two is poor [5]. This is not surprising. Changes in protein concentration in response to regulatory stimuli are

*Corresponding author.

E-mail address: fregnier@purdue.edu (F. Regnier).

0378-4347/00/$ – see front matter
PII: S0378-4347(00)00192-4

frequently achieved through post-translational modifications not tightly coupled to expression. Direct measurement of the concentration of specific proteins is obviously the most definitive approach to the study of cellular regulation.

Cells may express a few thousand to 20 000 proteins, depending on the cell type [6]. Determining which protein(s) changed in response to a stimulus is like looking for a needle in a haystack. Elegant studies using two-dimensional gel electrophoresis have shown that regulatory changes in a few to several hundred proteins can be seen in the *E. coli* proteome [7]. The power of two-dimensional gel electrophoresis in such studies is that under proper circumstances it can resolve 4000 to 6000 protein components [8]. Relative changes in the concentration of specific proteins among samples are generally measured by comparing the difference between two gels based on staining; one from the control and the other from the experimental trial. The possibility that the concentration of different proteins in regulatory flux varies 10^3–10^4 in the same gel exaggerates the detection and quantitation problem. It is very difficult to locate small amounts of protein in a gel by staining. Comparing concentrations between two gels has a very high level of uncertainty. The current protocol for identifying proteins in gels is as follows. Subsequent to locating protein components, either by staining or autoradiography of biosynthetically labeled species, they are (i) excised from the gel, (ii) reduced and alkylated, and (iii) trypsin digested. The tryptic digest is then subjected to matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) and the parent protein can often be identified by matching the mass of peptides against a sequence database [9]. This process frequently takes 2–5 days, depending on the exact protocol. When peptides cannot be identified from a sequence database, they must be at least partially sequenced, generally by MS–MS. Limitations of this approach are quantification, reproducibility, difficulty in dealing with very high molecular mass and basic proteins, the process is lengthy, and it is difficult to automate.

This paper proposes a new strategy for recognizing changes in the concentration of a protein when it is in a complex milieu of compounds having similar structures. The technique is based on proteolysis of proteins and isotopic labeling of the resulting peptides. Following selection of labeled peptides containing rare amino acids or post-translational modifications, the peptide mixtures were fractionated chromatographically and fractions analyzed by MALDI-MS. This approach to identifying peptides in flux is based on changes in isotope ratios of peptides differentially labeled in control and experimental samples. Peptide identification was achieved by comparing experimentally derived properties to those of tryptic peptides in sequence databases.

2. Materials and methods

2.1. Materials

N-Hydroxysuccinimide, *N*-acetoxysuccinimide, monobasic sodium phosphate, dibasic sodium phosphate, tris(hydroxymethyl)aminomethane (Tris base), iodoacetic acid, 4-vinyl pyridine, tris(hydroxymethyl)aminomethane hydrochloride (Tris acid), cysteine, dithiothreitol (DTT), *N*-tosyl-L-lysyl chloromethyl ketone (TLCK), α-cyano-4-hydroxycinnamic acid, *N*-acetylglucosamine (NacGlc) were purchased from Sigma (St. Louis, MO, USA). Acetic-d_3-anhydride was purchased from Aldrich (Milwaukee, WI, USA). All the peptides used were purchased from Bachem Bioscience (CA, USA). Nuclear extracts were prepared from human cell line U937 as previously described [10]. High-performance liquid chromatography (HPLC)-grade trifluoroacetic acid (TFA) was purchased from Pierce (Rockford, IL, USA). HPLC-grade water and acetonitrile (ACN) were purchased from EM Science (Gibbstown, NJ, USA). All reagents were used directly without further purification. *Escherichia coli* was cultured in our laboratory.

2.2. Synthesis of N-acetoxy-d_3-succinimide

A solution of 4.0 g (34.8 mmol) of *N*-hydroxysuccinimide in 10.7 g (105 mmol) of d_3-C^1 acetic anhydride was stirred at room temperature. White crystals began to deposit in 10 min. After 15 h the liquid phase was allowed to evaporate and the crystalline residue was treated with hexane and dried

in vacuum. Product yield was 5.43 g (100%), m.p. 133–134°C.

2.3. Acetylation of peptides

A three-fold molar excess of *N*-acetoxysuccinimide and *N*-acetoxy-d_3-succinimide was added individually to the two equal aliquots of 1 mg/ml peptide solution in phosphate buffer at pH 7.5, respectively. The reaction was carried out at room temperature. After stirring 4–5 h, equal aliquots of the two samples were mixed and purified on a C_{18} reversed-phase chromatography column. Collected fractions were then subjected to MALDI–time-of-flight (TOF)-MS and electrospray ionization (ESI)-MS.

2.4. MALDI–TOF-MS

MALDI–TOF-MS was performed using a Voyager DE-RP BioSpectrometry workstation (PE Biosystems, Framingham, MA, USA). Samples were prepared by mixing a 1-μl aliquot with 1 μl of matrix solution. A 1-μl sample volume was spotted into a well of the MALDI sample plate and allowed to air-dry before being placed in the mass spectrometer. The matrix for acetylated peptides was a solution of 3% TFA, ACN–water (50:50) solution saturated with α-cyano-4-hydroxycinnamic acid. The matrix solution for glycopeptides was prepared by saturating a water–ACN–TFA (50:47:3) solution with sinipinic acid. Peptide quantitation was performed with MALDI–TOF-MS in the linear, positive ion and reflector mode by delayed extraction using an accelerating voltage of 20 kV. Ten spectra were collected from each sample spot and the peak intensities averaged for each spot. A linear equation was deduced from the ion current intensity ratio of the deuterium-labeled and the unlabeled acetylated peptides versus the ratio of the amount of these two peptides.

2.5. Quantification of peptides by LC–ESI-MS

An Integral HPLC micro-analytical workstation was linked to the electrospray interface of a Mariner ESI mass spectrometer (PE Biosystems) by an auxiliary six-port splitter valve. Analyte was introduced into the sample loop and eluted with a 5 min mobile phase gradient ranging from 0.2% formic acid to aqueous 56% acetonitrile in 0.2% formic acid, followed by a 2 min isocratic elution with 80% ACN in 0.2% formic acid. Nitrogen was used as spray gas. Spectra were acquired at the rate of 2 s per spectrum with an ion count threshold of 1.

2.6. Proteolysis

Human serotransferrin (5 mg), nuclear extract from cell line U937, human serum or *E. coli* were reduced and alkylated in 1 ml of 0.2 *M* Tris buffer (pH 8.5) containing 8 *M* urea and 10 m*M* DTT. After a 2-h incubation at 37°C, iodoacetic acid (vinyl pyridine was used for ovalbumin) was added to a final concentration of 20 m*M* and incubated in darkness on ice for 2 more hours. Cysteine was then added to the reaction mixture to a final concentration of 40 m*M* and the reaction allowed to proceed at room temperature for 30 min. After dilution with 0.2 *M* Tris buffer to a final urea concentration of 3 *M*, TLCK-treated trypsin (2%, w/w, enzyme to that of protein) was added and incubated for 24 h at 37°C. Digestion was stopped by adding TLCK in slight molar excess to that of trypsin.

2.7. Affinity selection of glycopeptides by a two-dimensional chromatographic method

All chromatographic steps were performed using an INTEGRAL Micro-Analytical Workstation from PE Biosystems. Tryptic digested human serotransferrin (0.1 ml) was injected onto a Con A affinity column that had been equilibrated with a loading buffer containing 1 m*M* $CaCl_2$, 1 m*M* $MgCl_2$, 0.2 *M* NaCl and 0.1 *M* Tris–HCl (pH 7.5). The Con A column was eluted sequentially at 1 ml/min with two column volumes of loading buffer and then 0.2 *M* methyl-α-D-mannopyranoside in 0.1 *M* Tris (pH 6.0). Glycopeptides were directed to a 250×4.6 mm Peptide C_{18} (PE Biosystems) analytical reversed-phase HPLC column, which had been equilibrated for 5 min at 1.0 ml/min with 5% ACN containing 0.1% aqueous TFA. The glycopeptides were then eluted at 1.0 ml/min in a 35 min linear gradient to 50% ACN in 0.1% aqueous TFA. Eluted peptides were monitored at 220 nm and fractions manually

collected for MALDI–TOF-MS analysis. A 100-μl volume of tryptic digested nuclear extract was injected into a *Bandeiraea simplicifolia* (BS-II) lectin affinity column. The nonglycopeptides were washed away and the glycopeptides were transferred to the reversed-phase column. Washing with 20 ml starting buffer regenerated the BS-II column. The glycopeptides were separated by a 50 min gradient from 0.1% TFA in aqueous 1% ACN to 0.1% TFA in 90% ACN. The glycopeptides were manually collected and lyophilized before storage.

2.8. Partial sequence of peptide mixture by carboxypeptidase

A carboxypeptidase sequencing kit (PE Biosystems) was reconstituted and the reagent separated into five different solutions varying 10-fold in concentration. Carboxypeptidase concentration in the most concentrated solution was 1 pmol/μl. Lyophilized glycopeptides from nuclear extracts were reconstituted in 5 μl of water. A 0.5-μl sample aliquot was deposited on the MALDI plate at each sample well. In total, six deposits of each sample were made on the MALDI plate. After solvents had evaporated from the deposited samples, carboxypeptidase solutions (0.5 μl) of varying concentration were deposited on five of the sample deposits. Carboxypeptidase digestion of peptides was performed directly on the MALDI plate. Again the solution deposited on the plate was allowed to air dry. After water had evaporated from the digesting sample on the MALDI plate, 0.5 μl of matrix solution saturated with α-cyano-4-hydroxycinnamic acid was added to each dried spot.

2.9. Immobilized metal affinity chromatography

Immobilized metal affinity chromatography (IMAC) was performed using a copper-saturated 4.6 mm diameter POROS MC column (PE Biosystems). A 90-min gradient from 1 to 20 m*M* imidazole in 20 m*M* sodium phosphate containing 0.5 *M* sodium chloride, pH 7.0 was used for this study. A 70-min gradient from 1% ACN to 95% ACN in 0.1% TFA was used to elute peptides from the reversed-phase column. All the buffers and solutions were prepared in distilled water, degassed and filtered through a 0.2-μm filter. Tryptic digested peptides were acetylated with a three-fold molar excess of *N*-acetoxysuccinimide in phosphate buffer at pH 7.5.

3. Results and discussion

Recognizing concentration changes in biological systems involving small numbers of species in large numbers of similar, fixed components is a major problem. This is particularly important when the objective is to examine changes in the cellular milieu in response to disease or external chemical stimuli, such as drugs. The challenge is exacerbated by the fact that the components in flux are frequently unknown. In fact, recognizing that a substance is in concentration flux would generally precede identification. A strategy is proposed below for both recognizing and identifying proteins in flux in complex systems.

3.1. The internal standard strategy for quantification of proteins

Internal standard quantification is frequently used to determine the concentration of substances in complex systems. This technique is based on the addition of a known substance, similar to the analyte in structure, and determining analyte concentration by ratio, i.e.:

$$\Delta = \frac{AR}{\Lambda\Re} \tag{1}$$

where Δ is the relative difference in concentration, A is instrument response to the analyte, R is specific molar response to the analyte, Λ is instrument response to the internal standard, and $\Re$ is specific molar response to the internal standard. When the analyte and internal standard only differ isotopically, it may be assumed that a mass spectrometer will respond equally to both and. In this case, Eq. (1) reduces to $\Delta = A/\Lambda$ and relative difference in analyte and isotopically labeled internal standard concentration may be determined from peak height ratios in the mass spectrum.

The problem with this approach to quantification in proteomics is that the concentration of only a small number of proteins will change in response to

a stimulus and the identity of these proteins is seldom known. How does one create internal standards for unknowns? One solution is to create internal standards for all polypeptides in a sample. Because the objective is to recognize changes in protein concentration between control and experimental samples, all the components in the control could be used as internal standards for the experimental sample if they were labeled. The issue is how to label every polypeptide in a sample. Although metabolic incorporation of labeled amino acids has been widely used to label polypeptides, it is not very reproducible and is objectionable in human subjects. Post-sampling strategies for incorporation of labels would be much more attractive.

Proteolysis is widely used in protein chemistry to identify proteins [11]. During the course of trypsin digestion, peptide bonds in proteins are cleaved on the C-terminal side of basic amino acids. This produces a large number of peptides ranging in size from a few to roughly 20 amino acids, all with an N-terminal amino group. The single exception would be peptides arising from N-terminally derivatized proteins. Primary amino groups at the N-terminus and on lysine residues of peptides react readily with a variety of reagents and are easily labeled [12]. The labeling strategy used in these studies is illustrated in Fig. 1. The labeling agent, *N*-acetoxysuccinimide [13], was chosen because it can be used to acetylate peptides in water. Acetylating peptides from both the control and experimental samples was used to create internal standards of peptides from tryptic digests. The general strategy was to acetylate internal standard peptides with C^1H_3-CO–. This procedure was reversed in some cases where the internal standard appeared in the mixture at much higher concentrations and its isotope peaks are at the same mass as the analyte.

MALDI mass spectra of several model peptides (Fig. 2) show the spectrum of any particular peptide

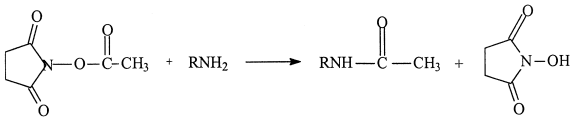

Fig. 1. The derivatization strategy used in labeling peptides.

to be a doublet. This is because analyte peptides are acetylated at the amino-termini and on ϵ-amino groups of lysines with CD_3-CO– whereas the internal standards are acetylated with the CH_3-CO– moiety. Arginine residues are not acetylated by this procedure. Analyte and internal standard peaks of peptides that contain no lysine will differ by 3 amu. For each lysine that is added to a peptide, the difference in mass increases an additional 3 amu. MALDI of peptides in the positive ion mode is most easily achieved when they have one or more positively charged groups. When they do not, peptides generally acquire charge from sodium or potassium ions in the system. This is the case with peptides that have neither histidine nor arginine. Subsequent to acetylation of amino-termini and ϵ-amino groups of lysine, they are no longer positively charged and cationize with some combination of sodium or potassium in MALDI mass spectrometry (Fig. 2b). This is not a problem in terms of the mass spectrometry, but it complicates data interpretation. When one is trying to identify the peptide based on mass, it is necessary to know whether it is cationized. Peptides lacking positive charge produce spectra with the expected mass in the negative ion mode of ionization if they contain a free carboxyl (data not shown).

Isotope ratios may be quantified with either MALDI or ESI-MS. It is seen in Fig. 3 that both the accuracy and dynamic range of ESI is superior to that of MALDI. However, MALDI has the advantage of being able to accommodate more complex samples. This is an advantage when dealing with biological extracts that may have more than 50 peptides in a single fraction.

Isotope ratio quantification has yet to be applied to cells in flux. When it is, it is expected that there will be no change in the relative concentration of most proteins between control and experimental samples. Thus Δ for peptides derived from these proteins will be the same in probably $>95\%$ of all cases. In a situation like this where only a small percentage of polypeptides are in regulatory flux, the average value of the peak height ratio for all components (Δ_{av}) will be nearly the same as that of an individual component that does not change. Any protein or peptide that deviates substantially from Δ_{av} is either being up- or down-regulated. The degree to which con-

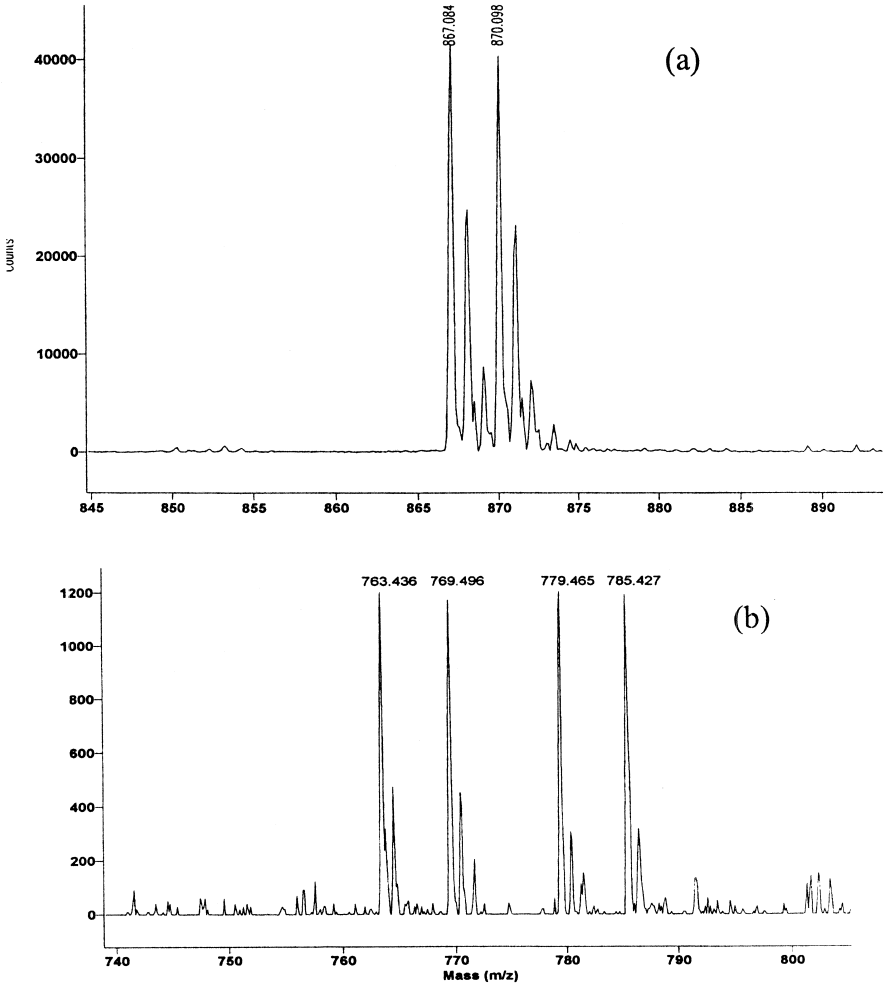

Fig. 2. MALDI mass spectra of acetylated model peptides.

centration has changed may be determine by the equation:

$$h_e = h_c \Delta_{av} \delta \quad (2)$$

where h_e is peak height of the component from the experimental sample, h_c is peak height from the control, and δ is the relative degree of up- or down-regulation.

3.2. The signature peptide approach to peptide and protein identification

It has been suggested above that using internal standards of peptides derived from proteins will be a valuable strategy for recognizing and quantifying species in regulatory flux. The issue with this approach is how to identify the source of a peptide after it has been recognized as having changed. Obviously, knowing the sequence would be of great value. But how can the sequence of small quantities of peptides in complex mixtures be determined?

The massive, ongoing effort to sequence the human genome in addition to those of most domestic species and pathogens is producing databases that are of enormous value in protein chemistry. Using genomic databases it is possible to predict the tryptic peptides that will be derived from each protein. Because the sequence coding for most of these

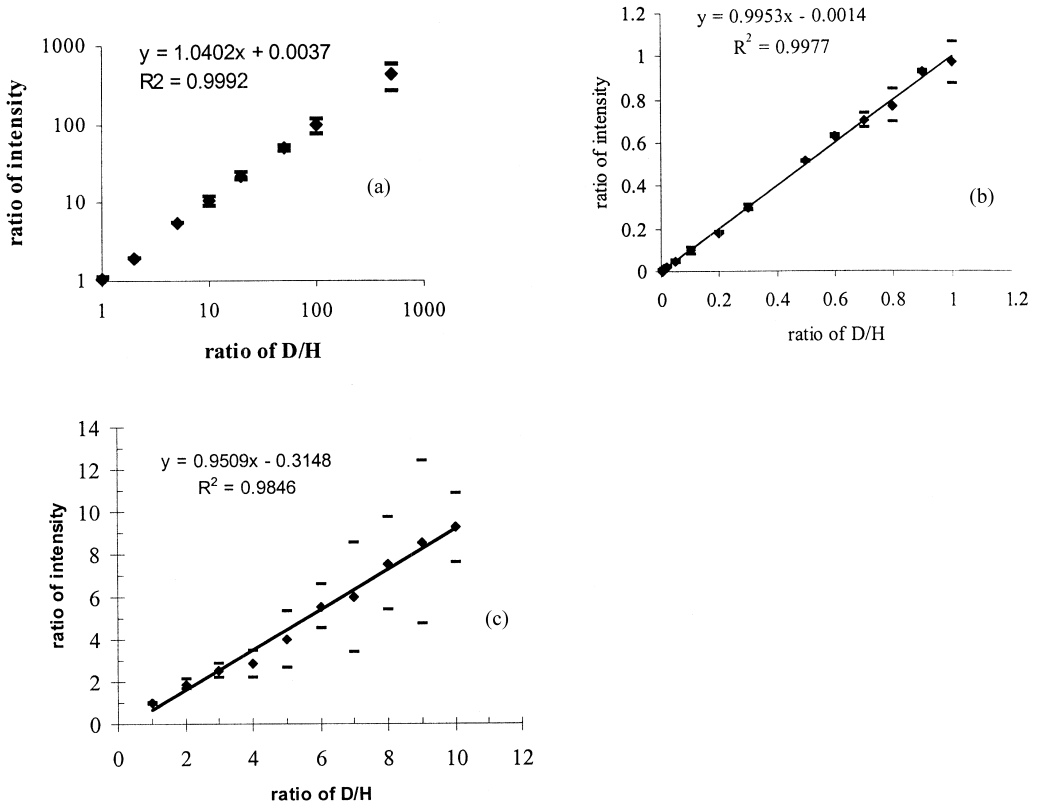

Fig. 3. Isotope ratio response curve of H–Ala–Ser–His–Leu–Gly–Leu–Ala–Arg–OH in electrospray ionization (a and b) and matrix assisted laser desorption ionization (c) mass spectrometry.

peptides is found only once in the genome, they are a signature of the protein from which they are derived. The dilemma is how to get sequence data from peptides in tryptic digests beyond the fact that they contain a C-terminal lysine or arginine. The problem is confounded further by the fact that tryptic digestion of all the proteins in a cell will produce thousands of peptides. The complexity of these mixtures is beyond the resolving power of liquid chromatography and mass spectrometry systems as has been shown in the ''shot-gun'' approach to peptide analysis in proteomics [14]. The shot-gun approach examines tryptic digests of cells by two-dimensional ion-exchange and reversed-phase chromatography followed by mass spectrometry.

The strategy chosen here was to simplify the tryptic digest by selecting peptides that contain either rare amino acids or sites of post-translational modification. Selecting for specific amino acids, or sequences, also adds to the knowledge of amino acid composition. Although it is potentially possible to select tryptic peptides that contain cysteine, tryptophan, methionine, histidine, tyrosine phosphate, serine phosphate, threonine phosphate, O-linked oligosaccharides, or N-linked oligosaccharides, only those containing histidine or glycosylation sites were studied.

3.2.1. Selecting histidine-containing peptides

IMAC is an established technique for separating histidine-containing polypeptides [15]. Copper loaded IMAC columns were used in these studies because they can bind histidine-containing peptides. Although cysteine and tryptophan-containing peptides also bind to copper loaded IMAC columns, the binding of these peptides may be controlled in several ways. Sulfhydryl alkylation eliminates cysteine binding. Binding of tryptophan-containing pep-

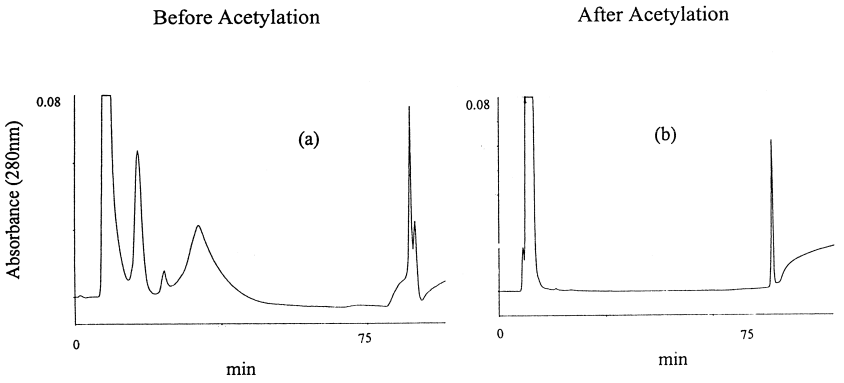

Fig. 4. IMAC chromatograms of native and acetylated tryptic peptides from ovalbumin.

tides is eliminated by acetylation of the N-terminal amino acid [16]. A similar observation was made in these studies with tryptic digests. Subsequent to acetylation of a tryptic digest of ovalbumin, early eluting peptides were eliminated from the IMAC chromatogram (Fig. 4a and b). In this respect, the reduction, alkylation, and acetylation procedures used above in quantification fortuitously enhance the selectivity of IMAC columns for histidine-containing peptides. MALDI mass spectra of ovalbumin peptides recovered from the IMAC column are seen in Fig. 5. All of these peptides contain histidine as expected.

Peptide fractionation of *E. coli* trypsin digests was achieved with a multidimensional chromatography system using IMAC and reversed-phase chromatography (RPC) columns. Elution of both the IMAC and RPC columns was accomplished with a continuous gradient. All peptides eluting from IMAC column were directly transferred to the RPC column where they were reconcentrated at the inlet of the RPC column. The only peptides that might not have been captured at the head of the RPC column would be small hydrophilic peptides of two or three amino acids. According to searches of the *E. coli* database this is not a problem because peptides of this size are

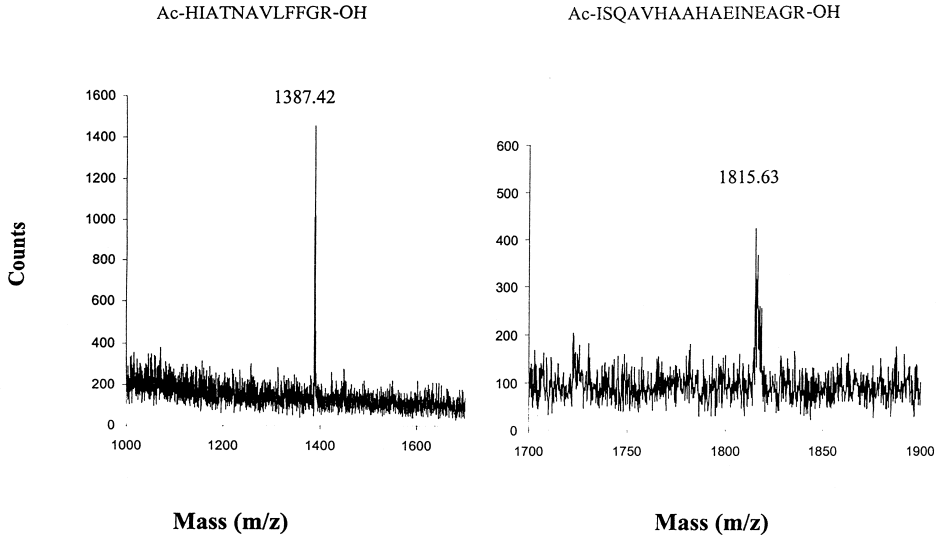

Fig. 5. MALDI mass spectra of histidine-containing peptides of ovalbumin recovered from the IMAC column.

generally not structurally unique. After sample transfer, the IMAC column was switched off-line, the solvent lines of the chromatography instrument purged at 10 ml/min for a few seconds with RPC solvent A, and finally the RPC column was gradient eluted at 1 ml/min. Fractions were collected from the RPC column and spotted onto a MALDI plate. After addition of MALDI matrix and solvent evaporation, samples were analyzed manually in a MALDI mass spectrometer.

The reversed-phase chromatogram of an *E. coli* tryptic digest fraction eluted from the IMAC column is seen in Fig. 6. Although the chromatogram does not appear to be enormously complex, database searches of the *E. coli* genome indicate that a thousand or more histidine-containing peptides could be contained in a single IMAC fraction. This would mean that individual fractions collected from the RPC column could contain 10, or more components. The apparent simplicity of the chromatogram probably results from very wide differences in peptide concentration in the sample. Concentration of the parent proteins from which these signature peptides were derived is thought to vary 10^4, or more [17]. Many of the peptides in the reversed-phase chromatogram (Fig. 6) are at too low concentration to be seen.

How can fractions containing 10 or more peptide components provide data of analytical value? Fortunately, MALDI-MS can accommodate samples of 50 or more peptides. Mass spectra of several representative fractions taken from the RPC column are shown in Fig. 7. One of the great advantages of MALDI-MS is that peptides are seldom multiply charged and when they are the abundance is low. Peaks in the spectra are almost exclusively of the parent ion. Multiple peaks in these MALDI-MS spectra indicate the presence of multiple peptides. However, it is possible that not all peptides in the sample are seen in the MALDI-MS spectra. It is known in tryptic digests of proteins containing a hundred or more peptides that some of the peptides are not seen in MALDI-MS due to quenching [18]. Mixtures encountered in these studies did not appear to be of

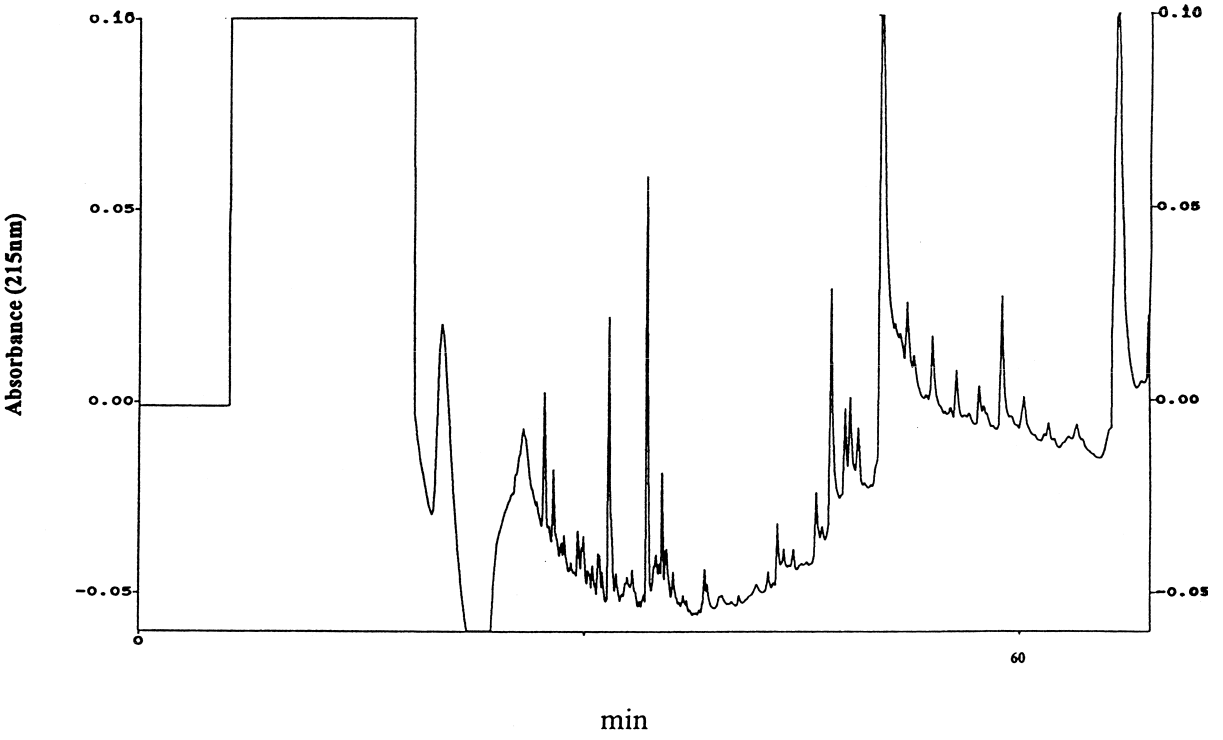

Fig. 6. The reversed-phase chromatogram of an *E. coli* tryptic digest fraction eluted from the IMAC column.

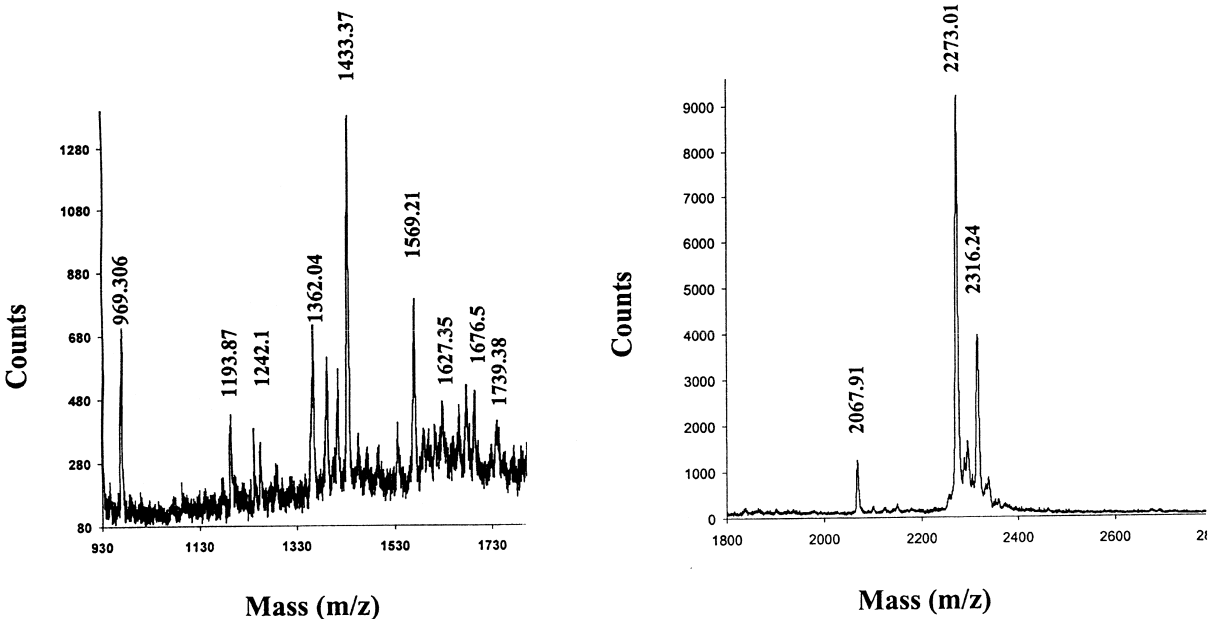

Fig. 7. Mass spectra of *E. coli* fractions taken from the RPC column.

sufficient complexity to show quenching, but it must be considered. It is also possible that peak intensity was too low to be seen, either because of poor ionization efficiency or limited concentration.

When internal mass standards are placed in a MALDI sample and the instrument is carefully calibrated, peptide mass may be determined to within a fraction of a mass unit. It is not as easy to include internal standard in samples and maintain the same degree of instrument calibration while screening thousands of samples. Mass accuracy in this case falls to ±several amu. Using a mass accuracy of several amu, it is frequently the case that several hundred tryptic peptides in the database may be of the same mass. The issue is to relate peptides in the MALDI spectrum to a sequence in the database. It will be shown below, that carboxypeptidase sequencing is one way to get sequence data. There is also the possibility that the parent protein may not be in the database, as in the case of organisms where the sequence of the genome is incomplete or unknown. A molecular biological approach would be necessary in this case. Sequence from the peptide would be used to synthesize a DNA probe complimentary to the cDNA of the mRNA coding for the parent protein. Subsequent to polymerase chain reaction (PCR) amplification, the cDNA would be sequenced to obtain the protein sequence.

3.2.2. Selecting post-translationally modified peptides

Post-translational modification plays an important role in regulation. For this reason, it is necessary to have methods that detect specific post-translational modifications. Among the more important are (i) *O*-glycosylation, (ii) *N*-glycosylation, and (iii) the phosphorylation of tyrosine, serine or threonine.

Although common in the cytosol, *O*-glycosylation in the nucleus evidently plays an important role in the control of transcription [19]. Glycosylation of serine and threonine with *N*-acetylglucosamine, in addition to deglycosylation, in the nucleus appears to be important in the synthesis and regulation of transcription factors. The biological significance of transcription factors, the fact that there are probably only a few thousand in the nucleus, and the ease with which they may be resolved from other nuclear and cytosolic proteins makes them attractive candidates for study.

Affinity selection of glycosylated peptides was

achieved with lectin columns. The lectin BS II from *Banderea simplifica* readily captures *O*-glycosylated proteins and peptides containing *N*-acetylglucosamine (glc-Nac). The procedure is essentially identical to the other affinity selection methods described above. Following reduction and alkylation, proteins were tryptic digested, the glycopeptides selected on the BS-II affinity column, and the glycopeptides then resolved by RPC (Fig. 8). The reversed-phase chromatogram in this case is much simpler than that of histidine peptides selected from *E. coli*. This is due to the fact that the number of *O*-glycosylated proteins in mammalian nuclei from which these peptides were derived is obviously much smaller than the entire set of proteins in *E. coli*. But still, there is a strong possibility that multiple peptides will coelute from the reversed-phase column.

MALDI mass spectra of a reversed-phase fraction indicate that peptides do coelute (Fig. 9). Because *O*-glycosylation with *N*-acetylglucosamine involves a single carbohydrate residue, there is no mass

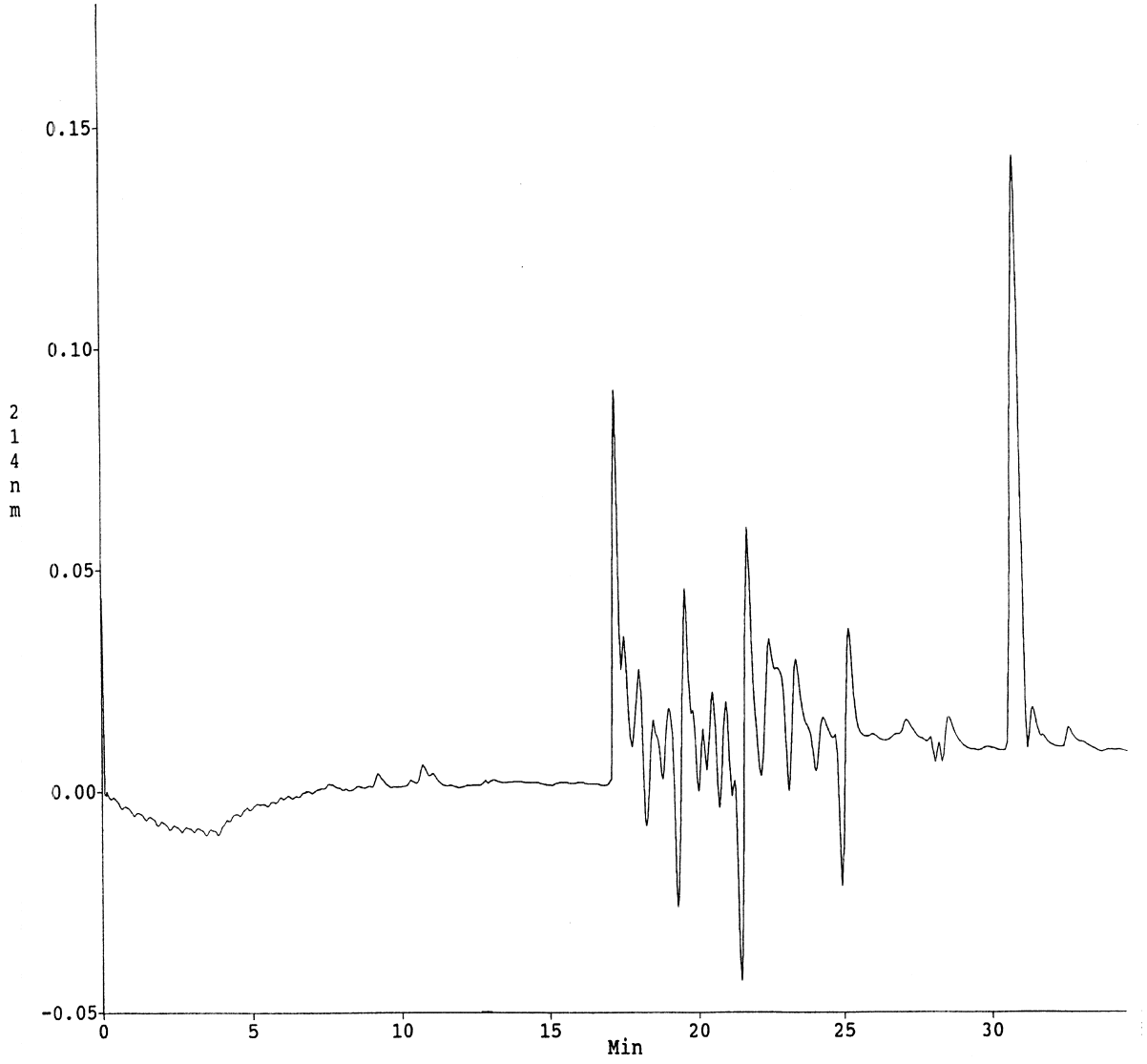

Fig. 8. The reversed-phase chromatogram of *O*-glycopeptides selected from a tryptic digest of nuclear extracts by the BS II affinity column.

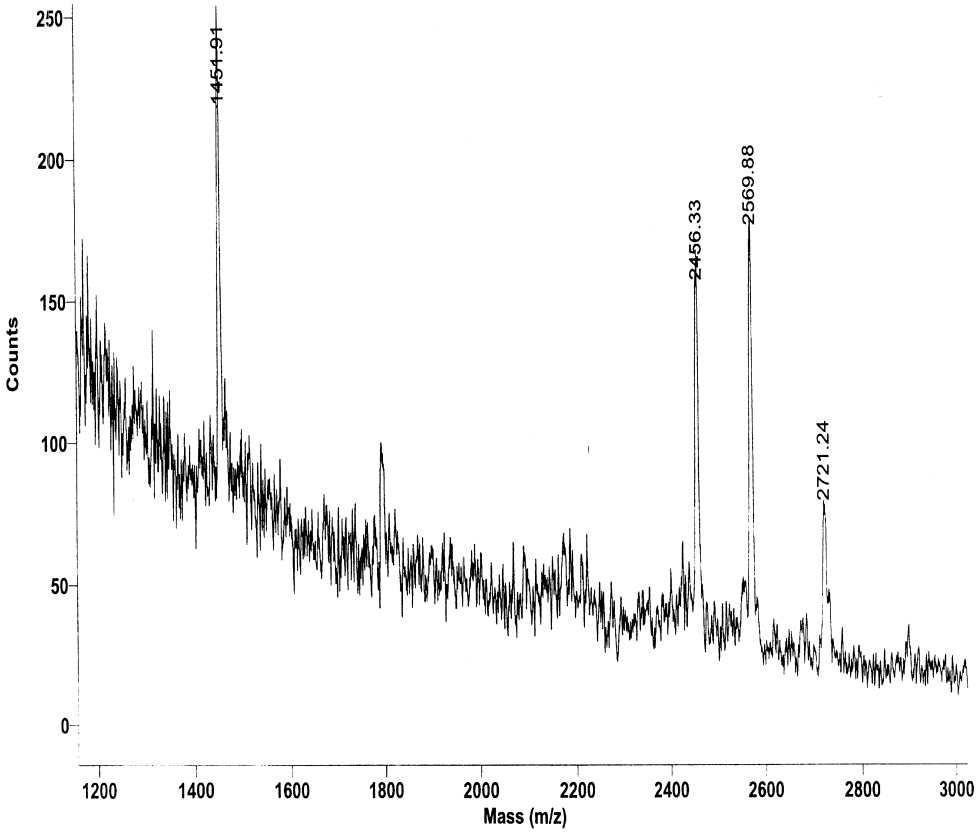

Fig. 9. MALDI mass spectra of a reversed-phase fraction of *O*-glycopeptides.

heterogeneity and each peak in the MALDI mass spectrum is from a different peptide. The mass of the carbohydrate residue is subtracted from the mass of the peptide before database searches. That will not be the case with the more complex O-linked glycopeptides from the cytosol where multiple oligosaccharides are attached at a single site. Enzymatic deglycosylation of peptides or cleavage of carbohydrate residues in strong base subsequent to affinity selection is necessary in these cases. *N*-Linked glycopeptides are not deglycosylated under these conditions.

Again there is the issue of deriving sequence data from peptides seen in the mass spectrum. It has recently been shown that carboxypeptidase may be utilized to partially sequence peptides using MALDI-MS to analyze the cleavage products [20]. Mixtures of carboxypeptidases sequentially cleave amino acids from the C-terminus of peptides. Because the rate at which various amino acids are cleaved varies widely, the cleavage process rapidly gets out of synchrony and a peptide ladder is generated. Differences in mass between the peptides in this ladder reveal the individual amino acids in the sequence of the parent peptide. Although this procedure has been used with pure peptides, it was reasoned that it might also work with simple mixtures. That it does is seen in Table 1. Five peptides in a mixture were sequenced simultaneously using a commercial carboxypeptidase sequencing kit. Using the same procedure, one of the peptides from the nuclear extract was sequenced and found to contain the sequence AGI (Table 2). Using the mass of the peptide and this sequence a database search revealed that this peptide was probably derived from the protein tyrosine phosphatase.

N-Linked oligosaccharide-containing peptides are

Table 1
Partial sequence of mixture of peptides by carboxypeptidase

Mass (m/z) of parent ion	Mass (m/z) of daughter ion	Partial sequence
5735.92	5024.01	
	4875.65	F
3659.83	3532.24	AEAFPLE
	3419.77	
	3324.07	
	3175.18	
	3101.76	
	2971.23	
	2900.49	
2466.41	2321.35	F
	2192.17	
2094.49		
1298.64	1185.17	PFHL
	1046.98	
	899.76	
	803.81	

more common. The lectin, concanavalin A, was used in this case to affinity select glycopeptides following reductive alkylation and proteolysis. Glycopeptides affinity selected from tryptic digests of both pure transferrin and human serum were examined. Heterogeneity of glycosylation in the pure transferrin sample peptides and transferrin peptides derived from serum was observed (Fig. 10). Although this heterogeneity complicates interpretation of the spectra of unknowns, it can be as an asset in identifying known peptides in serum. The presence of a series of peaks from a peptide can be used as a ''finger print''. Unfortunately, spreading the mass of a peptide across several species also lowers detection sensitivity. Interpretation of data from unknown peptides would be simpler if the peptide were deglycosylated [21].

4. Conclusions

Based on the data presented above several conclusions may be reached. One is that post-biosynthetic derivatization of natural peptides can give isotopically labeled internal standards that provide a basis for peptide quantification by mass spectrometry. However, the efficacy of this technique in monitoring regulatory flux remains to be determined. A second is that affinity selection can greatly simplify tryptic digests. This simplification is sufficient that two-dimensional liquid chromatography followed by MALDI-MS and database searches can offer a new route to identification of proteins in complex mixtures. Finally, sequence data derived from carboxypeptidase digestion can further enhance peptide identification in sequence databases.

Table 2
Partial sequence of a glycopeptide from nuclear extract by carboxypeptidase

Mass (m/z) of parent ion	Mass (m/z) of daughter ion	Partial sequence
2569.88	2092.65	AGL/N/I
	2163.38	
	2222.24	
	2335.89	

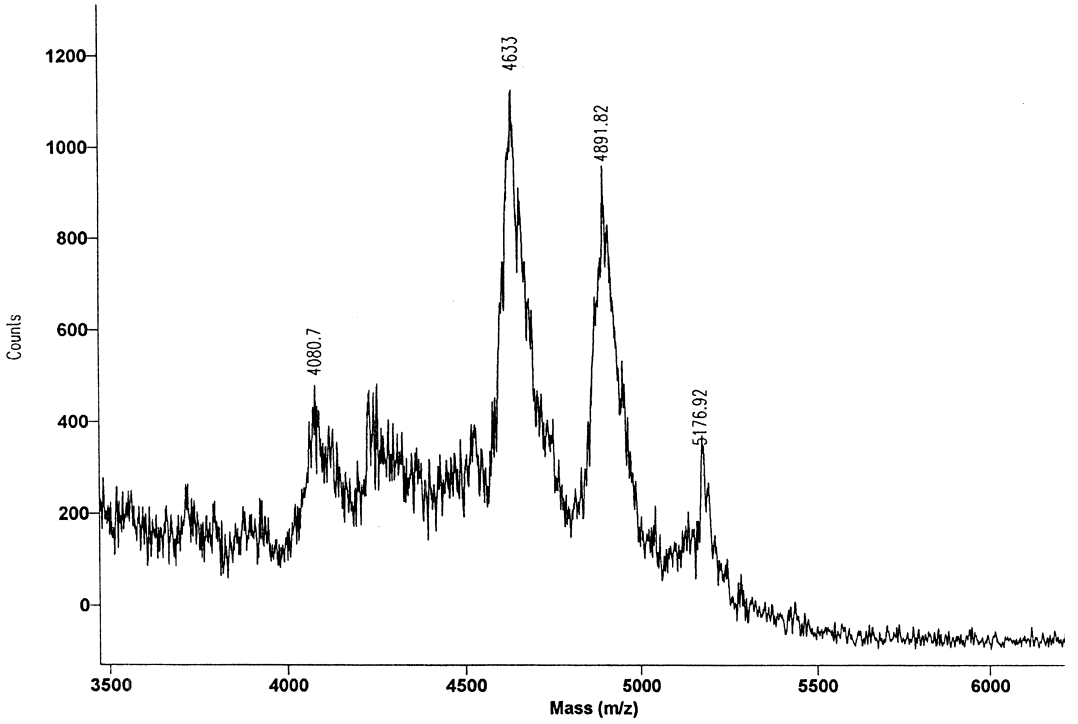

Fig. 10. Heterogeneity of glycosylation in a peptide selected from a trypsin digest of transferrin by concanavalin A affinity and reversed-phase chromatography.

Acknowledgements

The authors gratefully acknowledge support from the National Institute of Health (GM 59996) and PE Biosystems. The authors also wish to thank Dr. Tim Nadler of PE Biosystems for his assistance in obtaining the ESI-MS data presented in this paper.

References

[1] M.D. Adams, J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, J.C. Venter, Science 252 (1991) 1651–1656.
[2] M.R. Wilkins, Biotechnology 14 (1996) 61–65.
[3] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Proc. Natl. Acad. Sci. USA 93 (1996) 10614–10619.
[4] M. Schena, D. Shalon, R. Heller, A. Chai, P.O. Brown, R.W. Davis, Science 270 (1995) 467–470.
[5] L. Anderson, J. Seilhamer, Electrophoresis 18 (1997) 533–537.
[6] E. Celis, P. Gromov, Electrophoresis 10 (1999) 16–21.
[7] P. Dainese, W. Staudenmann, M. Quadroni, C. Korostensky, G. Gonnet, M. Kertesz, P. James, Electrophoresis 18 (1997) 432–442.
[8] P.H. O'Farrell, J. Biol. Chem. 250 (1975) 4007–4021.
[9] J.R. Yates, J. Mass Spectrom. 33 (1998) 1–19.
[10] K. Doyle, Y. Zhang, R. Baer, M. Bina, J. Biol. Chem. 269 (1994) 12099–12105.
[11] T. Nadler, C. Blackburn, J. Mark, N. Gordon, F.E. Regnier, G. Vella, J. Chromatogr. 743 (1996) 91–98.
[12] L.A. Holt, S.J. Leach, B. Milligan, Australian J. Chem. 21 (1968) 2115–2117.
[13] M. Geng, J. Ji, F. Regnier, J. Chromatogr. (1999) in press.
[14] J.R. Yates, J. Mass Spectrom. 33 (1998) 6–13.
[15] Y. Nakagawa, T.T. Yip, M. Belew, J. Porath, Anal. Biochem. 75 (1981) 168.
[16] P. Hensen, G. Lindeberg, L. Andersson, J. Chromatogr. 627 (1992) 125–135.
[17] G. Purnananda, BioEssays 17 (1995) 987–997.
[18] W.R. Wilkinson, A.I. Gusev, A. Proctor, M. Houalla, D.M. Hercules, Fresenius J. Anal. Chem. 357 (1997) 241–248.
[19] G. Hart, Annu. Rev. Biochem. 66 (1997) 315–335.
[20] D.M. Patterson, G.E. Tarr, F.E. Regnier, S.A. Martin, Anal. Chem. 67 (1995) 3971–3978.
[21] M. Geng, F. Regnier, manuscript in preparation.