JOURNAL OF MASS SPECTROMETRY
J. Mass Spectrom. 2003; **38**: 809–816
Published online 22 July 2003 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jms.493

JMS

'Mass defect' tags for biomolecular mass spectrometry

Michael P. Hall,* Siamak Ashrafi, Imad Obegi, Robert Petesch, Jeffrey N. Peterson and Luke V. Schneider

Target Discovery, Inc., 4015 Fabian Way, Palo Alto, California 94303, USA

Received 2 December 2002; Accepted 22 April 2003

We present a new class of 'mass defect' tags with utility in biomolecular mass spectrometry. These tags, incorporating element(s) with atomic numbers between 17 (Cl) and 77 (Ir), have a substantially different nuclear binding energy (mass defect) from the elements common to biomolecules. This mass defect yields a readily resolvable mass difference between tagged and untagged species in high-resolution mass spectrometers. We present the use of a subset of these tags in a new protein sequencing application. This sequencing technique has advantages over existing mass spectral protein identification methodologies: intact proteins are quickly sequenced and unambiguously identified using only an inexpensive, robust mass spectrometer. We discuss the potential broader utility of these tags for the sequencing of other biomolecules, differential display applications and combinatorial methods. Copyright © 2003 John Wiley & Sons, Ltd.

KEYWORDS: mass-defect tags; biomolecular mass spectrometry; in-source fragmentation; electrospray ionization; protein sequencing

INTRODUCTION

Applications of biomolecular mass spectrometry can be drawn from the areas of identification and sequencing of proteins,[1–4] polynucleic acids[5–7] and polysaccharides.[8,9] Mass spectrometry has been successfully applied to probing biomolecular structure–function relationships such as protein–ligand and protein–protein interactions.[10–12] The ability to resolve stable isotopes has also made mass spectrometry useful for differential display applications.[13,14]

However, chemical noise in mass spectra arising from matrix impurities, fragmentation products or unidentified constituents can compromise spectral analysis. Incorporation of one or more elements having atomic numbers between 17 (Cl) and 77 (Ir), and more effectively between 35 (Br) and 63 (Eu), into the biomolecules of interest produces a discernible difference between the masses of tagged and untagged biomolecules of the same nominal mass. The stable nuclei of these elements have substantially greater absolute mass defect values than those elements common to biomolecules (carbon, hydrogen, oxygen, nitrogen, sulfur and phosphorus). The difference in mass defect manifests itself as a resolvable mass shift in most high-resolution mass spectrometers. The use of this class of labels, called 'mass defect' tags, can significantly reduce the complexity of mass spectra and allow efficient tracking of desired tagged species.

*Correspondence to: Michael P. Hall, Target Discovery, Inc., 4015 Fabian Way, Palo Alto, California 94303, USA.
E-mail: mike_hall@targetdiscovery.com

EXPERIMENTAL

Materials

Laboratory chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless indicated otherwise.

N-terminal labeling of myoglobin with the succinimidyl ester of 5-bromo-3-pyridylacetic acid

The succinimidyl ester of 5-bromo-3-pyridylacetic acid (Lancaster Chemical, Lancaster, UK) was prepared by adding 12.7 mg (59 μmol) of 5-bromo-3-pyridylacetic acid, 14.6 mg (127 μmol) of *N*-hydroxysuccinimide and 20.2 mg (105 μmol) of *N*-(3-dimethylaminopropyl)-*N*′-ethylcarbodiimide hydrochloride to 0.24 ml of anhydrous DMSO. The solution was incubated in the dark at ambient temperature for 24 h. Formation of product was determined to be ~90% by standard addition using electrospray ionization time-of-flight mass spectrometry (ESI-TOFMS). Horse apomyoglobin (1.89 mg, 111 nmol) was added to 0.54 ml of 5% (w/v) sodium lauryl sulfate (SDS) in water and denatured by heating for 10 min at 90 °C. Upon cooling, 1.89 ml of 9 M urea in 20 mM sodium phosphate (pH 7.0) was added to the myoglobin mixture. The succinimidyl ester of 5-bromo-3-pyridylacetic acid (0.24 ml, 51 μmol) was added to the denatured myoglobin and the reaction mixture was incubated overnight at ambient temperature in the dark. The reaction was quenched by addition of 0.027 ml of 2 M hydroxylamine in water, followed by incubation for 1 h. The reaction mixture was spin dialyzed (YM-10 Centricon, molecular weight cutoff (MWCO) 10 000) (Millipore, Bedford, MA, USA) against 25 mM tris(hydroxymethyl)aminomethane, 0.1% (w/v) SDS (pH 8.3) with eight 2 ml buffer exchanges. The retentate was collected (~0.6 ml) and SDS was removed by chloroform

extraction.[15] The precipitated protein was dried with nitrogen gas and was resuspended in 0.4 ml of 10% (v/v) glacial acetic acid in water. Protein concentration was determined by amino acid analysis.

N-terminal labeling of myoglobin with bromobenzaldehyde

Myoglobin (102 nmol) was denatured as described in the preceding section. 4-Bromobenzaldehyde (40 mg, 216 μmol) was added to 0.8 ml of a buffer containing 0.05 M sodium carbonate, 0.1 M sodium citrate, 9 M urea, 0.5% (w/v) SDS (pH 9.5). Myoglobin was transferred to the label solution and mixed. Sodium cyanoborohydride (0.02 ml of a 5 M stock in 1 M sodium hydroxide) was added. Methyl sulfoxide (0.1 ml) and acetonitrile (0.2 ml) were added to the sample to aid in dissolution of the label. The solution was stirred overnight at ambient temperature in the dark. The sample was then centrifuged to pellet undissolved label (14 000 *g*, 10 min). The supernatant was transferred to a dialysis unit (Slide-A-Lyzer, MWCO 10 000 (Pierce Chemical, Rockford, IL, USA) and dialyzed against 2l of 0.1% (w/v) SDS in water. The dialysis buffer was exchanged four times over 24 h. The dialyzed sample was transferred to a microcentrifuge tube and dried to completion in a rotary concentrator. SDS was removed by chloroform extraction.[15]

In-source fragmentation of labeled myoglobin

Labeled myoglobin was fragmented in a Mariner ESI-TOF mass spectrometer (Applied Biosystems, Foster City, CA, USA) that was tuned and calibrated according to the manufacturer's protocols. The protein was diluted with 50 : 50 (v/v) acetonitrile–water to a final concentration of 0.3 mg ml^{-1}. Acetic acid was added to the sample to a final concentration of 1.2% (v/v). The sample was sonicated briefly in a bath sonicator and centrifuged (14 000 *g*, 10 min). The sample was introduced into the mass spectrometer by continuous infusion through a 20 μm i.d. microspray capillary at a flow-rate of 1 μl min^{-1}. Fragmentation was induced by elevation of the nozzle potential. Individual 3 s spectra were accumulated for a total of 3 min. This translates to the introduction of 53 pmol of protein for fragmentation and ensuing analysis.

Mass spectral filtering and sequencing algorithms

Mass defect tagged peptides were filtered from non-tagged peptides (chemical noise) using a modification of the chromatographic peak deconvolution method described by Felinger and Pietrogrande.[16] The modification involved basing the deconvolution kernel on the average peak shape over the range 150–900 amu. The average peak shape was determined by least-squares fit of all the peaks after autoscaling each peak in the spectrum between the minimum and maximum counts within each 1.00464 amu of the spectrum. This average peak spacing was independently determined from both fast Fourier transform and least-squares methods. A simplex algorithm[17] was used to fit the height and position for each of two kernels within each peak in the mass spectrum.

Sequencing was accomplished using a cumulative probability algorithm.[18] At each residue length (n), an exhaustive prediction of the masses of all possible peptide sequences was made (i.e. 19 residues at each position since L and I have the same mass). Only the masses of the b-ions were used since no empirical evidence of a- and c-type ions was seen in the spectra. The corresponding counts from the mass spectrum were returned for each peptide mass and assigned a probability (p_j) within a cumulative log-normal probability distribution centered around the mean peak height for all competing sequence possibilities and exhibiting the standard deviation of competing peak heights:

$$p_j = \frac{\log(\text{counts}_j) - \log_{\text{mean}}(\text{counts})}{\sigma} \quad (1)$$

where

$$\log_{\text{mean}}(\text{counts}) = \sum_{\text{all } j \text{ sequences}} \log(\text{counts}_j) \quad (2)$$

and

$$\sigma = \frac{\sqrt{\sum_{\text{all } j \text{ sequences}} \log^2(\text{counts}) - \left[\sum_{\text{all } j \text{ sequences}} (\text{counts}_j)\right] / 19^n}}{19^n} \quad (3)$$

The final rank of each sequence was determined from the product of the probabilities for each preceding sequence of the PST series (i.e. a cumulative probability).

RESULTS AND DISCUSSION

Origin of the mass defect and incorporation into biomolecular tags

The mass defect is related to the nuclear binding energy released upon formation and stabilization of the nucleus of a given isotope.[19] By convention, the mass defect of ^{12}C is defined as zero atomic mass units (amu), and the mass defect of any other stable elemental isotope is calculated as the difference between the actual mass of the isotope (relative to the exact defined mass of ^{12}C as 12.00000 amu) and the isotope's nominal mass (i.e. the integer sum of the numbers of protons and neutrons).[20] The mass defects of other elements commonly found in biomolecules differ negligibly from that of carbon: For ^{14}N 0.0031, ^{16}O −0.0051 and ^{1}H 0.0078 amu. Sulfur and phosphorus, which are generally at lower abundance in biomolecules, exhibit slightly larger mass defects of −0.0279 and −0.0262 amu, respectively, for the most abundant isotopes ^{32}S and ^{31}P. An analysis of the mass defects for the most abundant stable nuclei[20] of all of the elements (Fig. 1) shows a maximum mass defect value of ∼−0.1 amu for elements with atomic numbers between 35 (Br) and 63 (Eu) (corresponding to the range of stable isotope mass numbers 80–150).

For high-resolution mass spectrometers, the ability to distinguish masses is dominated by the mass accuracy of the instrument at the low end of the mass-to-charge scale. For example, a mass accuracy of 30 ppm can resolve a 0.1 amu mass difference to a total mass of 3300 amu for a single charge state, assuming that instrumental resolution is not limiting. Greater mass defect discrimination is possible if

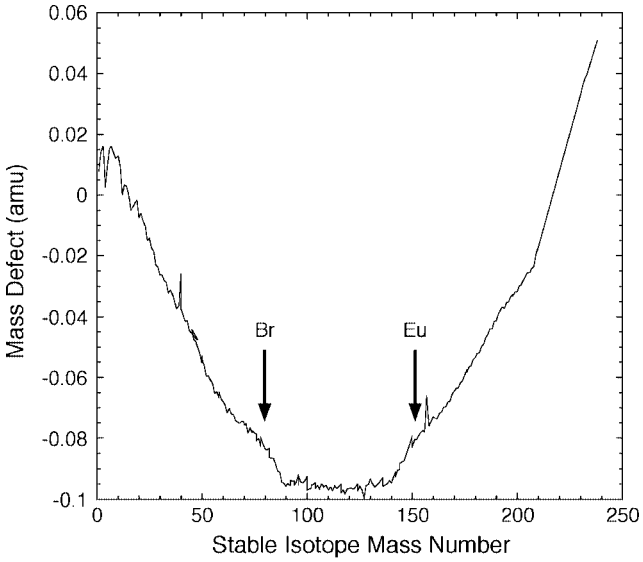

Figure 1. Trends in mass defect values for stable isotopes of the elements. Mass defect values (amu) are determined relative to ^{12}C, which is assigned a value of zero by convention. Stable isotopes with mass numbers from approximately 80 (Br) to 150 (Eu) have the largest absolute mass defect values of all the elements, differing by about −0.1 amu from carbon (range shown as arrows). The elements normally contained in biomolecules (e.g. C, H, N, O, S and P) have low stable isotope mass numbers resulting in less pronounced mass defects.

multiple mass defect elements are incorporated into a single tag. Bromine is a particularly good mass defect element in that it is easily incorporated into organic tags and has a nearly equivalent natural abundance of its two stable isotopes ^{79}Br and ^{81}Br.

Illustration of the use of mass defect tags in protein sequencing

Current mass spectral methods for protein identification include peptide mass fingerprinting[21–24] and sequencing by tandem MS.[25–27] Both techniques involve enzymatic or chemilytic digestion of the protein into smaller peptides prior to analysis. Although peptide mass fingerprinting is rapid, unequivocal protein identification becomes less likely as the size of the lookup database grows.[21,28,29] Sequencing via tandem MS generates a protein sequence tag (PST), which is a contiguous series of amino acids. Although the generation of a PST results in unambiguous identification, the processing time per protein can be prohibitively slow.[30]

Inverted mass ladder sequencing (IMLS) is a new methodology for high-speed determination of an N- or C-terminal PST from intact proteins by fragmentation in the ionization zone (i.e. in-source fragmentation) of an ESI-TOF mass spectrometer.[31,32] IMLS involves labeling the terminus of a protein with a unique mass tag that allows assembly of a PST by mass addition of fragment ions starting with the unique mass of the chemical tag. In-source fragmentation of whole proteins generates a multitude of fragment ions ('chemical noise'), giving rise to peaks at nearly every mass position in the spectrum with an average sequence-dependent peak spacing of 1.000464 amu (see Fig. 2(A)).

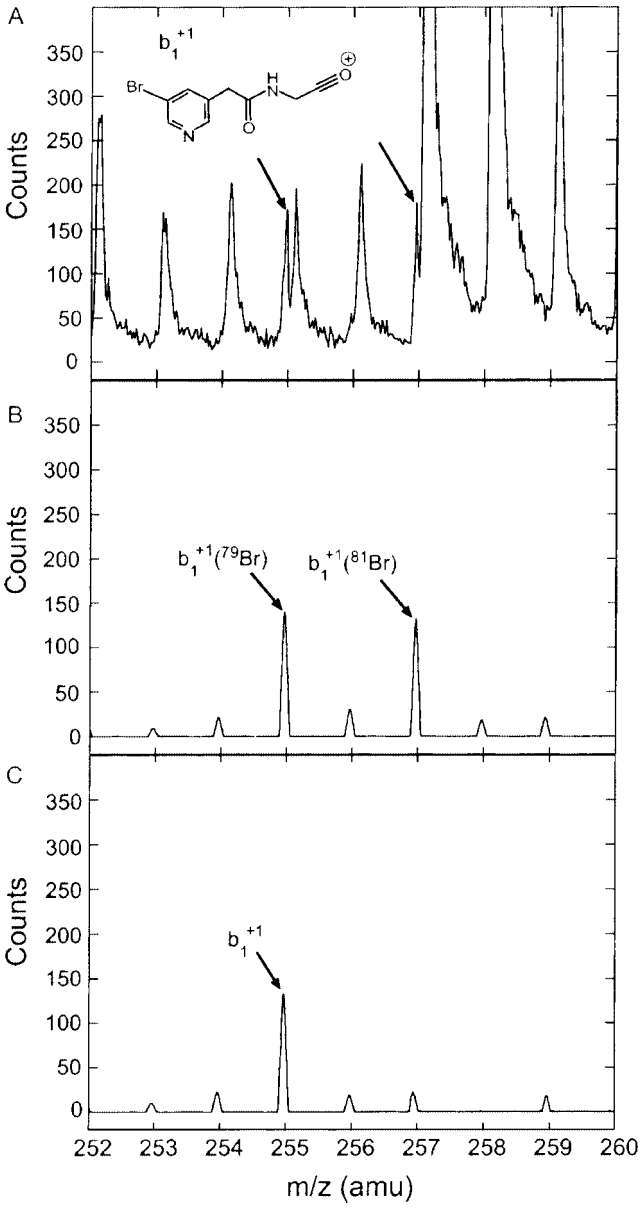

Figure 2. The region of the in-source fragmentation mass spectrum surrounding the $b_1{}^{+1}$ ion (the structure is shown) of myoglobin labeled at the N-terminus with the succinimidyl ester of 5-bromo-3-pyridylacetic acid. (A) The raw spectrum shows the periodic chemical noise, exhibiting a nearly 1 amu spacing, and the mass defect labeled $b_1{}^{+1}$ doublet shifted to the left of the chemical noise (arrows). (B) The same spectrum after algorithmic filtering of the chemical noise showing the $b_1{}^{+1}$ doublet peaks with approximately equal peak heights reflecting the natural 50 : 50 isotopic abundance of ^{79}Br and ^{81}Br. (**C**) The filtered spectrum from panel B with further peak qualification based on algorithmic peak pairing of the bromine doublets based on their expected relative abundance.

Although it has proved possible to assemble a terminal PST using any label distinguishable in mass from that of the amino acids found in the protein,[31] the use of a mass defect tag greatly enhances this process. Chemical noise peaks predominantly arise from unlabeled amino acid fragments, which exhibit minimal mass defects. Therefore, any fragment ion containing the mass defect tag should, on average, be

shifted by the value of the mass defect to the left (i.e. to lower m/z) of any chemical noise peak at the same nominal mass.

For example, myoglobin was labeled at the N-terminus with the succinimidyl ester of 5-bromo-3-pyridylacetic acid (50% N-terminal labeling as determined by one round of quantitative Edman sequencing). The labeled protein (53 pmol) was fragmented with a nozzle potential of 225 V. Figure 2(A) shows the region of the mass spectrum surrounding the tagged b_1-ion with the structure shown. This figure shows both the chemical noise peaks, separated by ~1 amu, and the resolved tagged b_1-ion, seen as a doublet shifted to the left of the periodic chemical noise by ~0.1 amu. These data can be algorithmically filtered both to baseline and to reduce a large portion of the periodic chemical noise (Fig. 2(B)).[18] The nearly 50 : 50 natural abundance of the two stable bromine isotopes allows further algorithmic peak pairing of the remaining mass defect peaks (Fig. 2(C)).[18] Thus, the N-terminal b_1-ion becomes the only major peak remaining in this region of the spectrum.

When the above algorithmic analyses are applied to the entire spectrum (Fig. 3(A)), the mass-defect peaks are easily identified as bromine doublets (Fig. 3(B)) and are seen as the only dominant peaks remaining in the spectrum after peak pairing (Fig. 3(C)). The remaining major peaks correspond to the b-ions of the tagged N-terminal sequence of myoglobin. Figure 3 shows the generation of the b-ion series for the four N-terminal amino acid residues of myoglobin. Using a sequencing algorithm[18] that tests and ranks every possible b-ion through the first six residues, we were able to recover the published sequence of myoglobin (GLSDGE) through six residues.

Similarly, we were further able to recover the N-terminal sequence of bovine ubiquitin (MQIFV) through five residues (data not shown). The sixth residue of ubiquitin, K, which can be side-labeled with the mass defect tag, was incompletely labeled, which confounded identification of that residue. We were also able to recover the first three residues of streptavidin (AEA) from IMLS spectra (data not shown). This shorter PST determination was a result of the subsequent discovery (by Edman sequencing) that the streptavidin used in this study was composed of an ~50 : 50 mixture of MAEA and AEA N-terminal isoforms that led to degenerate sequence possibilities at longer PST lengths.

To ascertain the minimum PST required for unambiguous protein identification, the percentage of unique proteins out of 4478 human protein sequences (minus any signal peptides) contained in SwissProt was plotted as a function of number of N-terminal residues (Fig. 4) (Z. Smilansky, Compugen, Tel Aviv, Israel, personal communication, 2000). It is clear that an asymptote is approached after approximately five residues; in other words, longer PSTs do not provide significantly more resolution of protein identities at the N-terminus. It should be noted that this asymptote falls short of 100% uniqueness for this database. This could indicate redundancy within the database or a high conservation of N-terminal sequences. A similar analysis was conducted for C-terminal PSTs predicted from 22 366 human gene clusters contained in the Compugen LEADS database (Fig. 4) (Z. Smilansky, Compugen, Tel Aviv, Israel, personal communication, 2000),

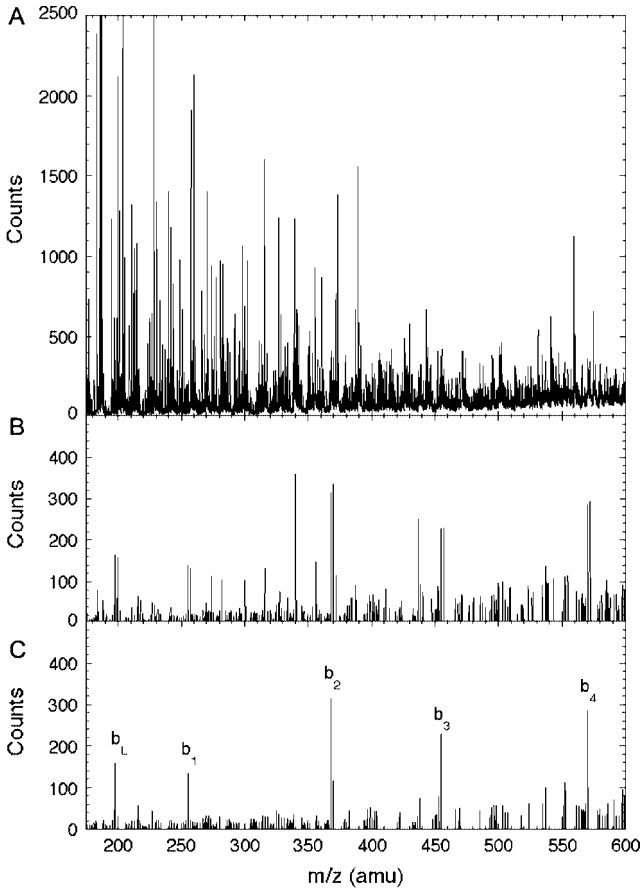

Figure 3. The 225 V (nozzle potential) in-source fragmentation mass spectrum (covering the *m/z* range between 150 and 600 amu) of myoglobin labeled at the N-terminus with the succinimidyl ester of 5-bromo-3-pyridylacetic acid. (**A**) The raw spectrum showing the substantial amount of chemical noise. (**B**) The spectrum after algorithmic filtering of chemical noise. The doublets corresponding to the bromine isotopes of the singly charged b-ions are now evident. (**C**) Further noise reduction based upon peak qualification resulting from bromine isotope pairing. The b-ions are now easily distinguished from the residual spectral noise. Peak b_L represents the b-type ion of the label itself, and b_1 through b_4 correspond to the calculated masses of the first four b-ions of the labeled N-terminal sequence of myoglobin (GLSD).

and a similar result obtained. Therefore, we conclude that negligible improvement is gained by sequencing more than 5–6 contiguous residues from either protein terminus. Since intact proteins are sequenced in IMLS it is also possible to use other coordinates such as molecular mass or isoelectric point to resolve ambiguous sequence identifications further. Alternatively, it may prove possible to sequence simultaneously from both termini by incorporating different numbers of mass defect elements into both N- and C-terminal tags.

Increased resolution with multiple mass defect elements

Myoglobin was doubly labeled at the N-terminus by reaction with excess 4-bromobenzaldehyde to examine the effect of incorporating multiple mass defect elements in IMLS.

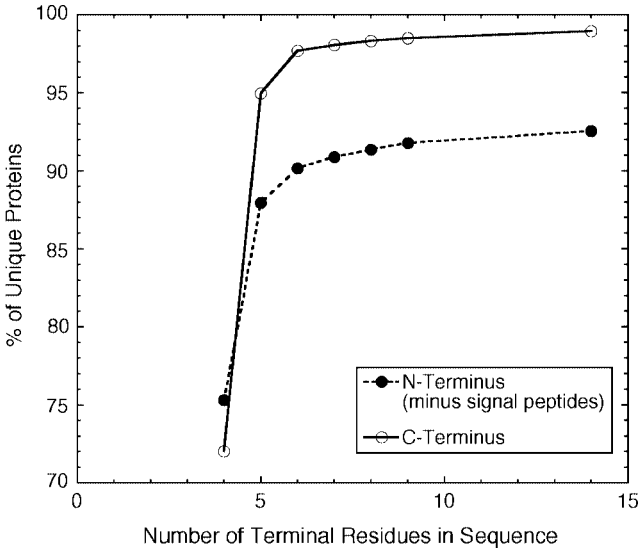

Figure 4. The percentage of unique proteins as a function of the number of contiguous terminal amino acid residues. The N-terminal data (●) were generated from 4478 human protein sequences contained in SwissProt. The C-terminal data (○) were generated from the Compugen LEADS database of 22 366 human gene clusters.

This sample was fragmented in-source as described in the previous section. The peaks corresponding to the doubly tagged, singly charged fragment ions appear as triplets shifted ~0.18 amu to the left of the chemical noise peaks. As an example, the mass spectral region around the tagged b_2-ion is shown (Fig. 5(A)). After algorithmic filtering of chemical noise, the triplet corresponding to the labeled b_2 fragment is immediately identifiable along with peaks arising from ^{13}C isotopes (marked with an asterisk). Peak identity is further corroborated by the splitting pattern, corresponding to the relative intensities (1 : 2 : 1) expected from the combinatorial pairs of two bromine isotopes. Multiple mass defect elements, therefore, increase the resolution of tagged species from chemical noise and would be expected to increase the ability to distinguish tagged from untagged species at higher mass-to-charge (m/z) values where instrumental resolution diminishes.

Generalization of the mass defect tag approach in IMLS

Although it is not possible to generate empirically spectra of all possible tagged sequences and chemical noise possibilities with multiple types of mass defect tags, it is possible to generalize this mass defect tag approach and estimate its limits using 'virtual' mass spectrometric data. TOF mass spectrometers have an intrinsic detector time resolution, or bin size, that is constant based on the square root of the m/z ratio. The mass spectrometer used to generate the protein sequencing data presented in this paper has a mass accuracy of ~45 ppm at 1000 amu. At lower m/z ranges used for IMLS, the virtual mass spectral peak width can be approximated by this mass accuracy, which is consistent with actual fragmentation spectra. It is possible then to calculate the exact mass of an ion and assign a count to the appropriate detector bin to create a 'virtual' mass spectrum of all singly charged a-,

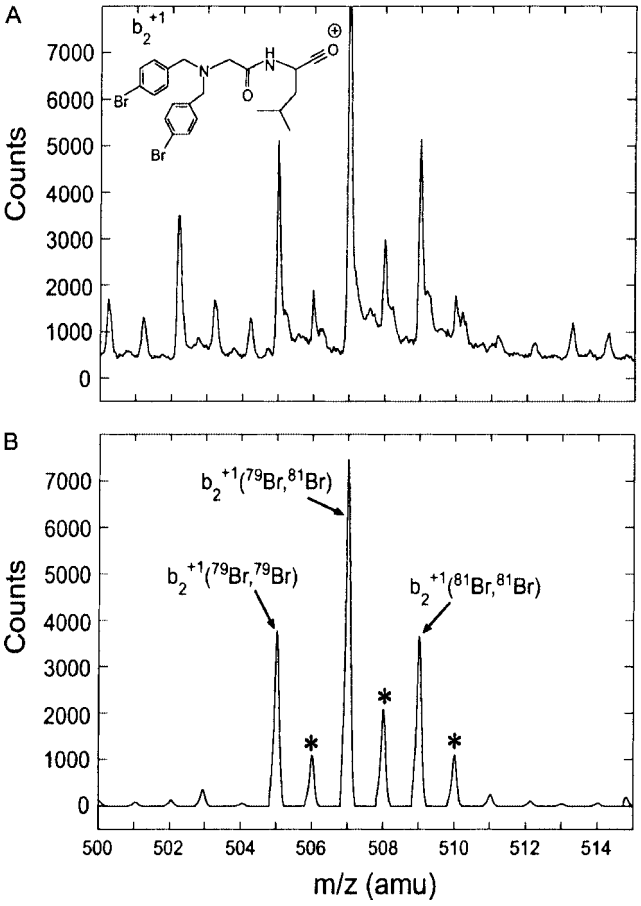

Figure 5. The region of the b_2^{+1} ion (the structure is shown) of the mass spectrum of fragmented myoglobin doubly labeled at the N-terminus with 4-bromobenzaldehyde. (**A**) The raw spectrum with chemical noise and (**B**) the filtered raw data. The triplet corresponding to the b_2^{+1} ion is unambiguous. A smaller triplet shifted to the right of the main triplet (marked with an asterisk) represents fragments containing one atom of ^{13}C.

b- and c-ions for every combinatorial possibility of peptide sequences up to 20 amino acids in length. This spectrum can be considered the chemical noise. A corresponding 'virtual' mass defect spectrum for various mass defect tagged analogs (i.e. pyridylacetic acid analogs shown in Fig. 6) of these peptide sequences can be constructed in the same manner.

The fraction of mass defect ions that do not overlap with any of the chemical noise ions is determined by comparing the contents of each bin in the mass defect spectrum with the corresponding bin in the chemical noise spectrum. The total number of mass defect sequences within each amu was determined from the sum of counts contained in all the bins spanning the amu. Similarly, the number of non-overlapping mass defect sequences was determined from the sum of counts in all bins within the same amu of the spectrum for which there are no counts in the corresponding bin of the chemical noise spectrum. Thus, the ratio of the number of non-overlapping to total number of mass defect peaks within each amu yields the fraction of non-overlapping mass defect sequences within each amu.

In Fig. 6 we plot the fraction of non-overlapping mass defect fragments within each amu of the 'virtual' spectrum

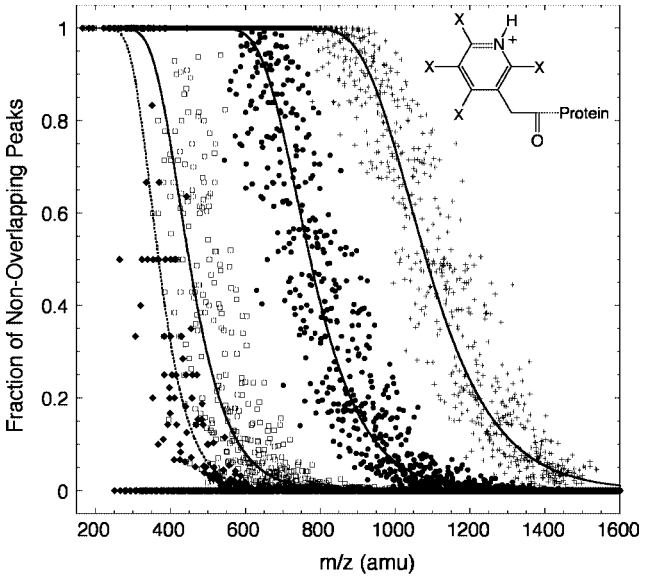

Figure 6. 'Virtual' mass spectral discrimination of mass defect tagged peptide ion fragments (incorporating differing types and numbers of mass defect elements (see structure) from chemical noise (i.e. all possible untagged fragment ions). Peptide fragments include all possible singly charged a-, b- and c-ions up to 20 amino acid residues in length. Closed diamonds (♦) are the data from the tag with four fluorine atoms (X = 4 F). Open squares (□), closed circles (●) and crosses (+) correspond to one bromine atom and three hydrogen atoms (X = 1 Br, 3 H), two bromine atoms and two hydrogen atoms (X = 2 Br, 2 H), and three bromine atoms and one hydrogen atom (X = 3 Br, 1 H), respectively. Within each nominal amu of the spectrum, the fraction of mass defect containing peptide ion fragments that do not overlap any chemical noise peak is determined. A value of 1 represents 100% discrimination of mass defect labeled ions peaks from chemical noise; a value of 0 represents the complete inability to discriminate any of the mass defect labeled ions within that amu. Each fraction is plotted as a function of the nominal *m/z* value of the amu. Only nominal amu regions that contain at least one mass defect peak are included.

that contains at least one mass defect sequence. This fraction quantifies the ability to discriminate mass-defect labeled peptide fragments from the chemical noise (i.e. unlabeled peaks) and is plotted as a function of the nominal mass of the amu for several possible acyl pyridylacetic acid derivatives. An exponential decay[33] can be fit to these data with the resulting fitted parameters providing an approximation of the range, $(m/z)_0$, over which each mass defect tag provides complete discrimination of the tagged peptides (Table 1). It is necessary to subtract the label mass from the fitted $(m/z)_0$ value to determine the practical dynamic range (Table 1). It is readily seen (Table 1) that a 45-fold improvement in the mass accuracy of the mass spectrometer (to 1 ppm at 1000 amu) only amounts to a doubling of the predicted $(m/z)_0$ of the mass defect and the superiority of a single Br versus four F is maintained. The upper limit to the number of bromine atoms that may be incorporated in the tag is most likely limited to four since the cumulative mass shift caused by five bromine atoms would begin to overlap double charge states in the mass spectrum.

Others have suggested the use of mass defect tags incorporating multiple fluorine atoms;[34] however, this 'virtual' mass spectrometric analysis shows that, even with the inclusion of four fluorine atoms in the tag, unambiguous discrimination of tagged peptides is not possible. This result is predictable since ^{19}F lacks significant mass defect (only −0.0016 amu per fluorine). The fitted parameters at 45 ppm shown in Table 1 are misleadingly optimistic for the four-F analog in that 38% of the 53 potential mass defect sequences within the predicted $(m/z)_0$ range completely overlap the 'virtual' chemical noise. These completely overlapped points are, of necessity, excluded from the curve fit. On the other hand, only 5.5% of the 115 potential mass defect sequences within the predicted $(m/z)_0$ range of the single Br analog completely overlap the 'virtual' chemical noise at 45 ppm. In addition, all these complete overlaps occur within a few amu of the predicted end of the $(m/z)_0$ range.

This 'virtual' mass spectrometric analysis provides a worst case scenario for mass defect tags since all possible peptide sequences are represented. In reality, only a few possibilities will exist at each amu, depending upon the sequence of the parent protein, which is the reason why we are able to discriminate empirically single mass defect fragments through the b_6-ion of myoglobin (756 amu) for a single Br mass defect tag at 45 ppm. Arguably, the 'virtual' mass spectrum includes many peptide sequences that likely do not exist in nature, such as polycysteine and polymethionine, which would approach the bromine mass defect after only four such residues because of the cumulative mass defect of sulfur in these amino acids.

CONCLUSION

Mass defect tags provide a powerful method for discriminating biomolecular ions of interest from chemical noise in the mass spectrum. Although we have demonstrated this with a protein sequencing example, this approach may also have utility in many other biomolecular mass spectrometric applications. For example, it would be possible to synthesize isotope-differentiated binding energy shift tags (IDBEST) with one or more isotopically pure bromine atoms to discriminate tagged peptides and proteins in differential display applications. The key advantage of a differential display strategy based on the mass defect is preservation of the relative abundance of each isotope peak because they are shifted away from any chemical noise. In addition, IDBEST tags automatically shift the peaks corresponding to labeled molecules in the mass spectrum, potentially eliminating the need for prior separation (e.g. affinity purification). Incorporating a non-extendable base containing one or more mass defect elements into DNA sequencing methodologies should allow discrimination of the resulting mass spectral sequence ladders from exogenous DNA in the sample. Furthermore, it should be possible to analyze the sequencing ladders for all four bases simultaneously if a different number of mass defect elements are incorporated into each terminal base. Just as a single mass defect element tag can be discriminated

Table 1. Theoretical probability of non-overlapping peptide ion fragments[a]

Equation:[30] $f = 1 - [1 - e^{-k(m/z)}]^n$; $(m/z)_0 = \frac{\ln n}{k}$

Mass defect element	No. of mass defect elements	Fitted parameters		$(m/z)_0$	Minimum effective dynamic range (amu)
		k (amu^{-1} × 10^2)	n		
For 45 ppm mass accuracy at 1000 amu:					
F	4	1.83 ± 0.09	600 ± 300	350 ± 30	156
Br	1	1.40 ± 0.06	400 ± 100	420 ± 30	228
Br	2	1.04 ± 0.02	2300 ± 500	740 ± 30	465
Br	3	0.80 ± 0.01	4000 ± 800	1040 ± 30	686
For 1 ppm mass accuracy at 1000 amu:					
F	4	0.55 ± 0.02	38 ± 5	690 ± 34	497
Br	1	0.56 ± 0.03	100 ± 20	784 ± 57	587

[a] An exponential decay equation with two adjustable parameters (k and n) can be used to fit the fraction of non-overlapping mass defect peptide sequences predicted from 'virtual' mass spectral data corresponding to each of the possible pyridyl mass defect element substitutions of 3-pyridylacetic acid cited in Fig. 6. The rate (in reciprocal amu) at which mass defect tag discrimination is lost at increasing m/z in the spectrum is given by the parameter k. The parameter n can be considered to be related to the potential number of mass defect peaks that can be fit into the spectrum before overlap with chemical noise occurs. The parameters k and n can be combined, by the equation shown, to yield an estimate of the mass-to-charge $(m/z)_0$ range over which the mass defect tag can be unambiguously discriminated. The effective dynamic range of each mass defect tag is then determined by subtracting the mass of the tag from the predicted $(m/z)_0$. The standard error of the estimate was partitioned to the standard deviation in each parameter using the Jacobian matrix.

from unlabeled chemical noise, tags containing different numbers of mass defect elements can be discriminated from one another within an amu. Therefore, mass defect tags may allow up to five different species with the same nominal mass to be discriminated in the mass spectrometer before hitting the double charge state limit. This suggests that mass defect tags may also extend the number of possible tags that can be discriminated simultaneously in combinatorial chemistry and high-throughput screening applications where mass tags are used.

REFERENCES

1. Li J, Tremblay TL, Wang C, Attiya S, Harrison DJ, Thibault P. Integrated system for high-throughput protein identification using a microfabricated device coupled to capillary electrophoresis/nanoelectrospray mass spectrometry. *Proteomics* 2001; **1**: 975.
2. Belghazi M, Bathany K, Hountondji C, Grandier-Vazeille X, Manon S, Schmitter JM. Analysis of protein sequences and protein complexes by matrix-assisted laser desorption/ionization mass spectrometry. *Proteomics* 2001; **1**: 946.
3. Ablonczy Z, Kono M, Crouch RK, Knapp DR. Mass spectrometric analysis of integral membrane proteins at the subnanomolar level: application to recombinant photopigments. *Anal. Chem.* 2001; **73**: 4774.
4. Ogorzalek Loo RR, Cavalcoli JD, VanBogelen RA, Mitchell C, Loo JA, Moldover B, Andrews PC. Virtual 2-D gel electrophoresis: visualization and analysis of the *E. coli* proteome by mass spectrometry. *Anal. Chem.* 2001; **73**: 4063.
5. Oberacher H, Wellenzohn B, Huber CG. Comparative sequencing of nucleic acids by liquid chromatography–tandem mass spectrometry. *Anal. Chem.* 2001; **74**: 211.
6. Edwards JR, Itagaki Y, Ju J. DNA sequencing using biotinylated dideoxynucleotides and mass spectrometry. *Nucleic Acids Res.* 2001; **29**: E104.
7. Walters JJ, Muhammed W, Fox KF, Fox A, Xie D, Creek KE, Pirisi L. Genotyping single nucleotide polymorphisms using intact polymerase chain reaction products by electrospray quadrupole mass spectrometry. *Rapid Commun. Mass Spectrom.* 2001; **15**: 1752.
8. Creaser CS, Reynolds JC, Harvey DJ. Structural analysis of oligosaccharides by atmospheric pressure matrix-assisted laser desorption/ionisation quadrupole ion trap mass spectrometry. *Rapid Commun. Mass Spectrom.* 2002; **16**: 176.
9. Deery MJ, Stimson E, Chappell CG. Size exclusion chromatography/mass spectrometry applied to the analysis of polysaccharides. *Rapid Commun. Mass Spectrom.* 2001; **15**: 2273.
10. Kim MY, Maier CS, Reed DJ, Deinzer ML. Site-specific amide hydrogen/deuterium exchange in *E. coli* thioredoxins measured by electrospray ionization mass spectrometry. *J. Am. Chem. Soc.* 2001; **123**: 9860.
11. Elviri L, Zagnoni I, Careri M, Cavazzini D, Rossi GL. Noncovalent binding of endogenous ligands to recombinant cellular retinol-binding proteins studied by mass spectrometric techniques. *Rapid Commun. Mass Spectrom.* 2001; **15**: 2186.
12. Last AM, Robinson CV. Protein folding and interactions revealed by mass spectrometry. *Curr. Opin. Chem. Biol.* 1999; **3**: 564.
13. Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnol.* 2001; **19**: 946.
14. Jimenez CR, Li KW, Dreisewerd K, Mansvelder HD, Brussard AB, Reinhold BB, Van der Schors RC, Karas M, Hillenkamp F, Burbach JP, Costello CE, Geraerts WP. Pattern changes of pituitary peptides in rat after salt-loading as detected by means of direct, semiquantitative mass spectrometric profiling. *Proc. Natl. Acad. Sci. USA* 1997; **94**: 9481.
15. Puchades M, Westman A, Blennow K, Davidsson P. Removal of sodium dodecyl sulfate from protein samples prior to matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* 1999; **13**: 344.
16. Felinger A, Pietrogrande MC. Decoding complex multicomponent chromatograms. *Anal. Chem.* 2001; **73**: 619A.
17. Besset DH. *Object-oriented Implementation of Numerical Methods: an Introduction with Java and Smalltalk*. Morgan Kaufmann: San Francisco, 2001.

18. Schneider LV, Petesch R, Hall MP. Methods for determining protein and peptide terminal sequences. PCT Patent Application WO 02/061661A2, 2002.
19. Bueche F. *Principles of Physics*. McGraw-Hill: New York, 1977.
20. Weast RC (ed). *CRC Handbook of Chemistry and Physics* (60th Edition). CRC Press: Boca Raton, FL, 1980; B236.
21. Bienvenut WV, Hoogland C, Greco A, Heller M, Gasteiger E, Appel RD, Diaz JJ, Sanchez JC, Hochstrasser DF. Hydrogen/deuterium exchange for higher specificity of protein identification by peptide mass fingerprinting. *Rapid Commun. Mass Spectrom.* 2002; **16**: 616.
22. Woo SH, Fukuda M, Islam N, Takaoka M, Kawasaki H, Hirano H. Efficient peptide mapping and its application to identify embryo proteins in rice proteome analysis. *Electrophoresis* 2002; **23**: 647.
23. Parker KC. Scoring methods in MALDI peptide mass fingerprinting: ChemScore, and the ChemApplex program. *J. Am. Soc. Mass Spectrom.* 2002; **13**: 22.
24. Lamer S, Jungblut PR. Matrix-assisted laser desorption–ionization mass spectrometry peptide mass fingerprinting for proteome analysis: identification efficiency after on-blot or in-gel digestion with and without desalting procedures. *J. Chromatogr. B* 2001; **752**: 311.
25. Adamczyk M, Gebler JC, Wu J, Yu Z. Complete sequencing of anti-vancomycin fab fragment by liquid chromatography–electrospray ion trap mass spectrometry with a combination of database searching and manual interpretation of the MS/MS spectra. *J. Immunol. Methods* 2002; **260**: 235.
26. Li L, Masselon CD, Anderson GA, Pasa-Tolic L, Lee SW, Shen Y, Zhao R, Lipton MS, Conrads TP, Tolic N, Smith RD. High-throughput peptide identification from protein digests using data dependent multiplexed tandem FTICR mass spectrometry coupled with capillary liquid chromatography. *Anal. Chem.* 2001; **73**: 3312.
27. Fernandez-de-Cossio J, Gonzalez J, Betancourt L, Besada V, Padron G, Shimonishi Y, Takao T. Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by 'SeqMS', a software aid for *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 1998; **12**: 1867.
28. Li G, Waltham M, Anderson NL, Unsworth E, Treston A, Weinstein JN. Rapid mass spectrometric identification of proteins from two-dimensional polyacrylamide gels after in gel proteolytic digestion. *Electrophoresis* 1997; **18**: 391.
29. Clauser KR, Hall SC, Smith DM, Webb JW, Andrews LE, Tran HM, Epstein LB, Burlingame AL. Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. *Proc. Natl. Acad. Sci. USA* 1995; **92**: 5072.
30. Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko A, Boucherie H, Mann M. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* 1996; **93**: 14 440.
31. Schneider LV, Hall MP, Peterson JN. Methods for sequencing proteins. US Patent 6379971, 2002.
32. Hall MP, Ashrafi S, Petesch R, Schneider LV. A method for sequencing intact proteins by in-source fragmentation using novel 'mass-defect' labels. In *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, 2002.
33. Schneider LV. Metabolic uncoupling in *Escherichia coli* during phosphate-limited growth. Doctoral Dissertation, Princeton University, Princeton, NJ, 1997.
34. Schmidt G, Thompson A, Johnstone R. Compounds for mass spectrometry comprising nucleic acid bases and aryl ether mass markers. PCT Patent Application WO 99/32501, 1999.