ELSEVIER

Journal of Chromatography B, 782 (2002) 363–383

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Artifacts and unassigned masses encountered in peptide mass mapping

Jonathan A. Karty[a], Marcia M.E. Ireland[b], Yves V. Brun[b], James P. Reilly[a,*]

[a]*Department of Chemistry, Indiana University, Bloomington, IN 47405, USA*
[b]*Department of Biology, Indiana University, Bloomington, IN 47405, USA*

Abstract

In peptide mass mapping of isolated proteins, a significant number of the observed mass spectral peaks are often uninterpreted. These peaks derive from a number of sources: errors in the genome that give rise to incorrect peptide mass predictions, undocumented post-translational modifications, sample handling-induced modifications, contaminants in the sample, non-standard protein cleavage sites, and non-protein components of the sample. In a study of the stalk organelle of *Caulobacter crescentus*, roughly one-third (782/2215) of all observed masses could not be assigned to the proteins identified in the gel spots (Karty et al., J. Proteome Res., 1 (2002) 325). By interpreting these masses, this work illuminates a number of phenomena that may arise in the course of peptide mass mapping of electrophoretically separated proteins and presents results from a number of related studies.

Keywords: Peptide mass mapping; Proteomics

1. Introduction

Since its description in 1993, protein identification by peptide mass mapping (PMM) has become the cornerstone of many proteomic studies [1–6]. In these experiments, proteins are digested with a proteolytic enzyme, and mass spectra of the resulting peptides are recorded. The experimentally observed masses are compared to a list of theoretical proteolytic fragment masses generated from the amino acid sequences of proteins that may be present in the sample. This list can be as short as the peptides arising from a single protein or as long as the total number of proteolytic peptides from every protein in a large sequence database such as SwissProt or NCBI. Most PMM algorithms rank protein identifications by the number of matching experimental and predicted masses [7]. High-resolution, high-accuracy MALDI-TOF mass spectrometers have become standard tools in proteomics research [7–10]. Complex samples containing multiple proteins often give rise to a large number of proteolytic fragments. Since mass spectra contain a limited number of peaks, all components of a complex mixture may not be represented in a MALDI-TOF mass spectrum. To ensure that all components are adequately represented, a sample containing relatively few proteins is required for successful MALDI-TOF peptide mass mapping. For this reason, two-dimensional gel electrophoresis commonly precedes peptide mass mapping of cellular lysates [7,8,11]. The principal advantage of

*Corresponding author. 800 E. Kirkwood Ave., Bloomington, IN 47405, USA.
E-mail address: reilly@indiana.edu (J.P. Reilly).

1570-0232/02/$ – see front matter
PII: S1570-0232(02)00550-0

MALDI-TOF MS for analyzing electrophoretically separated proteins is its speed. Spectra can be acquired in seconds, and increasing computer power allows for database searches to be performed in near real-time allowing for highly automated data acquisition and analysis [8,11–13].

Although experiments matching observed and calculated peptide masses have been underway for nearly 20 years [14], the process of predicting proteolytic fragment masses has been greatly aided by the publication of the complete genomes of many organisms. For a sequenced organism, it should be possible to predict the masses of all proteolytic fragments. Furthermore, if all features in a mass spectrum arise from proteolytic fragments of that organism, they should all be interpretable. This is not always the case. In previous work, our lab investigated 2D-gel-separated proteins from the stalk organelle of *Caulobacter crescentus* bacteria. Proteins were identified in 50 of the 62 gel spots analyzed. In those cases, 782 of the 2215 mass spectral peaks were not initially interpretable [1]. Motivated by the desire to improve the confidence of our identifications, we performed an exhaustive study of the origins of these uninterpretable masses.

Effective peptide mass mapping is based on several assumptions. The first is that the published genome is accurate. Frame-shifts can arise from incorrect sequencing of the genome, specifically the inclusion of an extra base or the deletion of a base in a gene's sequence [14–16]. Different strains of the same organism can have variations in their genomes [17]. Errors can also arise during genome annotation. For example, we previously identified several incorrectly assigned protein start codons [18]. Genome errors are transferred to the proteome upon in silico translation. Proteome errors give rise to predicted peptide masses that will be improperly interpreted. For example, a frame-shift in a gene may cause all codons downstream to be incorrectly translated [15]. Likewise, an incorrectly documented start codon could cause a protein sequence to be too short (if the true start codon is upstream of the published one) or too long (if the true start codon is downstream of one annotated). Both of these errors lead to improper prediction of the mass of the N-terminal protelytic fragment.

Even with an accurate genome, any of several protein modifications can introduce complications into peptide mass mapping. A common post-translational modification is the cleavage of signal peptides [19]. Many organisms use portions of protein N-terminal sequences as markers for export by any of a number of secretory pathways [20]. Often, the locations of the signal peptide cleavage points are not annotated. This is especially true for hypothetical proteins that have not been fully characterized. Signal peptide cleavage radically alters the mass of the N-terminal proteolytic fragment of a protein. One effect of signal peptide cleavage is that the originally predicted N-terminal peptide will not be observed. Likewise, its new mass will not be predicted and therefore, if observed in a mass spectrum, it will not be properly assigned. Signal peptide cleavage also affects a protein's predicted molecular mass and isolectric point (p*I*) as discussed below [21–23].

Protein or peptide modifications at single amino acids give rise to masses that may not be predictable from the genome. Although scores of protein modifications have been described [24,25], common ones include: phosphorylation [26], methylation [27], glycosylation [28], acetylation [29], oxidation [30–32], nitration [33], acrylamidation [34,35], vinylpyridinylation [36], guanidination [37–41], esterification [42], carbamidomethylation [43,44], formylation [35], carbamylation [24,29], deamidation [45–48], and formation of disulfide bonds [49]. These modifications can occur naturally in an organism, result from deliberate chemical derivatization, or inadvertently arise during sample handling. They affect peptide mass mapping in different ways: first, modified peptide masses will not be correctly matched. These ''false negatives'' decrease the confidence levels of protein identifications. Second, the masses of modified peptides will often match predicted fragment masses of proteins not present in the sample. These so-called ''random matches'' represent false positives that may suggest an incorrect or ambiguous identification. Similarly, other sources of false positives are impurities such as human keratins [50,51] or autolysis products of proteolytic enzymes [52] that also give rise to mass spectral features.

Another common assumption in peptide mass mapping experiments is that proteolytic specificity is perfect. Many commonly used proteolytic enzymes and chemical digestion protocols cleave proteins at

specific amino acid residues (e.g., trypsin at lysine and arginine). Some of these procedures can have side reactions, however. Trypsin often contains chymotrypsin as an impurity. Chymotrypsin primarily cleaves C-terminal to the aromatic residues phenylalanine, tyrosine, and tryptophan. Furthermore, trypsin can cleave itself giving rise to pseudotrypsin [53,54], which displays chymotryptic activity. Additionally, proteins can be cleaved by extremes of pH without addition of any other reagent; Asp–Pro linkages are extremely labile at acidic pH [55]. Non-specific or unexpected cleavages can give rise to peptides whose masses will not be predicted from the proteome.

Finally, a common assumption in peptide mass mapping is that all mass spectral features arise from peptides. In fact, low-mass matrix–alkali clusters are commonly observed in MALDI-TOF mass spectrometry [56,57]. These are not related to proteolytic fragments of proteins, and can therefore cause some confusion during data analysis. Although matrix–alkali cluster masses do not often match the masses of predicted tryptic fragments [57], they still add to the number of non-matched masses, and therefore lower the confidence of assignments.

The previously mentioned phenomena need to be considered during the construction of peptide mass maps. Many artifacts such as autolysis peptides [52,57], keratin peptides [50,51,58], and matrix–alkali cluster ions [56,57] can be predicted and removed from mass lists prior to analysis. Similarly, many of the above modifications, if deliberate or well annotated in the proteome, can be taken into account when generating the list of theoretical peptide masses. Some protein modifications are not predictable from primary sequences. Deamidation, for example, is highly dependent on both pH and the amino acids near the Asn or Gln residue in three-dimensional space [45,46,48]. Double oxidation of tryptophan can occur when peptides interact with certain types of surfaces [30] or with ozone in the laboratory air [32].

Most peptide mass mapping algorithms allow observations of protein molecular mass and p*I* to be used as parameters to restrict searches. Many of the modifications described above can impact a protein's molecular mass and p*I*. Signal peptides in Gram-negative prokaryotes tend to be 20–30 amino acids in length and contain a number of positively charged residues [20]. Cleavage of these peptides will have two major effects. First, the protein's molecular mass is reduced. Second, the loss of two or three positively charged residues can dramatically affect a protein's predicted p*I* [21,23]. Many single amino acid modifications can affect p*I* as well. Most naturally occurring modifications, such as methylation, phosphorylation, and deamidation reduce protein p*I*, although a handful, such as amidation of C-termini, can increase p*I*. Similarly, many deliberate modifications introduced during sample handling can change a protein's p*I*. Again, many of these such as alkylation with iodoacetic acid, deamidation, and carbamylation decrease protein p*I* values, although others, for example guanidination, can increase p*I*. The effects of a large number of protein modifications on p*I* predictions have been reviewed [22].

Guanidination, the conversion of lysine to homoargine by *O*-methylisourea, has been shown to increase the intensity of lysine-terminated tryptic peptides [37–41]. Furthermore, by comparing unguanidinated and guanidinated mass spectra, limited sequence information about a peptide, namely its lysine content, can be inferred [1,59]. The advantage of using the lysine content as a search parameter during peptide mass mapping for the identification of proteins in a large number of gel spots and the ability to assign absolute confidence limits to the identifications have recently been described [1]. The lysine count can also be used to discriminate between the various explanations for any unmatched mass discussed herein.

2. Experimental

The procedure for the isolation and electrophoretic separation of stalk proteins has been recently described [18]. Briefly, stalks were isolated from a stalk-shedding mutant strain of *Caulobacter crescentus* by differential centrifugation. They were lysed using multiple freeze–thaw cycles. Proteins were solubilized with urea and separated by isoelectric focusing using an IPGPhor apparatus (Amersham–Pharmacia Biotech, Piscataway, NJ). The IPG (immobilized pH gradient) strips (3-10NL, APBiotech) containing the focused proteins were soaked in a

buffer containing urea, sodium dodecyl sulfate (SDS), and dithiothreitol to reduce any disulfide bonds. They were then soaked in a similar buffer containing iodoacetamide to alkylate the newly generated free cysteines. The IPG strips were placed on top of a 2–3-mm layer of 0.5% agarose which in turn formed the stacking layer of 12% SDS–polyacrylamide gel. The second dimension gels were run for 4 h, and the proteins visualized with colloidal Coomassie Blue (Novex, San Diego CA).

Spots containing separated proteins were destained and digested with trypsin using a previously described procedure [1,60]. Briefly, spots were destained with multiple treatments of an acetonitrile, ammonium bicarbonate, water solution. They were then washed with water, and dehydrated with acetonitrile. A vacuum centrifuge (Jouan, Winchester, VA) was used to dry the spots completely. The dried gel spots were rehydrated with 15 μl of 16.67 mg/l (250 ng total) bovine trypsin (Sigma T8802, St. Louis, MO, USA) and incubated overnight at 37 °C. Digestion was stopped by the addition of 0.1% (v/v) trifluoroacetic acid (TFA). Tryptic peptides were extracted using successive 100-μl washes of 0.1% (v/v) TFA, 30% (v/v) acetonitrile, and 60% (v/v) acetonitrile. The three extracts from each spot were combined, vacuum centrifuged to dryness, and resuspended in 8 μl of Type I water (E-Pure, Barstead, Dubuque, IA) Half of the suspension was guanidinated following the procedure described by Beardsley and Reilly [41].

MALDI spots of unguanidinated samples were made by depositing 0.65 μl of a solution containing equal parts peptide extract and 15 g/l α-cyano-4-hydroxycinnamic acid (CHCA) in 75% (v/v) acetonitrile, 0.1% (v/v) TFA onto the stainless steel target. Guanidinated samples were purified using micro-extraction columns packed with Zorbax C_{18} medium (Sigma). The columns were eluted with 2 μl of 10 g/l CHCA in 50% (v/v) acetonitrile, 0.1% (v/v) TFA; 0.65 μl of the eluent was deposited onto the MALDI probe. All spots were allowed to air-dry prior to loading into the mass spectrometer. Positive ion mass spectra were recorded using a Bruker Reflex III reflectron MALDI-TOF mass spectrometer. Spectra were internally calibrated using three or four tryptic autolysis peaks.

Peak reports from both guanidinated and unguanidinated data were analyzed with a previously described, in-house algorithm, *Prodigies* [1]. The program deduces the number of lysines in a particular peptide by comparing experimental unguanidinated and guanidinated data. Guanidination quantitatively and selectively converts lysines to homoarginines, inducing a 42-u increase in mass for each lysine residue in a peptide [41]. *Prodigies* creates three output files, or ''master hit arrays'', (MHAs) for each pair of spectra. The first summarizes a peptide mass mapping database search comparing the observed unguanidinated masses to a list of all possible tryptic fragment masses for the entire *Caulobacter crescentus* proteome [61]. Assumptions implicit in the search include partial oxidation of all methionine residues (+16 u), partial cleavage of protein N-terminal methionine residues (−131 u), total alkylation of cysteines by iodoacetamide (+57 u), and partial conversion of peptide N-terminal glutamine residues to pyroglutamic acid (−17 u). The second MHA is qualitatively similar to the first. It summarizes the results of a database search in which all lysines have been converted to homoarginines. Often, only a small collection of open reading frames (ORFs) is common to both the unguanidinated and guanidinated MHAs. The lysine count mentioned above is not obtained until the two MHAs are compared. The true power of *Prodigies* is evident in the third master hit array. The program compares both the unguanidinated and guanidinated peak reports and looks for features that are shifted by integral multiples of 42 u. These pairs of peaks separated by multiples of 42 u will be referred to hereafter as ''consistent peaks''. The multiple of 42 u by which a peak shifts upon guanidination corresponds to the number of lysine residues in the peptide. This ''lysine count'' is a parameter in the database search used to prepare the third MHA. A match in this MHA is indicated only when a theoretical peptide's mass matches an observed mass and it contains the previously determined number of lysines. The product of the ORF with the most matches in an MHA is assumed to be present in the gel spot. Statistical methods were used to assess the confidence with which identifications were made. Absolute confidence limits for assignments were derived through Monte Carlo simulations of over 5.5 million peptide mass mapping experiments [1].

Despite the fact that all but two of the *Caulobacter* stalk protein assignments were made with greater than 95% confidence, roughly one-third of the observed masses were initially unassigned [1]. In order to interpret these masses, all data files were analyzed a second time using a slightly different database-searching algorithm. This time, guanidinated and unguanidinated data were considered individually. *Prodigies* attempts to interpret unmatched masses in three ways. First, starting with all theoretical peptides associated with each ORF listed in an MHA, it considers nine different types of peptide modifications. These include: oxidation (+16 or 32 u) [7,30–32,62], carbamylation (+43, 86, or 129 u) [24], phosphorylation (+80 u) [26], methylation (+14 u) [27], acetylation (+42 u) [29], deamidation (+1 u) [45–48,63], formylation (+28 u) [35], acrylamidation (+78 u) [34,35], and carbamidomethylation (+57 u) [43,44]. The mass of each modified theoretical peptide is shifted appropriately, and a new list of theoretical masses is generated. This list is compared to the unmatched experimentally observed masses. All reported potential peptide modifications are then verified to ensure that the modifications and the peptide sequences are compatible (e.g., a predicted deamidation needs to correspond to a peptide with an N or a Q in its sequence). Second, the algorithm checks for ''random cleavage peptides''. Random cleavage peptides are generated by protein hydrolysis at any amino acid, not just lysine and arginine. In this mode, any contiguous sequence of amino acids with a predicted mass equal to an unmatched mass is noted. Third, the algorithm attempts to identify signal peptides by removing amino acids from the N-terminus of each ORF listed in an MHA. If any new predicted N-terminal tryptic fragment mass matches one of the unmatched experimentally observed masses, it is recorded. Putative signal peptide cleavage sites are then checked against the predictions made by the *SignalP* 2.0 program (www.cbs.dtu.dk/services/SignalP-2.0) [19].

Localization predictions for all proteins identified were made using the *PSORT* algorithm (http://psort.nibb.ac.jp) [64]. This algorithm analyzes protein sequences and suggests their preferred subcellular environments. PSORT did not predict localizations for several proteins identified in the *Caulobacter* stalk gels. These were further examined using the *CODONPREFERENCE* program (GCG, Madison, WI) that analyzes ''GC bias,'' the occurrence of G or C in the third position of DNA triplet codons, and tracks the frequency of so-called ''rare codons''. Many organisms preferentially use specific codons to encode amino acids despite degeneracies in the codon translation table [65]. Codons other than those preferred are called rare codons. A properly annotated gene should have a GC bias consistent with that of the rest of the ORFs and should have a low occurrence of rare codons. These analyses allow one to examine the annotation of a particular gene. In this work, they were used to assess the start codon predictions for several proteins.

MALDI-QTOF CID MS^2 mass spectra for some peptides were recorded using a QTOF Global instrument (Micromass, Manchester, UK). The argon collision gas inlet pressure was set to 25 p.s.i. and the collision voltage setting varied with the parent mass of the peptide fragmented. All MS^2 spectra were externally calibrated, and a 0.1-u mass error window was used for all fragment peak assignments.

3. Results

Proteins were identified in 50 out of 62 gel spots analyzed. A complete discussion of the proteins identified, confidence levels of assignments, and the biological significance of the assignments are presented in two other publications [1,18]. There were 2215 mass spectral peaks observed in the 100 mass spectra (50 unguanidinated and 50 guanidinated samples) recorded for the 50 gel spot digests in which proteins were identified. Nearly two-thirds of those masses, 1433, could be assigned to proteins identified in the gel spots. Although little evidence was seen for multiple proteins migrating in the same spot, several proteins appeared in more than one spot [1,18]. The remaining 782 masses were not immediately interpretable, and these represent the focus of this paper.

The simplest explanation for unmatched masses in a gel spot digest would be the presence of a second protein in the spot [7,9,66]. However, the only time co-migrating proteins were observed was for two proteins that shared 82% sequence identity [1]. The

lack of co-migrating proteins is not surprising since a single bacterial organelle was studied.

3.1. Post-translational modifications: signal peptide cleavage

Signal peptides are cleaved after translation during translocation into the cytoplasmic membrane [20]. *SignalP 2.0* predicted signal peptide cleavages for the proteins in 40 of the 50 gel spots analyzed. Although this may seem like a large fraction, most of the proteins identified were associated with the outer membrane, and would be expected to be processed by the general secretory pathway [18,20]. Eleven previously uninterpreted masses corresponding to N-terminal tryptic fragments created by cleavage of signal peptides were observed in these experiments, including three pair of ''consistent peaks''. The sequence of one of the new N-terminal fragments was confirmed with MS^2 sequencing (data not shown) [67].

3.2. Sample handling induced modifications: deamidation of asparagine

Deamidation of asparagines and glutamines to aspartates and glutamates, respectively, is an effect of protein aging, and is often observed in 2D gel electrophoresis [22,47,68]. It leads to a series of spots with the same apparent mass, but measurably different isoelectric points. Deamidation can also arise when samples are kept under basic conditions [45,46,48,69]. Deamidation causes a peptide to increase in mass by 1 u and the apparent p*I* to decrease [22,70]. Close examination of the 782 unmatched masses revealed 38 that could be explained by a single deamidation, and two more that involved the deamidations of two asparagines in the same peptide. Interestingly, 34 of the deamidations involved peptides with NG in their sequences. Asparagines followed by glycine are known to be the most rapidly deamidated [46,48,70]. Furthermore, deamidation of NG-containing peptides was strongly guanidination-dependent. While 32 NG-containing peptides were observed in unguanidinated mass spectra, only three of these were deamidated. In contrast, of 34 NG-containing peptides identified in the guanidinated mass spectra, 31 were deamidated. This striking result is consistent with previously published studies demonstrating strong pH and temperature dependences for the rate of deamidation [46,48,70]. Deamidation under basic conditions has been shown to proceed through a succinimidyl intermediate created when the backbone amide nitrogen attacks the carbonyl of the asparagine side chain [48,63,70]. The pH of solution during guanidination is approximately 10.6, and the reaction is performed at 65 °C [41]. These conditions should favor deamidation of NG-containing tryptic peptides. An important ramification of deamidation on peptide mass mapping is that all 31 cases of guanidination-induced deamidation resulted in the non-identification of consistent peaks.

A peak at m/z 2045.96 was observed in the unguanidinated mass spectrum of a gel spot containing ORF CC163, a hypothetical protein. The mass corresponded to that of the predicted tryptic fragment YADNGSEDAEEFTVTANGR. Upon guanidination, a peak at m/z 2047.84 was observed. This new peak could be explained as the same tryptic fragment with both NGs converted to DGs. Fig. 1B displays a CID mass spectrum of the m/z 2045.96 peak in the unguanidinated sample with the y ions labeled. Fig. 1A displays the CID mass spectrum of the peak observed at m/z 2047.84 upon guanidination. The y_{16} ion in the top spectrum is 2 u heavier than the y_{16} ion in the bottom spectrum, and the y_3–y_{15} ions in the top spectrum are 1 u heavier than the y_3–y_{15} ions in the bottom spectrum. These two fragment ion series are consistent with the sequences indicated in Fig. 1 and clearly implicate deamidated Ns 3 and 16 residues from the C-terminus.

3.3. Sample handling induced modifications: guanidination of peptide N-terminal glycine

Derivatization reactions often cause unintended side reactions. Almost all peptides have primary amines at their N-termini that can be possible substrates for guanidination. Beardsley and co-workers twice reported some unexpected guanidinations on peptide N-terminal glycine residues [38,41]. Furthermore, Beardsley and Reilly demonstrated partial guanidination at the N-terminal glycine residue of a model tryptic peptide after 5 min reaction time, and limited guanidination at several other N-terminal residues after several hours [41]. In our

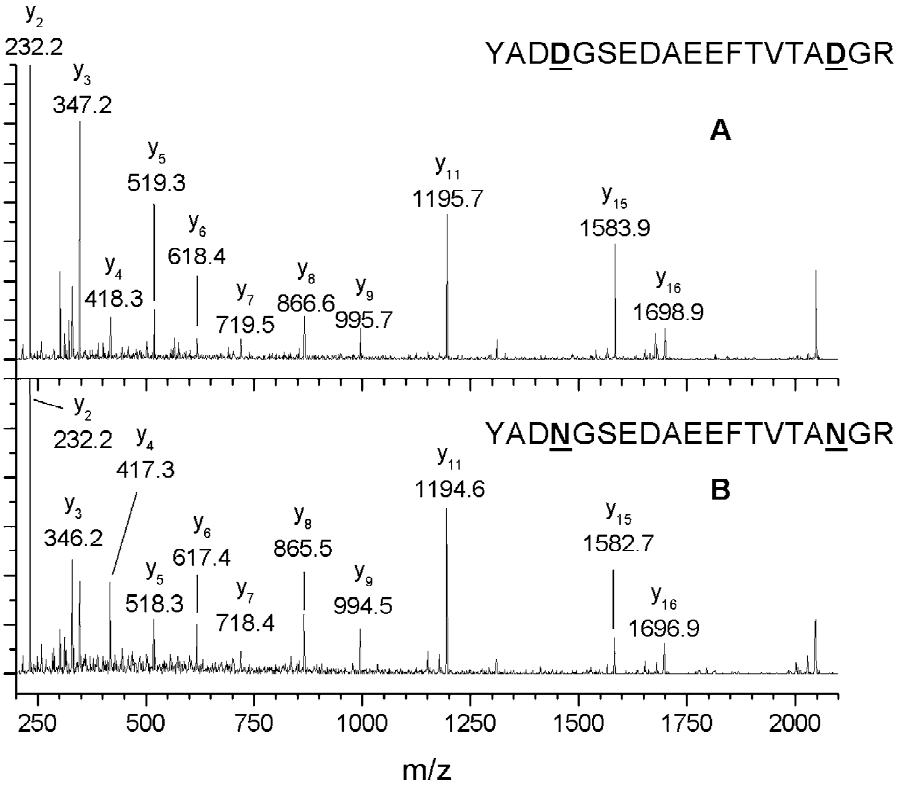


Fig. 1. MALDI-QTOF CID mass spectra of a tryptic peptide before (B) and after (A) guanidination-induced deamidation of its NG residues. The y ions and their masses are labeled in each spectrum.

experiments, the reaction occurred over 20 min, and only glycine was N-terminally guanidinated. Seventy-six peptides with N-terminal glycine residues were identified in the unguanidinated mass spectra. Fifty-five of these were observed in the guanidinated mass spectra. Of those, 14 showed partial guanidination at their N-termini (i.e., peaks for both the normally guanidinated and N-terminal guanidinated peptide were observed). Four others were totally guanidinated at their N-termini. All four cases of total guanidination at N-terminal glycine residues resulted in non-identification of ''consistent peaks'' since their masses shifted by 42 u more than expected. Total N-terminal guanidination only occurred 7% of the time, and is therefore seen as only a minor problem. The 14 partial N-terminal glycine guanidinations are actually useful. Since both the normally guanidinated and N-terminal guanidinated peaks were observed, one can infer the presence of an N-terminal glycine, thus learning even more about the peptide. The splitting of a single peak into two separate features does cost some sensitivity, but in general the sequence information gained outweighs the sensitivity loss [1].

3.4. Artifacts: matrix–alkali clusters

Keller and Li described the observation of matrix–alkali clusters between m/z 600–1500 in mass spectra of dilute peptide mixtures [56]. Similar observations were described by Harris et al. who noted the unusual fractional parts of their masses [57]. These ''fractional masses'' result from a superposition of the mass defects of all atoms in each molecule. In the present work, 65 matrix–alkali clusters were observed in unguanidinated mass spectra, and 17 were observed after guanidination. This reduction is most likely due to the removal of sodium and potassium during the micro-extraction step in the guanidination procedure; unguanidinated

samples were not purified. Although it may seem advantageous to purify both unguanidinated and guanidinated samples, loss of peptides at sub-picomole sample loadings has been reported when μC_{18} Ziptips were employed for this purpose [71]. In the guanidinated case, these losses are compensated by the increased ionization efficiency of derivatized peptides and the information derived from the lysine count [1,41].

Mann commented that the fractional masses for all random peptides with nominal masses below 2000 u fell into very narrow (roughly 0.25 u) regions [72]. Gay et al. demonstrated similar results for all tryptic peptides calculable from release 37 of the SwissProt database [73]. Fig. 2 displays the fractional masses of all theoretical tryptic peptides for the *Caulobacter* proteome between 700 and 3500 u (predictions were made using the constraints described in Section 2). The most commonly observed matrix–alkali clusters were at m/z 855.07 (4 CHCA, 2 K, 1 Na), 871.05 (4 CHCA, 3 K), 877.08 (4 CHCA, 2 Na, 2 K) and 1060.09 (5 CHCA, 3 K). At m/z 870, for example, peptide ions have fractional masses of about 0.4 ± 0.1 u, in contrast with 0.07 for the M_4K_2Na. The fractional masses for all of the matrix alkali clusters observed in these experiments (denoted by squares in Fig. 2) were well outside the band of predicted peptide masses. The narrow distribution of fractional masses arises from the unique combinations of carbon, nitrogen, oxygen, hydrogen, and sulfur that make up the amino acids [57]. Fig. 2 can be used to discard non-peptide masses since the fractional masses of unmodified tryptic peptides of *Caulobacter*, or any other organism according to Gay et al. [73], should fall within a narrow band.

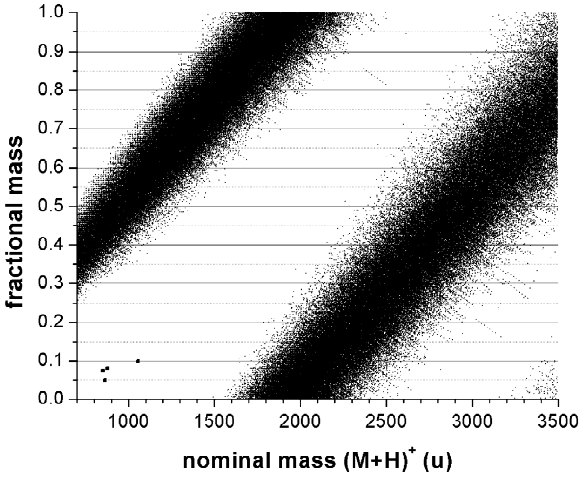


Fig. 2. Dependence of the fractional masses of predicted *Caulobacter* tryptic peptides on their nominal masses. The squares represent the locations of the matrix–alkali cluster masses.

3.5. Artifacts: uncommon autolysis peptides

Several tryptic autolysis peaks were used for calibration as described in Section 2. Four calibrants ($m/z = 805.416$, 1153.574, 2163.057, and 2289.155) were observed in every unguanidinated mass spectrum. Likewise, four guanidinated autolysis peptides ($m/z = 990.549$, 1195.596, 2205.079, and 2331.177) were seen in all guanidinated mass spectra. A handful of other known autolysis fragments were removed from peak reports prior to interpretation, including m/z 906.505, 1020.504, 1046.596, and 1493.600 in unguanidinated samples. Other masses that were not initially recognized as autolysis peaks were observed with sufficient frequency to warrant further investigation. A peak occurring at m/z 3211.467 and its methionine-oxidized counterpart at m/z 3227.492 corresponds to amino acids 140–170 of completely processed bovine trypsin with a disulfide bond between C148 and C162 and hydrolyzed between K149 and S150. They were observed six and nine times, respectively, as unmatched masses in the course of these experiments. Unfortunately, they were often too weak or not well enough resolved for routine use as high mass calibrants. The predicted masses for the $(M+H)^+$ ions of these peptides are 3211.475 and 3227.469 u. A peptide at m/z 1774.87 was observed in 13 different unguanidinated samples. It corresponds to the tryptic autolysis fragment from amino acids 1 to 14 and 137 to 139 with a disulfide bond between C7 and C137 ($(M+H)^+ =$ 1774.851 u). It should be noted that the N-terminus of this peptide results from a chymotryptic cleavage between Y14 and Q15. An ion at m/z 1433.78 appeared in eight different unguanidinated mass spectra, and it corresponds to amino acids 187–200 with C197 reduced; its predicted mass is 1433.721 u. The ions listed above had been reported previously [52]. Other, recently uncovered, autolysis peaks [57] were observed in these experiments. Two previous

reports indicate the presence of a tryptic autolysis fragment from amino acids 100–125 at m/z 2552 [52,57]. This interpretation requires that both cysteines in the peptide are reduced. An unmatched peak at m/z 2550.26 was observed in 16 different unguanidinated tryptic digest mass spectra, and an unassigned mass at m/z 2592 was observed in one guanidinated mass spectrum. The tryptic autolysis peptide, 100–125 contains one lysine and two cysteines. The species responsible for the m/z 2550.26 ion could be this tryptic autolysis peptide with a disulfide bond between C109 and C116 ($(M+H)^+ =$ 2550.233 u). This peak is often well resolved and it could be used as a calibrant when its identity is confirmed. Two other ubiquitous ions can be added to the list of tryptic autolysis peptides. An ion at m/z 1676.78 was observed in 14 different mass spectra (seven unguanidinated and seven guanidinated). The lack of mass shift implies that it corresponds to a peptide with no lysines. A tryptic autolysis peptide from amino acids 50 to 64 ($(M+H)^+ = 1676.777$ u) is consistent with these data. The C-terminus of this peptide arises from a chymotryptic cleavage between F49 and I50. Similarly, an ion at m/z 1756.88 was observed in eight different unguanidinated mass spectra. A tryptic autolysis peptide from P203 to K217 ($(M+H)^+ = 1756.909$ u) probably corresponds to this ion, but the cleavage between K202 and P203 is not favored by trypsin. These results are summarized in Table 1. Two other more complete listings of tryptic autolysis peaks have been published [52,57]. At present, sequencing experiments on these ions have been frustrated by their low intensities.

3.6. Artifacts: keratin peptides

Human keratin is a common contaminant in proteomics experiments. The keratin can either be present in the commercial trypsin preparation [51] or be introduced during sample handling [50,58]. 23 of the unmatched masses were consistent with previously described keratin peptides [50,58]. The most commonly observed keratin masses were m/z 832.4 in the unguanidinated data (10 times) and m/z 1151.5 (5 times) in the guanidinated data. Another keratin peak at m/z 1108.6 [50] was observed in a number of guanidinated mass spectra as well.

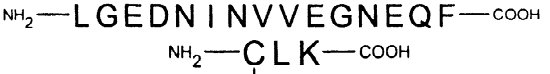
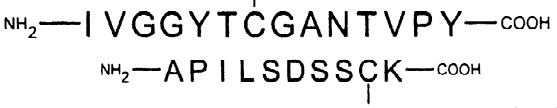
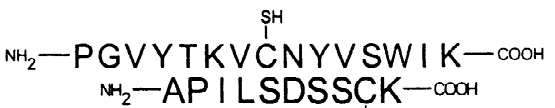
3.7. Post-translational modifications: double oxidation of tryptophan

Tryptophan can be converted to *N′*-formylkynurenine by photooxidation with near-UV light in an aqueous environment [31]. This oxidation increases the mass of a tryptophan-containing peptide by 32 u. Furthermore, the oxidation is slightly accelerated by mildly basic conditions. Gobom et al. reported double oxidation of tryptophan containing peptides when they analyzed samples using Anchor Chip® MALDI targets [30]. Cohen and Ward demonstrated that this can be induced by ozone in laboratory air [32]. Double oxidation of tryptophan residues was not considered during the initial analyses of our mass spectra, and several probable cases of it were observed upon closer examination of the data. In total, 112 tryptophan-containing peptides were observed in the mass spectra, of which 24 were probable cases of double oxidation. Eight of the 24 occurred as four pair of ‘‘consistent’’ peaks. In contrast to deamidation, tryptophan oxidation and guanidination were not correlated. Thirteen putative tryptophan oxidations were observed in the unguanidinated samples; 11 were seen in the guanidinated samples. There also appeared to be no correlation between tryptophan oxidation and sample storage time.

3.8. Artifacts: random cleavages

The assumption that proteins are cleaved at protease-specific sites is built into most PMM algorithms. Proteases present in the bacterial cell that have not been properly inhibited can lead to additional cleavages. Similarly, some of the conditions encountered during sample handling can promote hydrolysis of peptide bonds. In order to inhibit contaminating chymotrypsin, *N*-tosyl-L-phenylalanine chloromethylketone (TPCK), a potent irreversible inhibitor of chymotrypsin [74], is added to many sequencing grade trypsin preparations. TPCK is not stable at high pH, however, as chloromethyl ketones are rapidly hydrolyzed under basic conditions. For example, *N*-tosyl-L-lysine chloromethyl ketone loses 90% of its ability to inhibit trypsin after 5 min at pH 9.0 [75]. The pH of the ammonium bicarbonate solution used for digestion was approximately 8.5,

Table 1
Uncommon autolysis peptides[a] and ubiquitous ions

m/z[b]	Observations		Sequence
	Unguan	Guan	
Autolysis peptides			
2550.26	16	1	
1676.78	7	7	
1774.87	13	0	
3227.49[c]	9	0	
1433.78	8	0	
1756.88	8	0	
3211.47	6	0	
Ubiquitous non-autolysis ions			
2092.01	17	0	Unknown
1335.76	6	6	Unknown
2383.98	7	0	Unknown
2311.145	6	0	Unknown

[a] Trypsin sequence from SwissProt entry P00760 (www.expasy.ch) assuming cleavage of a 20 amino acid signal peptide.
[b] All masses are averages of all observations.
[c] ***M*** represents a singly oxidized methionine.

therefore, the TPCK should hydrolyze during digestion. It should be noted that pseudotrypsin displays low affinity for small polypeptide substrates [54], and therefore may not be inhibited by TPCK. Deamidation can cause bond cleavage C-terminal to asparagine residues [45,46,52], and, as noted above, deamidation was shown above to be accelerated during guanidination. In order to investigate the possibility of these processes, a list of all possible random cleavages that could explain unmatched masses was generated. Only one random cleavage per peptide was allowed; the other terminus had to be generated by a tryptic or chymotryptic cleavage or be the N or C-terminus of a protein.

Using the criteria described above, *Prodigies* found 510 possible random cleavage peptides that could explain 354 unique, unmatched masses. Evidently, some measured masses could be explained by more than one random cleavage peptide. To probe whether random cleavages correlate with specific amino acids, the residues at the N- and C-termini of all proposed random peptides were tabulated. Results

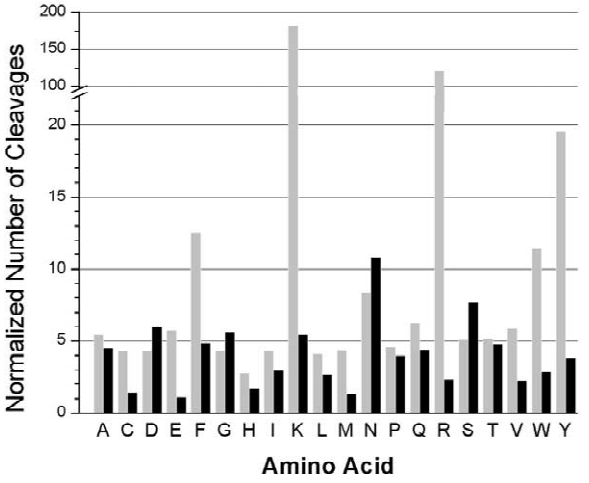


Fig. 3. Distribution of cleavage sites in the stalk gel spot proteomics study. All values have been generated by dividing the number of cleavages by the amino acid's abundance in the proteome (expressed in percent). Gray bars represent the frequency of cleavages C-terminal to a particular amino acid. Black bars represent the frequency of cleavages N-terminal to a particular amino acid. The bars for cleavages C-terminal to K and R correspond to normal tryptic fragments.

normalized with respect to the relative abundances of each amino acid in the proteome are summarized in Fig. 3. The gray bars for K and R at the C-termini of peptides represent normal tryptic cleavages for the 1433 tryptic peptides that were assigned to the proteins identified in these experiments. They are included to put the possible random cleavages in perspective (note the broken vertical axis in the graph). The bar corresponding to the normalized number of lysine cleavages is about 1.5 times as tall as that corresponding to arginine. Considering that the relative abundances of K and R residues have already been taken into account one might expect these bars to be of equal height. The observation of a higher efficiency for detecting lysine-terminated peptides is due to the significant enhancement of their intensities of upon guanidination [1,41].

Four bars in Fig. 3 stand out from the others. These bars correspond to F, Y, and W occurring at the C termini of the ''random'' peptides, and N appearing at the *N*-termini of the ''random'' peptides. The cleavages C-terminal to F, Y, and W represent chymotryptic activity, either due to contaminating chymotrypsin or pseudotrypsin formed by autolysis. The latter is more likely since the trypsin was treated with TPCK prior to packaging. Overall, 101 possible chymotryptic cleavages were identified. Of these, 30 are associated with peptides created by chymotryptic cleavage at both ends. It should be noted that two of the uncommon autolysis peptides mentioned in a previous section arose due to chymotryptic cleavages. Nine unmatched, consistent masses could be explained as peptides arising from a tryptic cleavage on one side, and a chymotryptic cleavage on the other. For example, a peak at m/z 2465.09 was observed in three different unguanidinated and three different guanidinated samples. It was seen as three consistent peaks in mass spectra of tryptic digests from gel spots containing ORF CC2149. Similarly, a feature at m/z 1800.97 was seen in three different unguanidinated and three guanidinated tryptic digests of ORF CC2149, and was observed as a consistent peak twice. The data suggest that neither of the peptides responsible for these ions contains lysine. Both can be explained as proteolytic peptides of ORF CC2149 created by one tryptic and one chymotryptic cleavage with predicted masses of 2465.086 and 1800.961 u.

The enhanced cleavage N-terminal to asparagine apparent in Fig. 3 was unexpected. Six of the random peptides that could be explained as cleavages N-terminal to asparagines were observed as three pairs of consistent peaks. We have no mechanistic proposals for cleavage N-terminal to asparagine at present. Neither the possible chymotryptic cleavages nor the cleavages N-terminal to asparagine were noticeably enhanced by guanidination.

3.9. Genome errors: improper start codon

Genome errors lead to the prediction of masses that will never be experimentally observed. In the course of the *Caulobacter* stalk analysis [1,18], it was determined that four-fifths of the proteins identified were associated with the outer membrane of the bacterium. Some of the proteins identified on the stalk 2D gel, like TonB-dependent receptors, were expected to localize to the outer membrane. However, analyses of the protein sequences associated with 20 identified gel spots using the *PSORT* program did not strongly predict subcellular localizations [18]. Furthermore, if these are outer membrane

proteins and are processed by the general secretory pathway, signal peptides should be predicted [20]. However, *SignalP* yielded no such predictions for many of these. A possible explanation for this dichotomy would be that the sequences of the N-terminal regions of these proteins were incorrect. In *Caulobacter*, more than two-thirds of all codons have G or C in the third position; a correctly annotated gene should reflect this GC bias [65,76]. This led us to examine the GC bias and codon preference near the proposed start site for each of the 20 proteins using *CODONPREFERENCE* (GCG, Madison, WI). We found that, in many of these cases, the GC bias increased a number of nucleotides downstream from the reported initiation codon, suggesting that the initiation site was incorrectly assigned. Similarly, the occurrence of rare codons in many of these cases did not drop until several residues downstream from the reported start codon. The increase in GC bias and decrease in rare codon usage coincided in several cases allowing for the prediction of new start codons. *SignalP* was then used to determine if the new initiation sites predicted by GC bias and codon preference would also result in the prediction of signal peptides. *PSORT* suggested the probable localizations of the proteins resulting from the new initiation sites. In eight of the 20 cases, the new initiation sites resulted in definitive predictions of signal peptides and subcellular localizations. In seven additional cases, the original initiation codons led to decisive predictions of signal peptides, but no localization predictions were suggested. A possible reading frame shift was identified in ORF CC3444 during the GC bias analysis. Furthermore, all 14 peptides matched to ORF 3444 were from the portion of the protein sequence C-terminal to the putative frame-shift. However, since no peptides upstream from of the frame-shift were observed, the exact location of the frame-shift could not be established.

3.10. Artifacts: unexplained yet commonly observed masses

A small number of unmatched masses were observed in several different gel spot digests yet they could not be explained. For example, an unmatched feature at m/z 2092.01 appeared in 17 guanidinated mass spectra. Its fractional mass suggests that it may be a peptide ion (see Fig. 2). Since it only appears in guanidinated mass spectra, it most likely contains at least one lysine. Another common, yet uninterpretable, mass observed only in unguanidinated data was at m/z 2311.15 (eight times). An unmatched peak at m/z 1335.75 was observed in six different unguanidinated spectra and six different guanidinated spectra. It was observed as a ''consistent'' peak in two different samples. An ion at m/z 2383.98 was observed in five different unguanidinated mass spectra and two different guanidinated mass spectra. The fractional masses of all the above listed ions are consistent with them being peptide in nature (see Fig. 2). They appear too ubiquitously to be peptides from co-migrating proteins, especially since little evidence of co-migration was observed [1]. It is much more likely that contaminants introduced during sample handling are responsible for these masses. As with autolysis and keratin peptides, they could be placed on an exclusion list in future experiments to reduce the number of false identifications. The masses and occurrences of all unexplained peaks observed six or more times in unrelated spectra are summarized in Table 1. Work is currently underway to sequence and identify them. Unfortunately, many of these ions were of low intensity, making MS^2 experiments difficult.

4. Discussion

Since peptides were not identified by sequencing, the credibility of our assignments was based primarily on recurrence of observation and consistency with previously published results. For example, the signal peptide analysis was rather unambiguous; signal peptide cleavage is well documented and two different algorithms led to compatible assignments. The deamidation of NG by guanidination, the guanidination of N-termini, and the observation of unexpected tryptic autolysis peptides were all examples of experimental observations that occurred often enough that their interpretation is compelling. The occurrence of matrix–alkali clusters had been previously reported [56,57], but not in the PMM literature. Because the number of proposed ''random cleavage peptides'' exceeds the number of unmat-

ched masses that they explain, their proposed interpretations cannot all be correct. However, the chymotryptic cleavages were supported by both repetition (nine consistent peaks could be explained as having chymotryptic cleavages at one terminus), and previous documentation (both in tryptic autolysates [52,57] and experiments specifically designed to evaluate pseudotrypsin [53,54]).

4.1. The impact of start codon reassignment and signal peptide cleavage

The effects of start codon reassignments and signal peptide cleavages on predicted protein p*I* values and molecular masses were examined. Molecular masses were calculated by summing the masses of all amino acids in a protein. Protein p*I* values were calculated using an algorithm designed after the work of Bjellqvist et al. [23,77]. Table 2 summarizes data for all ORFs that either had start codons reassigned or signal peptide cleavages predicted. The predicted p*I* values of several proteins changed rather dramatically, especially ORFs 163, 2149, 3272, and 3494, all of which shifted by more than 1 p*I* unit. The portion of a protein sequence that is removed when the start codon is reassigned may contain any number of charged residues, the removal of which can change the predicted p*I*.

Table 2
ORFs for which a new start codon was assigned, and/or a signal peptide cleavage was predicted

ORF	Original mass	Modified mass	Original p*I*	Modified p*I*
163	**50 225**	**43 539**	**9.48**	**6.43**
170	53 185	51 375	4.43	4.52
171	97 310	94 421	5.47	5.38
210	92 480	85 581	5.98	5.61
288	117 218	109 786	5.31	4.85
300	44 051	41 878	7.05	6.51
454	88 902	86 374	5.53	5.34
455	61 598	59 244	6.17	6.06
600	29 168	26 754	6.38	5.54
902	61 209	59 148	4.64	4.64
925	112 906	108 627	5.8	5.63
1375	49 865	47 278	5.59	5.24
1750	**69 353**	**67 237**	**5.98**	**5.81**
1914	22 360	20 177	9.71	9.51
2149	92 459	82 985	6.19	5.15
2257	29 567	27 155	9.66	9.39
2925	**39 763**	**37 921**	**6.27**	**6.65**
3146	88 828	86 718	8.56	8.27
3147	89 274	87 034	8.76	8.58
3228	30 511	28 351	9.23	8.91
3272	37 947	35 378	8.93	6.58
3392	**22 525**	**19 842**	**9.41**	**8.93**
3494	47 605	40 531	9.06	5.47
3584	99 145	96 622	6.1	5.91

Rows in bold represent cases in which the new N-terminal tryptic peptide was observed. Underlined rows represent those ORFs for which a new start codon was predicted. ORFs are numbered according to the latest release of the *Caulobacter* genome [61].

In isoelectric focusing, each protein should migrate to the gel position that has the same pH as its isoelectric point. Fig. 4A is a plot of predicted p*I* versus migration distance in the first dimension of the 2D separation. In cases where multiple gel spots were identified as the same protein, only the most basic spot was used to generate the plot, since, as discussed above, most modifications make proteins more acidic [22]. Fig. 4A has a substantial number of outlier proteins that migrate to positions more acidic than their sequences would predict. Fig. 4B displays the same data after accounting for signal peptide cleavages and start codon reassignments. The number of outliers is somewhat reduced. One of the remaining outliers is the spot containing ORF3444, the ORF with the putative frame-shift. Since the exact position of the frame-shift is not known, a precise p*I* prediction is impossible. Three of the other outliers are spots containing ORFs 792, 1460, and 1461. All three of these proteins are flagellins, and the p*I* values of flagellins in *Methanococcus jannaschii* have been shown to vary somewhat with different growth conditions [78]. Unfortunately, the shifts observed by Giometti et al. are not enough to explain the discrepancy between predicted p*I* values and observed gel migrations for these three spots. It is important to note that different algorithms yield varying p*I* predictions for the same protein sequence. All such programs derive their predictions by taking into account the number of charged residues and their unique pK_a values. Unfortunately, there does not appear to be universal agreement on which pK_a values to use. For example, chicken lysozyme has an observed p*I* of 11.1 [79]. The *Compute pI/MW* algorithm [80] found at SwissProt (us.expasy.org/tools/pi_tool.htm) predicts a p*I* of 9.3 for this

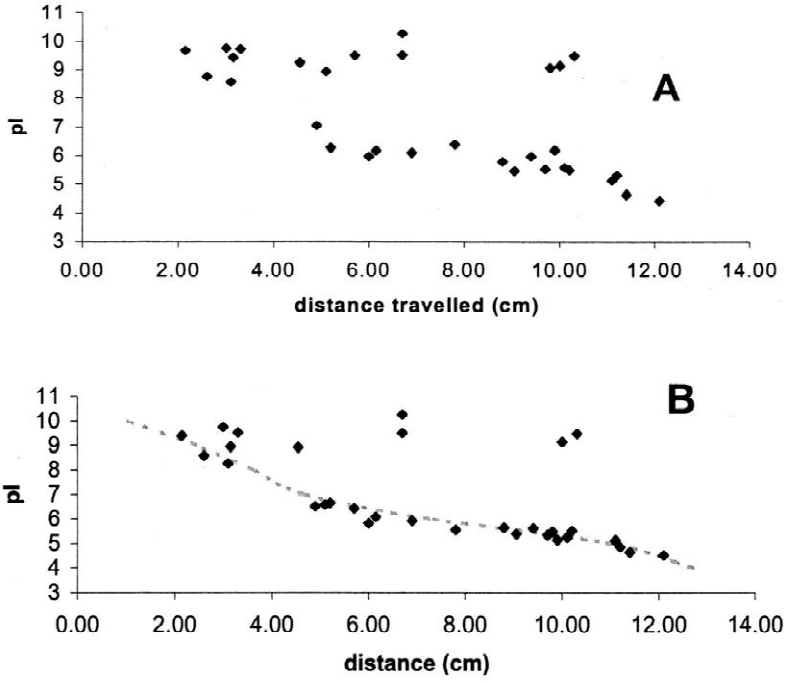


Fig. 4. Relation between predicted p*I* and migration distance (A) before and (B) after correcting for improper start position and signal peptide cleavage. The dashed line is the approximate pH gradient for the 3-10NL IPG strip. (Amersham Pharmacia Biotech technical note *Immobiline Drystrip Visualization of pH Gradients* found at: http://www2.apbiotech.com/applic/upp00738.nsf/vLookupDoc/185267373-F640/$file/18114060ab.pdf).

protein, and the *ProteinTools* program from ChemSW (www.chemsw.com) predicts 9.2. *Prodigies* predicts a p*I* of 10.9 for lysozyme. It should be noted that none of the programs took into account lysozyme's four disulfide bonds. Furthermore, for proteins in which the number of acidic and basic residues is nearly equal, any modification can have a dramatic effect on p*I*. For example, ORF CC1460 is one of the outliers in Fig. 4B. It has 22 basic residues and 21 acidic residues [61]. Removal of one basic residue changes the p*I* predicted by *Prodigies* from 9.13 to 7.59, and removing two basic residues drops the predicted p*I* to from 9.13 to 5.22. *Compute pI/MW* predictions reached similar conclusions.

The effect of reassigning start codons or cleaving signal peptides on the mass of a protein tends to be relatively insignifcant. In the SDS dimension, the migration distance is proportional to the log of the molecular mass [81]. Fig. 5 plots the log of the predicted molecular masses of the identified proteins as a function of spot migration distance. The mass changes induced by start codon reassignments and signal peptide cleavages hardly affected the log plot (data not shown).

The excellent agreement between predicted intact protein mass and gel migration distance in the SDS dimension suggests that the latter can provide useful information for identifying proteins. It can be used either to limit the number of candidates for a gel spot or to increase the confidence of an identification. Based on Fig. 4, predicted p*I* does not appear to be so useful for protein identification. Although considerable data scatter are evident, the plot of predicted p*I* versus gel migration distance would look even worse if cases of the same protein in multiple gel spots were not removed. That simplification was introduced only after all spots were analyzed. Clearly, predicted p*I* should be used judiciously as a parameter for peptide mass mapping studies of

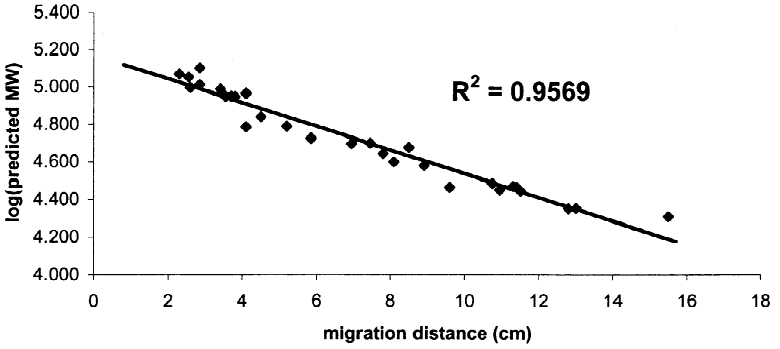


Fig. 5. Dependence of log(predicted MW) and SDS gel migration before correcting for start codon reassignment and signal peptide cleavage.

poorly annotated genomes, while intact molecular mass predictions are generally quite useful.

4.2. Effect of unmatched masses on peptide mass mapping

Fig. 6 contains mass spectra of the tryptic peptide extracts of gel spot 36. The letters denote masses corresponding to some of the phenomena described in the Section 3, and are defined in Table 3. This example demonstrates that both intense and weak unmatched peaks can be interpreted as modified tryptic peptides of the protein(s) already identified in a gel spot. Tables 3–5 are the master hit arrays

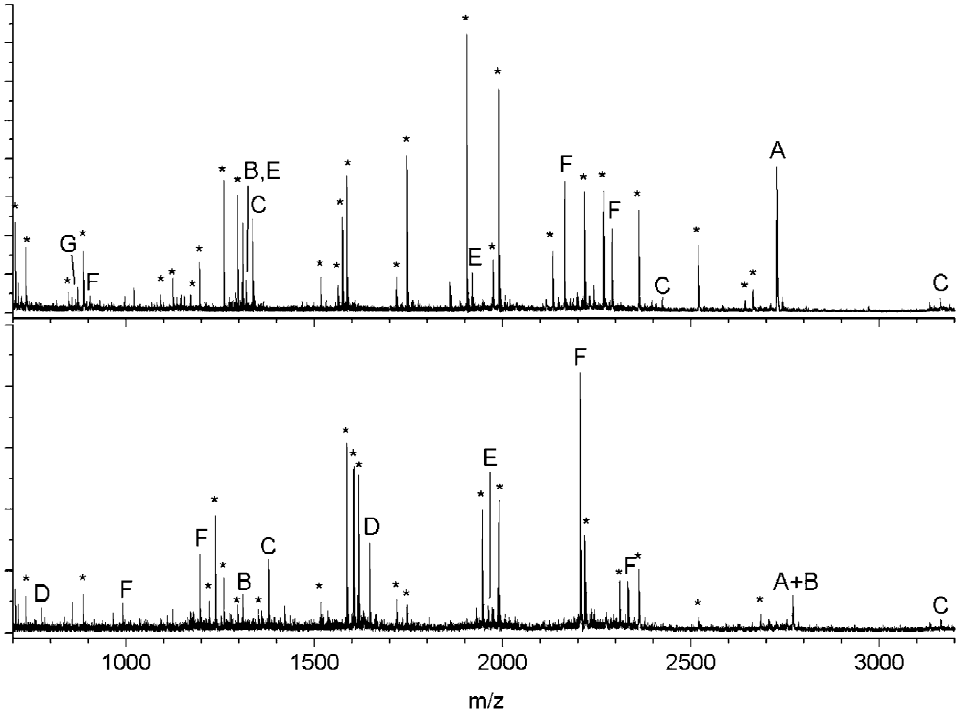


Fig. 6. Mass spectra obtained from the (top) unguanidinated and (bottom) guanidinated tryptic peptide extracts from gel spot 36. Asterisks denote peaks with masses corresponding to predicted tryptic peptides from ORF CC1750. The letters are defined in Table 3.

Table 3
Master hit array (MHA) for the unguanidinated tryptic digest of the proteins in gel spot 36

	1750 {27}	878 {9}	13 {8}	339 {8}	652 {8}	701 {8}	804 {8}
1902.87	−0.05	*	*	*	*	*	*
1988.95	−0.04	*	*	*	0.12	*	*
1744.98	−0.02	−0.06	*	−0.03	−0.08	−0.01	*
1258.64	0.01	0.06	−0.02	*	*	−0.01	0.03
1584.83	−0.04	*	*	*	0.02	*	*
2216.08	−0.04	*	*	0.10	*	*	0.07
1295.57	−0.01	*	*	*	*	*	*
2267.16	−0.03	*	*	−0.14	*	*	0.05
1573.73	0.00	*	*	*	*	0.03	−0.04
1335.71	C	0.00	*	*	*	*	*
2726.42	A	*	*	*	*	*	*
705.31	0.03	*	*	0.07	*	*	*
1309.69	B	*	−0.04	*	*	−0.11	*
2360.22	−0.03	*	*	*	*	*	−0.06
733.43	0.01	*	−0.13	*	*	*	*
885.41	0.01	*	0.12	*	*	*	*
2132.01	−0.04	*	*	*	*	*	*
1974.17	−0.07	*	*	*	*	*	*
1194.51	0.07	0.1	*	*	0.1	*	*
2518.37	−0.06	−0.03	*	*	0.09	*	*
1918.87	E	*	*	*	*	*	*
1515.85	−0.05	*	−0.04	*	−0.02	0.00	*
1123.51	0.07	0.01	*	*	*	*	*
1317.73	0.01	*	*	*	−0.12	*	*
1307.67	**E**	*	*	*	*	*	0.11
1561.84	0.00	*	*	0.02	*	*	*
871.03	G	*	*	*	*	*	*
2663.34	−0.05	0.11	*	*	*	*	*
846.33	0.06	*	0.11	0.14	*	*	*
1170.62	0.02	0.01	−0.02	*	*	0.01	0.10
1090.52	0.05	*	*	*	*	*	*
1715.85	0.00	0.03	*	*	*	0.11	*
3227.52	F	*	*	0.08	0.07	*	*
2422.3	C	*	0.01	*	*	−0.03	−0.09
2641.36	−0.06	*	*	−0.03	*	*	*
3160.51	C	*	*	*	*	*	*

The top row lists ORFs ranked by number of matches between theoretical peptides and observed masses. The experimentally measured masses are listed in order of decreasing intensity in the left-hand column. The total numbers of measured masses matched to each ORF are in braces. The fractional numbers represent differences between observed and theoretical masses. Capital letters represent initially unmatched masses that can be explained as: (A) new N-terminal tryptic peptide after signal peptide cleavage; (B) peptide with a deamidation at an NG sequence; (C) doubly oxidized tryptophan peptide; (D) peptide with N-terminal glycine guanidinated; (E) peptide contains one tryptic and one random cleavage; (F) trypsin autolysis or keratin peptide; (G) matrix–alkali cation cluster.

generated during the analysis of gel spot 36 [1]. The top row lists ORFs ranked by number of matches between theoretical peptides and observed masses. The experimentally measured masses are listed in order of decreasing intensity in the left-hand column. The total numbers of measured masses matched to each ORF are in braces. The fractional numbers represent differences between observed and theoretical masses. Most of the masses in the three MHAs that do not match ORF1750 can be explained by seven of the effects described above. Guanidinated data allow confirmation of putative assignments, and Table 4 is the MHA generated from the guanidinated mass spectrum. The ‘‘random’’ cleavage peptide

Table 4
MHA from the guanidinated tryptic digest of the proteins in gel spot 36: letters are defined in Table 3

	1750 {22}	503 {8}	373 {7}	701 {7}	826 {7}	1604 {7}	1778 {7}
1584.84	−0.05	*	*	*	*	*	*
1603.95	−0.09	*	*	*	*	*	*
1615.83	−0.08	*	−0.02	−0.05	0.06	*	*
1236.55	0.05	−0.02	*	*	0.11	*	0.08
1988.91	0.00	*	0.12	*	*	*	*
1944.86	−0.01	*	*	*	*	*	*
1645.91	D	*	−0.03	−0.06	−0.13	0.02	*
1377.66	C	*	0.11	*	*	*	*
2216.06	−0.01	*	*	*	*	*	*
1258.62	0.03	0.1	*	0.01	−0.04	*	*
2309.14	0.02	*	*	*	*	*	*
1986.9	*	*	*	*	*	*	*
705.51	E	−0.13	*	*	*	*	*
885.31	0.11	*	*	*	*	*	*
1309.63	B	*	0.03	−0.05	0.06	*	*
733.58	−0.14	*	*	*	*	*	*
2360.21	−0.02	*	*	*	*	*	*
1717.83	−0.13	0.04	*	*	*	*	*
1220.53	0.08	0.09	*	*	*	*	0.1
1515.84	−0.04	*	*	0.01	*	−0.09	*
1943.93	0.01	*	*	*	*	*	*
2769.44	A+B	*	*	*	*	0.01	*
712.39	*	*	*	*	*	0.08	*
1295.54	0.02	*	*	*	*	0.08	*
1960.87	E	*	*	*	*	*	*
1745.01	−0.05	−0.08	*	−0.05	*	*	−0.07
775.49	D	*	−0.10	*	*	−0.03	−0.12
1419.69	E	−0.07	−0.11	*	−0.05	*	*
1349.63	*	*	*	*	*	*	*
1123.67	−0.10	*	*	*	*	*	−0.08
1359.77	−0.01	*	*	−0.02	−0.09	−0.07	*
2683.34	−0.02	−0.07	*	*	*	*	0.00
3160.55	C	*	*	*	*	*	*
2518.38	−0.07	*	*	*	*	*	−0.08

proposed for the m/z 1307.67 ion (indicated in bold in Table 3) in the unguanidinated sample contained two lysines. A peak at m/z 1391.69 would be expected in the guanidinated mass spectrum. However, a peak at m/z 1349.63 (indicated in bold in Table 4) was seen in the guanidinated sample, implying that the peptide responsible for the m/z 1307.67 peak contained only a single lysine. These data suggest that the proposed ''random cleavage'' peptide is not present in the sample. This demonstrates the importance of a having a second technique such as guanidination to increase the credibility of assignments. For example, two of three tryptophan oxidations proposed in Table 3 were supported by guanidination as was the proposed deamidation peptide. Similarly, the proposed N-terminal fragment that arises after cleavage of the predicted signal peptide was observed in both unguanidinated and guanidinated data. Five new consistent peaks were identified following interpretation of the previously unmatched masses.

Consistent masses are the most important source of information for identifying proteins using the *Prodigies* algorithm. These pairs of mass spectral features that appear in both the unguanidinated and guanidinated spectra convey both mass and limited sequence information. The advantages of using consistent masses for peptide mass mapping have been

Table 5
Consistent MHA for gel spot 36: letters are defined in Table 3

	1750 {18}	652 {5}	701 {5}	804 {4}	2698 {4}	2806 {4}	3462 {4}
1902.87	−0.05	*	*	*	*	*	*
1988.95	−0.04	*	*	*	*	0.12	*
1744.98	−0.02	−0.08	−0.01	*	*	*	*
1258.64	0.01	*	−0.01	0.03	0.03	*	*
1584.83	−0.04	0.02	*	*	*	*	−0.01
2216.08	−0.04	*	*	*	*	*	0.04
1295.57	−0.01	*	*	*	*	*	*
2267.16	−0.03	*	*	0.05	−0.03	*	*
1573.73	0.00	*	0.03	−0.04	*	*	*
1335.71	C	*	*	*	0.02	*	*
1309.69	B	*	−0.11	*	*	*	*
2360.22	−0.03	*	*	−0.06	*	*	*
733.43	D	*	*	*	*	*	*
885.41	0.01	*	*	*	*	*	*
2132.01	−0.04	*	*	*	*	*	*
1194.51	0.07	0.1	*	*	0.13	*	*
2518.37	−0.06	0.09	*	*	*	*	*
1918.87	E	*	*	*	*	*	*
1515.85	−0.05	−0.02	0.00	*	*	*	*
1317.73	0.01	*	*	*	*	−0.08	0.06
1307.67	*	*	*	*	*	−0.05	*
1561.84	0.00	*	*	*	*	*	*
2641.36	−0.06	*	*	*	*	−0.06	−0.08
3160.51	C	*	*	*	*	*	*

thoroughly described [1]. Briefly, their use decreases the minimum number of matched peptides required to identify a protein by about 35% on average, in the process making experiments less sensitive to sources of chemical noise. In addition, use of consistent peaks limits the influence of random matches that arise because peptides from different proteins can have identical or nearly identical masses. Any loss of consistent peptides significantly frustrates the identification of proteins. Guanidination-induced deamidation of asparagine caused 31 consistent peaks to be missed in the analyses of the *Caulobacter* stalk gel spots. Conversely, three new consistent peaks were discovered when signal peptide cleavage was considered. Although spectra of gel spot 36 digests contained a great deal of information and the identification of the protein associated with it was not significantly affected by the phenomena described in this report, one or two additional consistent peptides can have a significant impact on the analyses of mass spectra with fewer peaks. For example, the mass spectra from the gel spot identified as ORF CC2369 contained only four consistent peaks, limiting the confidence of its assignment to less than 90%. A reinterpretation of the unmatched masses uncovered two additional consistent peaks allowing the identification to be made with 98% confidence.

4.3. Using the knowledge gained from interpreting unmatched masses:

The phenomena described in this document can be advantageously exploited in future PMM studies. The observation of a mass shift of 1 or 2 u upon guanidination could indicate the presence of one or two NG sequences in a peptide. Similarly, the observation of multiple guanidinated peaks for the same peptide suggests the presence of G at its N-terminus. This is analogous to the observation of two masses spaced by 16 u often implicating a methionine [7]. The ‘‘random signal peptide cleavage’’ algorithm was helpful in confirming the reas-

signed start codon for ORF163. Furthermore, many of the signal peptide cleavages sites proposed by *Prodigies* were confirmed by *SignalP 2.0.* These are all examples of unique diagnostic information gained without extra sample processing, such as digestion of a replicate sample with a second proteolytic enzyme. This type of analysis simply requires a second interpretation of the data that can be included in peptide mass mapping algorithms. One can envision improved scoring algorithms that account for these new effects when assigning confidence limits to identifications.

Autolysis peptides and matrix–alkali clusters are all excellent candidates for internal calibrants in tryptic digest mass spectra [57]. Uncommon autolysis peptides may prove useful as calibrants in some experiments, especially those above m/z 2300. The problem with these ''contaminant'' masses (autolysis peptides, matrix–alkali clusters, ''ubiquitous'' masses, and keratin peptides,) is that any analyte ion of similar mass is discarded without further analysis. For example, any *Caulobacter* peptide with a mass near m/z 2163.06 in an unguanidinated mass spectrum will not be identified since all signal at that m/z ratio is assumed to arise from the autolysis peptide of the same mass. Furthermore, in the case of low intensity peptide ions, the entire isotopic envelope near a ''contaminant'' ion may be rendered useless. Obviously, steps should be taken to limit the introduction of keratins and other contaminants during proteomics experiments to reduce the amount of extraneous or unproductive information.

Although not previously discussed, single amino acid substitutions can explain a number of our unmatched masses. These could arise from either genome sequencing errors or single point mutations. In order to assess the impact such substitutions might have on peptide mass mapping, *Prodigies* was directed to sequentially change each amino acid in every theoretical peptide for all ORFs listed in each MHA. This new list of peptide masses was then compared to the unmatched masses in each spectrum. Unfortunately, almost every unmatched mass that could be assigned to a peptide created by a single amino acid substitution yielded multiple proposed substitutions. Because it would be impossible to know which of these, if any, is valid without peptide sequencing, this approach is probably not fruitful.

5. Conclusions

Explanations for 590 of 782 previously unmatched masses encountered during peptide mass mapping experiments on electrophoretically separated *Caulobacter* stalk proteins were presented. These were derived only after proteins were identified based on other mass spectral data. Eleven masses could be explained as N-terminal tryptic fragments of proteins following signal peptide cleavage, and eight proteins were suggested as having improperly assigned start codons. Some of the shortened protein sequences led to radically altered p*I* predictions that were consistent with gel migration data. One putative frame-shift was identified. Twenty-four of the unmatched masses were assigned as tryptic peptides containing doubly oxidized tryptophans. Some of these occurred as consistent pairs of peaks. There was no correlation between tryptophan oxidation and guanidination or storage time. Forty possible deamidations were uncovered, 34 of which arose from peptides with NG in their sequences. The deamidation of NG was significantly enhanced by guanidination. Eighteen peptides had their N-termini guanidinated. Most importantly, the N-terminal residue in each was glycine. Nevertheless, the guanidination of N-terminal glycine does not always occur.

Three hundred and fifty-four of the masses could be explained as arising from 510 peptides with a ''normal'' tryptic cleavage on one side and a random or non-tryptic cleavage on the other. Eighty-six of these ''random'' cleavage peptides showed cleavage C-terminal to F, Y, and W residues, consistent with chymotryptic activity. Nine of these chymotryptic peptides were observed as consistent pairs. There was an anomalously high frequency of putative cleavages N-terminal to asparagine. Three of the ''random cleavage'' peptides displaying hydrolysis N-terminal to asparagine were observed as consistent pairs. Several uncommon autolysis peptides were described and summarized in Table 1. Twenty-three keratin peptides were observed with the most common being m/z 832 in unguanidinated mass spectra, and m/z 1108 and 1151 in guanidinated mass

spectra. Sixty-five masses consistent with matrix–alkali clusters were identified in unguanidinated mass spectra. Only 17 were identified in guanidinated samples, and the reduction most likely resulted from the purification step of the guanidination protocol.

Several ubiquitous yet unexplained mass spectral peaks (17 occurrences of m/z 2092.01, 12 of m/z 1335.76, seven of m/z 2383.98, and six each of m/z 1800.97, 2311.15, and 2465.09) were observed. Experiments are currently underway to identify these ions. Sequencing experiments needed to accomplish this have their own costs and benefits.

Acknowledgements

This work was supported by NIH grant number GM61336-02 and NSF grant CHE-0094579. The authors thank Dr. William A. Harris, Dr. Dariusz Janecki and Richard L. Beardsley for insightful discussions in the course of this work.

References

[1] J.A. Karty, M.M.E. Ireland, Y.V. Brun, J.P. Reilly, J. Proteome Res. 1 (2002) 325.
[2] W.J. Henzel, T.M. Billeci, J.T. Stults, S.C. Wong, Proc. Natl. Acad. Sci. USA 90 (1993) 5011.
[3] P. James, M. Quadroni, E. Carafoli, G. Gonnet, Biophysical and Biochemical Research Communications 195 (1993) 58.
[4] M. Mann, P. Hojrup, P. Roepstorff, Biol. Mass Spectrom. 22 (1993) 338.
[5] D.J.C. Pappin, P. Hojrup, A.J. Bleasby, Curr. Biol. 3 (1993) 327.
[6] J.R. Yates, S. Speicher, P.R. Griffin, T. Hunkapiller, Anal. Biochem. 214 (1993) 397.
[7] O.N. Jensen, M.R. Larsen, P. Roepstorff, Proteins: Struct. Funct. Genet. Suppl. 2 (1998) 74.
[8] A. Shevchenko, O.N. Jensen, A.V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H. Boucherie, M. Mann, Proc. Natl. Acad. Sci. USA 93 (1996) 14440.
[9] O.N. Jensen, A.V. Podtelejnikov, M. Mann, Anal. Chem. 69 (1997) 4741.
[10] P. Jungblut, B. Thiede, Mass Spectrom. Rev. 16 (1997) 145.
[11] M. Quadroni, P. James, Electrophoresis 20 (1999) 664.
[12] E. Nordhoff, V. Egelhofer, P. Giavalisco, H. Eickhoff, M. Horn, T. Przewieslik, D. Theiss, U. Schneider, H. Lehrach, J. Gobom, Electrophoresis 22 (2001) 2844.
[13] A. Stensballe, O.N. Jenson, Proteomics 1 (2001) 955.
[14] B.W. Gibson, K. Biemann, Proc. Natl. Acad. Sci. USA 81 (1984) 1956.
[15] D. Wassenberg, N. Wuhrer, N. Beaucamp, H. Schurig, M. Wozny, D. Reusch, S. Fabry, R. Jaenicke, Biol. Chem. 382 (2001) 693.
[16] M. Belghazi, K. Bathany, C. Hountondji, X. Grandier-Vazeille, S. Manon, J.-M. Schmitter, Eur. J. Mass Spectrom. 7 (2001) 101.
[17] J.C. Betts, P. Dodson, S. Quan, A.P. Lewis, P.J. Thomas, K. Duncan, R.A. McAdam, Microbiology 146 (2000) 3205.
[18] M.M.E. Ireland, J.A. Karty, J.P. Reilly, Y.V. Brun, Mol. Microbiol. 45 (2002) 1029.
[19] H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, Prot. Eng. 10 (1997) 1.
[20] A.P. Pugsley, Microbiol. Rev. 57 (1993) 50.
[21] N.D. Phadke, M.P. Molloy, S.A. Steinhoff, P.J. Ulnitz, P.C. Andrews, J.R. Maddock, Proteomics 1 (2001) 705.
[22] E. Gianazza, J. Chromatogr. A 705 (1995) 67.
[23] B. Bjellqvist, B. Basse, E. Olsen, J.E. Celis, Electrophoresis 15 (1994) 529.
[24] R.L. Lundblad, in: Techniques in Protein Modification, CRC Press, Boca Raton, FL, 1995.
[25] C. Kannichts, in: Posttranslational Modification Reactions: Tools for Function Proteomics, Humana Press, Clifton, NJ, 2002.
[26] M.J. Huddleston, R.S. Annan, M.F. Bean, S.A. Carr, J. Am. Soc. Mass Spectrom. 4 (1993) 710.
[27] R.J. Arnold, B. Polevoda, J.P. Reilly, F. Sherman, J. Biol. Chem. 274 (1999) 37035.
[28] S.A. Carr, M.J. Huddleston, M.F. Bean, Prot. Sci. 2 (1993) 183.
[29] V.N. Lapko, D.L. Smith, J.B. Smith, Prot. Sci. 10 (2001) 1130.
[30] J. Gobom, M. Schuerenberg, M. Mueller, D. Theiss, H. Lehrach, E. Nordhoff, Anal. Chem. 2001 (2001) 434.
[31] J.P. McCormick, T. Thomason, J. Am. Chem. Soc. 100 (1978) 312.
[32] S.L. Cohen, G. Ward, in: 50th ASMS Conference on Mass Spectrometry, Orlando, FL, 2002.
[33] A.-S. Petersson, H. Steen, D.E. Kalume, K. Caidahl, P. Roepstorff, J. Mass Spectrom. 36 (2001) 616.
[34] S.C. Hall, D.M. Smith, F.R. Masiarz, V.W. Soo, H.M. Tran, L.B. Epstein, A.L. Burlingame, Proc. Natl. Acad. Sci. USA 90 (1993) 1927.
[35] M. Hamdan, E. Bordini, M. Galvani, P.G. Righetti, Electrophoresis 22 (2001) 1633.
[36] M. Raftery, R. Cole, J. Biol. Chem. 241 (1966) 3457.
[37] F.L. Brancia, S.G. Oliver, S.J. Gaskell, Rapid Commun. Mass Spectrom. 14 (2000) 2070.
[38] R.L. Beardsley, J.A. Karty, J.P. Reilly, Rapid Commun. Mass Spectrom. 14 (2000) 2147.
[39] J.E. Hale, J.P. Butler, M.D. Knierman, G.W. Becker, Anal. Biochem. 287 (2000) 110.
[40] T. Keough, M.P. Lacey, R.S. Youngquist, Rapid Commun. Mass Spectrom. 14 (2000) 2348.
[41] R.L. Beardsley, J.P. Reilly, Anal. Chem. 74 (2002) 1884.
[42] D.R. Goodlett, A. Keller, J.D. Watts, R. Newitt, E.C. Yi, S. Purvine, J.K. Eng, P. von Haller, R. Aebersold, E. Kolker, Rapid Commun. Mass Spectrom. 15 (2001) 1214.

[43] M. Scigelova, P.S. Green, A.E. Giannakopulos, A. Rodger, D.H.G. Crout, P.J. Derrick, Eur. J. Mass Spectrom. 7 (2001) 29.
[44] E.S. Boja, H.M. Fales, Anal. Chem. 73 (2001) 3576.
[45] H.T. Wright, Crit. Rev. Biochem. Mol. Biol. 26 (1991) 1.
[46] S. Clarke, R.C. Stephenson, J.D. Lowenson, in: T.J. Ahern, M.C. Manning (Eds.), Stability of Protein Pharmaceuticals, Part A, Plenum, New York, 1992, p. 1.
[47] H. Sarioglu, F. Lottspeich, T. Walk, G. Jung, C. Eckerskron, Electrophoresis 21 (2000) 2209.
[48] K. Patel, R.T. Borchardt, Pharm. Res. 7 (1990) 703.
[49] D.L. Smith, A.R. Zhou, in: J.A. McCloskey (Ed.), Methods in Enzymology, Academic Press, New York, 1990, p. 374.
[50] J.S. Andersen, B. Kuster, A. Podtelejnikoc, E. Morts, M. Mann, in: 47th ASMS Conference on Mass Spectrometry and Applied Topics, Dallas, TX, 1999, p. 405.
[51] Y. Zhang, D. Figeys, R. Aebersold, Anal. Biochem. 261 (1998) 124.
[52] M.M. Vestling, C.M. Murphy, C. Fenselau, Anal. Chem. 62 (1990) 2391.
[53] R.L. Smith, E. Shaw, J. Biol. Chem. 17 (1969) 4704.
[54] V. Keil-Dlouha, N. Zylber, J.M. Imhoff, N.T. Tong, B. Keil, FEBS Lett. 16 (1971) 291.
[55] M. Landon, in: C.H.W. Hirs (Ed.), Methods in Enzymology, Academic Press, New York, 1977, p. 145.
[56] B.O. Keller, L. Li, J. Am. Soc. Mass Spectrom. 11 (2000) 88.
[57] W.A. Harris, D.J. Janecki, J.P. Reilly, Rapid Commun. Mass Spectrom. 16 (2002) 1714.
[58] K.C. Parker, J.I. Garrels, W. Hines, E.M. Butler, A.H.Z. McKee, D. Patterson, S. Martin, Electrophoresis 19 (1998) 1920.
[59] F.L. Brancia, A. Butt, R.J. Beynon, S.J. Hubbard, S.G. Oliver, S.J. Gaskell, Electrophoresis 22 (2001) 552.
[60] M. Fountoulakis, H. Langen, Anal. Biochem. 250 (1997) 153.
[61] W.C. Nierman, T.V. Feldblyum, M.T. Laub, I.T. Paulsen, K.E. Nelson, J. Eisen, J.F. Heidelberg, M.R. Alley, N. Ohta, J.R. Maddock, I. Potocka, W.C. Nelson, A. Newton, C. Stephens, N.D. Phadke, B. Ely, R.T. DeBoy, R.J. Dodson, A.S. Durkin, M.L. Gwinn, D.H. Haft, J.F. Kolonay, J. Smit, M.B. Craven, H. Khouri, J. Shetty, K. Berry, T. Utterback, K. Tran, A. Wolf, J. Vamathevan, M. Ermolaeva, O. White, S.L. Salzberg, J.C. Venter, L. Shapiro, C.M. Fraser, Proc. Natl. Acad. Sci. USA 98 (2001) 4136.
[62] W. Mo, Y. Ma, T. Takao, T.A. Neubert, Rapid Commun. Mass Spectrom. 14 (2000) 2080.
[63] P. Bornstein, Biochemistry 9 (1970) 2408.
[64] K. Nakai, M. Kanehisa, Proteins: Struct. Funct. Genet. 11 (1991) 95.
[65] Y.G. Nakamura, T. Ikemura, Nucleic Acids Res. 28 (2000) 292.
[66] S.P. Gygi, G.L. Corthals, Y. Zhang, Y. Rochan, R. Aebersold, Proc. Natl. Acad. Sci. USA 97 (2000) 9390.
[67] J.A. Karty, M.M.E. Ireland, Y.V. Brun, J.P. Reilly, in: American Society for Mass Spectrometry 50th Conference on Mass Spectrometry, Orlando, FL, 2002.
[68] Z. Zhang, D.L. Smith, J.B. Smith, Eur. J. Mass Spectrom. 7 (2001) 171.
[69] Y.-R. Hsu, W.-C. Chang, E.A. Mendiaz, S. Hara, D.T. Chow, M.B. Mann, K.E. Langley, H.S. Lu, Biochemistry 37 (1998) 2251.
[70] K. Patel, R.T. Borchardt, Pharm. Res. 7 (1990) 787.
[71] I.I. Stewart, T. Thomson, D. Figeys, Rapid Commun. Mass Spectrom. 15 (2001) 2456.
[72] M. Mann, in: 43rd ASMS Conference on Mass Spectromtery and Applied Topics, Atlanta, GA, 1995, p. 639.
[73] S. Gay, P.-A. Bins, D.F. Hochstrasser, R.D. Appel, Electrophoresis 20 (1999) 3257.
[74] G. Schoellman, E. Shaw, Biochemistry 2 (1963) 252.
[75] J.C. Powers, in: B. Weinstein (Ed.), Chemistry and Biochemistry of Amino Acids, Peptides, and Proteins, Marcel Dekker, New York, 1977, p. 65.
[76] P.V. Schoenlein, L.S. Gallman, M.E. Winkler, B. Ely, Gene 93 (1990) 17.
[77] B. Bjellqvist, G.J. Hughes, C. Paquali, N. Paquet, F. Ravier, J. Sanches, S. Frutiger, D. Hochstrasser, Electrophoresis 14 (1993) 1023.
[78] C.S. Giometti, S.L. Tollaksen, G. Babnigg, C.I. Reich, G.J. Olsen, H. Lim, J.R. Yates III, Eur. J. Mass Spectrom. 7 (2001) 207.
[79] D.E. Kuehner, J. Engmann, F. Fergg, M. Wernick, H.W. Blanch, J.M. Prausnitz, J. Phys. Chem. B 103 (1999) 1368.
[80] M.R. Wilkins, E. Gasteiger, A. Bairoch, J.-C. Sanchez, K.L. Williams, R.D. Appel, D.F. Hochstrasser, in: A.J. Link (Ed.), 2-D Proteome Analysis, Humana Press, Clifton, NJ, 1998, p. 531.
[81] T. Creighton, in: Proteins, Structures and Molecular Properties, W.H. Freeman, New York, 1993.