Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Mass Spectrometry 234 (2004) 79–87
International Journal of Mass Spectrometry
www.elsevier.com/locate/ijms

Utility of accurate monoisotopic mass measurements to confidently identify lambda exonuclease generated single-stranded amplicons containing 7-deaza analogs by electrospray ionization FT-ICR mass spectrometry

Jennifer L. Frahm [a,b], Christopher J. Mason [b], David C. Muddiman [b,c,*]

[a] *Department of Chemistry, Virginia Commonwealth University, P.O. Box 842006, 1001 West Main Street, Richmond, VA 23284, USA*
[b] *W.M. Keck FT-ICR Mass Spectrometry Laboratory, Mayo Proteomics Research Center, Rochester, MN 55905, USA*
[c] *Department of Biochemistry and Molecular Biology, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA*

Received 24 October 2003; accepted 4 February 2004

Available online 18 March 2004

Abstract

A 53-base pair region on the long arm of chromosome 22 was amplified using PCR with 7-deaza-modified deoxynucleotides. Increased amplification efficiency was achieved by doubling the concentration of the modified deoxynucleotide triphosphates. Incorporation of 7-deaza purines has been previously shown to selectively eliminate fragmentation pathways during gas-phase sequencing of nucleic acids by sustained off-resonance irradiation collision-induced dissociation (SORI-CID) and infrared multiphoton dissociation. However, 7-deaza analogs result in significant duplex stability precluding interrogation of the single-stranded species by tandem mass spectrometry. Herein, we demonstrate the use of lambda exonuclease to successfully overcome this problem and are able to obtain single-stranded PCR products containing 7-deaza adenine and guanine nucleobases. Mass accuracy was used as our metric to determine complete incorporation of 7-deaza residues in PCR products $>$ 15 kDa; $\leq$ 3 ppm neutral monoisotopic mass measurement accuracies were routinely achieved. High mass measurement accuracy was obtained using a dual electrospray source and subsequently, using an isotopic fitting algorithm, the best fit between the theoretical and experimental isotopic distributions was determined using a chi-square value. Theoretical isotopic distributions were generated using an average nucleotide (*averatide*) chemical formula developed herein which was based on the relative frequencies of AT and GC base pairs in the human genome. Single-stranded PCR products were fragmented using SORI-CID and as expected, cleavage at the 7-deaza modified sites was not observed. Collectively, this integrated approach can facilitate top–down sequencing of PCR products by a variety of tandem mass spectrometry methods.

Keywords: FT-ICR; Dual-electrospray source; 7-Deaza; Mass measurement accuracy; Average nucleotide

1. Introduction

The completion of the Human Genome Project [1,2] and progress toward a better understanding of disease mechanisms have resulted in a focus on developing DNA "re-sequencing" methods. An advancement of any technique demands more rapid, accurate, and sensitive technologies. Nucleic acid sequencing by mass spectrometry offers these qualities over other methods as it affords rapid analysis times and it measures an intrinsic property of the molecule (i.e., mass).

There are two general approaches to nucleic acid sequencing by mass spectrometry. One approach is to measure the mass of enzymatic or chemical cleavage products and the second approach is gas-phase dissociation of oligonucleotides. Both approaches have been reviewed extensively [3–6]. Although enzymatic or chemical cleavage products of larger oligonucleotides can be routinely analyzed by mass spectrometry [7–9], the incubation or reaction times increase the overall analysis time. Gas-phase sequencing methods in principle can be faster due to the fact that no additional sample preparation is required.

* Corresponding author. Tel.: +1-507-284-1997; fax: +1-507-284-9261.
E-mail address: muddiman.david@mayo.edu (D.C. Muddiman).

1387-3806/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2004.02.004

Ubiquitous acceptance of gas-phase sequencing has been impeded by the high proton affinity of the purines [10,11]. High proton affinity results in the precursor ion current (abundance) being dispersed into numerous product ion channels thereby lowering the signal-to-noise ratio of each product ion. The majority of gas-phase sequencing of nucleic acids has been of oligonucleotides containing $\leq$20 nucleotides [12–17]. Nonetheless, the success of gas-phase sequencing is not limited to oligonucleotides less than twenty base pairs in length. Little et al. [18] have reported complete gas-phase sequencing of a 50-mer and partial sequencing of a 108-mer.

Although gas-phase oligonucleotide sequencing has been demonstrated, an overall strategy that allows routine sequencing of larger nucleic acids needs to be explored. One strategy to aid gas-phase sequencing by mass spectrometry would be a selective reduction in the number and type of fragmentation pathways through the incorporation 7-deaza purines [19]. The 7-deaza analogs contain a vinylinic carbon at the most basic site in guanine and the third most basic site in adenine [11], N7, in place of nitrogen. When the N7 nitrogen is replaced by carbon, the latter of which can no longer hydrogen bond, the phosphate group is unable to protonate the N7 position and hence cannot induce cleavage of the N-glycosidic bond [15,20,21]. As a result, cleavage does not occur at positions containing 7-deaza analogs.

The replacement of the N7 nitrogen with carbon changes the number of π-electrons, consequently affecting the aromaticity of the ring. This, along with other factors that affect stability [20,22,23], make it difficult to denature PCR products of any appreciable length containing 7-deaza purines into single stranded species. Ordinarily, changes in solution composition and temperature present a rapid means of denaturing PCR products, thereby affording the ability to measure the complementary single-stranded species [24]. However, due to the increased stability of the double stranded amplicons containing 7-deaza residues, an alternative approach (e.g., enzymatic) to generate single-stranded species is required.

Lambda exonuclease is a DNA repair enzyme that selectively digests a 5′ phosphorylated terminus from a duplex structure leaving the complementary strand intact [25–27]. Our laboratory has found that lambda exonuclease is the most reproducible method to selectively generate single-stranded PCR products [28].

The detection of single-strand species reduces the overall mass by roughly a factor of two, thus increasing the likelihood of achieving isotopic resolution [28]. Isotopic resolution can dramatically improve mass measurement accuracy (vide infra), which can be used to calculate the base composition and length of an amplicon [29]. A base composition that differs from the wild-type sequence would warrant gas-phase sequencing to elucidate the position of the sequence variation.

Accurate mass measurements (low to sub ppm) in Fourier transform-ion cyclotron resonance (FT-ICR) mass spectrometry are achievable using internal calibration [13,30–32]; the latter ensures that the analyte and calibrant ions experience the same experimental conditions (e.g., ion number). It has also been shown that tandem mass spectrometry data can be internally calibrated to obtain high mass measurement accuracies [13]. Even with the precision and accuracy of internally calibrated FT-ICR data, assignment of the most abundant isotopic mass may actually differ by $\pm$1 Da due to the experimental variation in ion abundances [33]. Moreover, average mass measurements are affected by isotopic variability, especially the abundance of carbon isotopes which are a major component of oligonucleotides [33,34], as well as by the dependence of average mass on measured isotope abundances.

Herein, we describe a strategy to facilitate top down sequencing that combines both the incorporation of 7-deaza residues to selectively eliminate fragmentation pathways and the generation of single-stranded amplicons by means of lambda exonuclease digestion. Moreover, we demonstrate how a simple isotopic fitting procedure can generate accurate experimental monoisotopic masses, which can then be used to delineate base composition and length to determine if tandem mass spectrometry should be invoked to elucidate the position of the sequence variation compared to the wild-type sequence. It is important to note that the ability of the isotopic fitting algorithm to correctly determine monoisotopic mass of internally calibrated data proved essential to routinely obtain low ppm mass measurement accuracies of large PCR products.

2. Experimental

2.1. PCR amplification

A 53-base pair region containing a single nucleotide polymorphism (SNP) (dbSNP, rs#717336) was amplified from synthetic template designed from a sequence on the long arm of chromosome 22 (SNP 22). Each PCR reaction contained 1× *AmpliTaq* Gold PCR buffer, 3 mM $MgCl_2$, 0.2 mM each deoxynucleotide triphosphate (dNTP), 12.5 pmol of each primer (forward sequence, 5′-GGG GCC ATC AGA GAA ATA TAC CT-3′, reverse sequence, 5′-CCA GGG TGG TAA TTT CCA GGT-3′) (Midland Certified Reagent Co., Midland, TX), 1 fmol template (Midland Reagent Co., Midland, TX), and 1.25 U of *AmpliTaq* Gold (Applied Biosystems, Foster City, CA) in a total volume of 50 μL. Amplification was carried out in a 96-well MJ Research PTC 200 DNA Engine System Peltier Thermal Cycler (Watertown, MA) using a hot-start PCR program. An initial 10-minute heat step at 95 °C is required for enzyme activation, followed by 34 cycles of 95 °C for 30 s (denaturation), 58 °C for 45 s (annealing), and 72 °C for 45 s (elongation). A final elongation step was accomplished at 72 °C for 10 min.

2.2. PCR amplification with 7-deaza analogs

PCR reactions containing 7-deaza analogs were prepared in accordance with the above except that 7-deaza-2′-deoxy-GTP (c^7-dGTP) (Roche Diagnostics Corp., Indianapolis, IN) and 7-deaza-2′-deoxy-ATP (c^7-dATP) (TriLink BioTechnologies, San Diego, CA) were substituted for deoxyguanosine triphosphate (dGTP) and deoxyadenosine triphosphate (dATP), respectively in separate reactions with a final concentration of 400 μM per reaction.

2.3. Primer phosphorylation with T4 polynucleotide kinase

A 5′ phosphorylated reverse primer was employed for PCR products that were to be treated with lambda exonuclease. Primer phosphorylation was accomplished with the enzyme T4 polynucleotide kinase (PNK) [35]. Each 50 μl phosphorylation reaction contained 1× PNK reaction buffer, 25 nmol ATP, 1.25 nmol reverse primer, and 5 U of PNK (Epicentre, Madison, WI). Reactions were incubated for 30 min at 37 °C followed by an enzyme deactivation step at 70 °C for 5 min. Primers were completely phosphorylated when analyzed by ESI-FT-ICR mass spectrometry (data not shown).

2.4. Generation of single-stranded amplicons

PCR products were ethanol precipitated prior to lambda exonuclease treatment and then digested according to the previously published protocol [28]. The resulting single-stranded PCR products were purified before ESI-FT-ICR MS analysis according to standard laboratory practice comprised of ethanol precipitation [36,37] and microdialysis [38,39].

2.5. Mass spectrometry

Mass spectra were acquired in negative-ion mode on a modified ESI-FT-ICR mass spectrometer (IonSpec Inc., Irvine, CA) with a 7 T superconducting magnet (Cryomagnetics, Oak Ridge, TN). PCR products were electrosprayed from a 50 μm fused-silica capillary pulled to a fine tip and remotely coupled to a potential of −2000 V [40]. PCR products were electrosprayed from a buffer consisting of 60:20:20 acetonitrile/isopropanol/10 mM ammonium acetate in water with a final concentration of 20 mM each piperidine and imidazole and infused at a rate of 3 nl s^{-1}. Single-acquisition spectra were collected with 1024 k data points with a digitization rate of 1 MHz and a Blackman window function applied then zero-filled three times prior to fast-Fourier transform, unless otherwise noted. All spectra of intact PCR products were internally calibrated using a dual-electrospray source fashioned in our laboratory and described elsewhere [30]. All SORI experiments were conducted using a 1000 Hz off-resonance excitation waveform.

Internally calibrated data for an individual charge state was used to generate an experimental monoisotopic mass, as determined by an algorithm similar to that described by Senko et al. [34], from which mass measurement accuracies were determined. Software was written in C++ on Linux.

3. Results and discussion

3.1. ESI-FT-ICR mass spectrometry of wild-type sequence

A model system was used to develop and evaluate a methodology to generate and accurately measure the mass of single-stranded PCR products. For this evaluation we chose the SNP 22 locus as our model system since we were familiar with the locus from previous experiments [41]. PCR products for wild-type SNP 22 were amplified and then the noncoding strand was digested. The sequence for the coding strand with a theoretical neutral monoisotopic mass of 16 221.739 Da is shown in Fig. 1a. A single-acquisition ESI-FT-ICR mass spectrum of the coding strand is shown in Fig. 1b, charge states are labeled and asterisks indicate the internal mass calibrant peaks. Polyethylene glycol with an average molecular weight of 1000 Da (PEG-1000) was used as the internal mass calibrant.

It is advantageous to use neutral monoisotopic masses opposed to the most abundant isotopic mass (vide supra). In addition, monoisotopic mass is additive, therefore, changes in mass relative to the wild-type mass (e.g., polymorphisms, mutations, base modifications) are readily assignable. Monoisotopic mass was determined using a chi-square procedure, described by Senko et al. [34], which aligns a theoretical isotopic distribution to the experimental isotopic distribution. Theoretical isotopic distributions were generated using the Mercury algorithm [42]. A number of possible distribution alignments are obtained by shifting the theoretical distribution in $\sim 1/z$ increments relative to the experimental distribution. Then, for each alignment of distributions, a chi-square value between experimental and theoretical abundances was calculated. Only peaks where the experimental abundance was greater than 20% of the most abundant isotopic peak in the distribution were included. The alignment with the lowest chi-square value was taken to be optimal. Monoisotopic mass was then calculated by an abundance weighted average of differences for each isotopic peak between experimental mass and theoretical monoisotopic mass [43].

Senko et al. utilized a model amino acid, *averagine*, whose chemical composition was obtained by averaging the relative occurrence of each amino acid in a database. We adopted a similar procedure and since the base composition of the wild-type sequence was known, a sequence-specific approach was used to validate our average nucleotide, "*averatide*," approach.

The frequencies of the bases in the human genome are known with reasonable certainty, which enabled us to

(a) 5' GGG GCC ATC AGA GAA ATA TAC CTC TTA TTA CCA CCT GGA AAT TAC CAC CCT GG 3'

(b)

∇	χ^2 Averatide	χ^2 Sequence
1	2.856	3.024
2	1.014	1.084
3	0.259	0.281
4	0.035	0.036
5	0.183	0.168
6	1.051	1.006
7	4.250	4.114
8	17.199	16.707
9	28.467	27.677

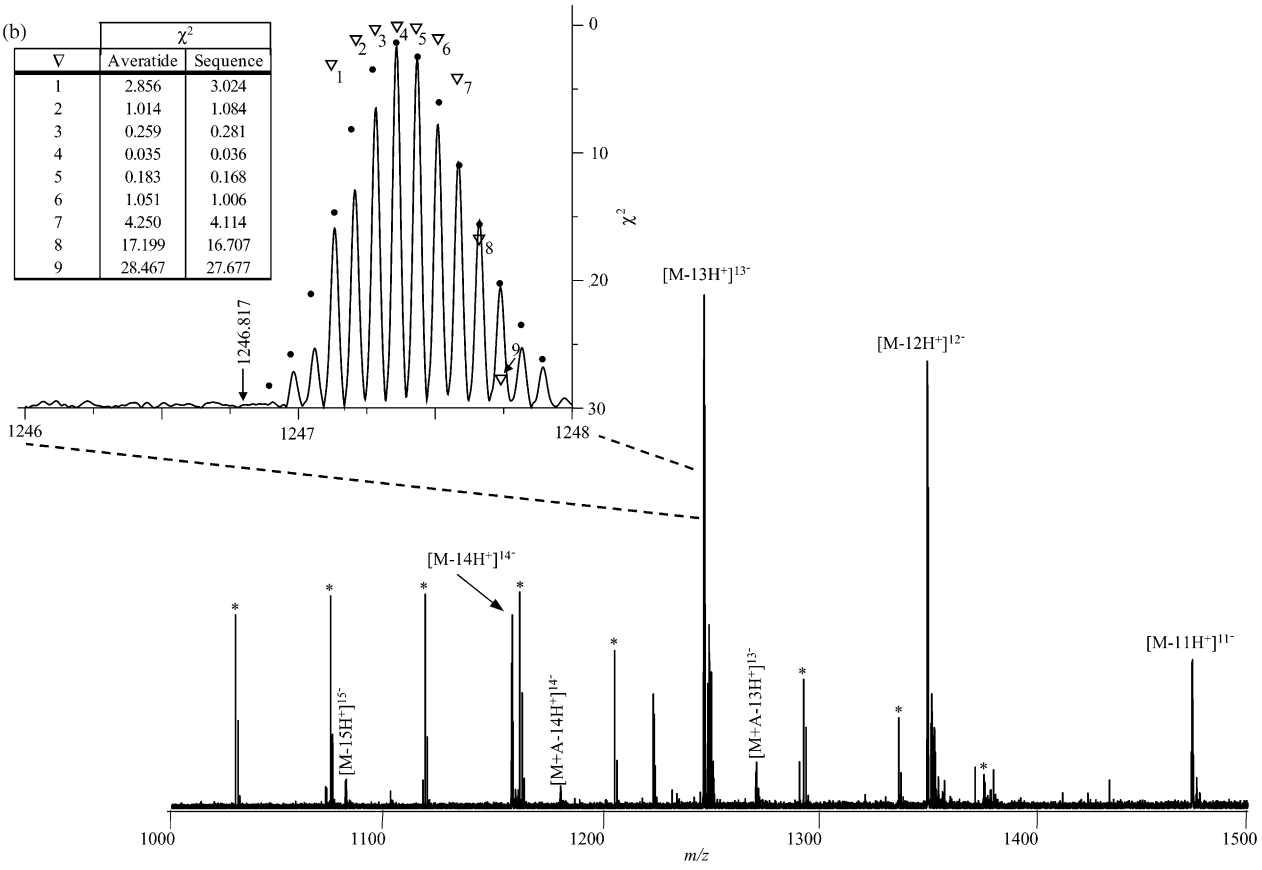

Fig. 1. (a) Sequence of the coding strand of the SNP 22 PCR product is shown. Underlined portion represents the forward primer sequence. Theoretical neutral monoisotopic mass for the wild-type sequence is 16 221.739 Da. (b) ESI FT-ICR mass spectrum of the unmodified coding strand of the SNP 22 PCR product. Charge states are labeled for both the blunt-ended and mono-adenylated products. Asterisks indicate peaks resulting from the internal mass calibrant polyethylene glycol (PEG) 1000. The 13^- charge state is enlarged to compare both the experimental and theoretical isotopic distribution. Solid circles indicate the theoretical isotopic distribution. Triangles represent χ^2 values generated from an algorithm using the *averatide* sequence. Both sets of χ^2 values resulting from the *averatide*- and sequence-based approaches are listed in the table. The experimentally calculated monoisotopic *m*/*z* value is labeled with an arrow.

develop an average nucleotide. The chemical formula of *averatide* was calculated based on the frequencies of AT and GC found in the genome [2], shown in Eq. (1). In this calculation we assumed that the frequency of AT and GC are consistent throughout the genome. It should be noted that CpG islands (regions of the genome with >50% GC content) and repetitive regions in the genome will not be accurately represented by an *averatide* chemical formula thus warranting a locus-specific approach afforded by the specificity inherent in the PCR phase.

$$\begin{aligned}\text{Formula}_{\text{Averatide}} &= 0.58\left(\frac{C_{10}H_{12}N_5O_5P_1 + C_{10}H_{13}N_2O_7P_1}{2}\right) \\ &\quad + 0.42\left(\frac{C_{10}H_{12}N_5O_6P_1 + C_9H_{12}N_3O_6P_1}{2}\right) \qquad (1)\end{aligned}$$

$$\text{Formula}_{\text{Averatide}} = C_{9.79}H_{12.29}N_{3.71}O_6P_1$$

Basically, this calculation is a weighted average of the chemical formulas for the four nucleotides. Presently, only the abundances of the complimentary nucleotides are known (i.e., the percentage of AT and GC). Therefore, the average chemical formula for each pair is weighted according to the frequency of the pair found in the human genome. The first term in Eq. (1) is an average of the chemical formula for deoxyadenosine monophosphate and deoxythymidine monophosphate multiplied by the frequency of the two nucleotides in the human genome, which is 58%. Similarly, the second term is a weighted average of deoxyguanosine monophosphate and deoxycytidine monophosphate. The resulting chemical formula is an average nucleotide found in DNA; due to the structural differences between the nucleobases and sugar moieties, RNA would have a different resulting average nucleotide. Previously, Zubarev and Demirev [44] developed an average nucleotide "*averabaseine*" assuming an equal distribution of the bases.

The average molecular weight of *averatide* is used to calculate the number of nucleotides in an oligonucleotide, which is accomplished by dividing the experimental average molecular weight by the average molecular weight of *averatide*. The average molecular weight of *averatide* is 308.908 Da. Since the *averatide* average molecular weight includes the mass of a phosphate group, the number of nucleotides calculated will not be an integer. However, small differences in experimental mass will not affect the length calculation (data not shown).

The number of nucleotides is then used to determine the model chemical formula of the amplicon, shown in Eq. (2).

$$\text{Formula} = \#\text{Nucleotides}\,(C_{9.79}H_{12.29}N_{3.71}O_6P_1) - (x)PO_2 \quad (2)$$

The variable x represents the number of strands, $x = 1$ for single-stranded DNA and $x = 2$ for double-stranded DNA.

Rounding the number of each element to an integer will result in a mass discrepancy between the theoretical and experimental average molecular weight. Adding or subtracting hydrogens from the model chemical formula corrects for this difference in average molecular weight. Large correction factors are a result of the range of nitrogen and oxygen in the various nucleotides. The number of nitrogen atoms in the four different nucleotides ranges from three to five, which corresponds to a 28 Da difference, and the number of oxygen atoms ranges from five to seven, corresponding to a 32 Da difference.

Once the corrected model chemical formula is obtained, this formula is used to generate a theoretical isotopic distribution. The isotopic fitting procedure then uses this distribution to determine the experimental monoisotopic mass (vide supra).

The inset in Fig. 1b shows the aligned theoretical (solid circles) and experimental isotopic distribution for the wild-type sequence generated using the *averatide* approach. Chi-square values are represented by triangles and are listed in the table. In this case, the most abundant peak in the experimental isotopic distribution is determined by the fitting procedure to correspond to the theoretical most abundant isotope. The experimental monoisotopic m/z value was found to be 1246.817 Da for the 13^- charge state, corresponding to a neutral monoisotopic mass of 16 221.710 Da (mass error: 1.8 ppm).

Applying the isotopic fitting procedure to acquired mass spectra will generate experimental neutral monoisotopic masses. The experimental neutral monoisotopic mass can then be compared to the theoretical neutral monoisotopic mass of the wild-type sequence. If large discrepancies (mass errors ≥ 3 ppm) between theoretical and experimental neutral monoisotopic mass exist, a base composition and length comparison to the wild-type sequence can be used as a preliminary determination regarding the nature of the sequence variation. In cases where the base composition differs from wild type sequence, three possibilities exist: a substitution, insertion, or deletion is present. A comparison of the calculated length to the length of the wild-type sequence will easily distinguish between the three possibilities. No change in length indicates a substitution, for example a single nucleotide polymorphism. Longer or shorter oligonucleotide sequences compared to the wild-type sequence would indicate insertions or deletions, respectively, and identify a need for tandem mass spectrometry.

In order to determine the ability of the algorithm to routinely identify monoisotopic mass, ten sequential internally calibrated mass spectra of the coding strand SNP 22 amplicon were acquired. Isotopic distributions of the 13^- charge state were employed to generate experimental monoisotopic masses with the isotopic fitting procedure using both a sequence-specific and *averatide*-based approach. When monoisotopic masses were generated using the sequence-specific approach the average mass accuracy was found to be 1.8 ± 1.3 ppm (confidence interval of the mean at the 95% confidence level). The *averatide* approach produced an average mass measurement accuracy of 1.8 ± 1.3 ppm (confidence interval of the mean at the 95% confidence level). These results are summarized in Table 1. The consistent mass measurement accuracies obtained from both the *averatide* approach and the sequence-specific approach validates the *averatide* approach. However, a sequence (locus)-specific approach may prove more useful for shorter amplicons or for regions of the genome that are highly invariable (i.e., CpG islands, repeat regions).

Table 1
Comparison of neutral monoisotopic mass measurements for 7-deaza modified and unmodified SNP 22 amplicons generated using a sequence-specific and *averatide* approach

Species	Neutral monoisotopic mass				
	Theoretical	Sequence-specific approach		*Averatide* approach	
		Experimental	Error (ppm)	Experimental	Error (ppm)
Wild-type	16221.739	16221.710[a]	1.8 ± 1.3[b]	16221.710[a]	1.8 ± 1.3[b]
7-Deaza guanine	16217.758	16217.792	−2.1	16217.791	−2.0
7-Deaza adenine	16213.777	16213.812	−2.1	16213.813	−2.2

[a] Average of monoisotopic mass measurements ($n = 10$).
[b] Average mass error ($n = 10$) ± CI of the mean at the 95% confidence level.

3.2. ESI-FT-ICR mass spectrometry of 7-deaza guanine products

After the development of the isotopic fitting algorithm, an application became apparent. The use of 7-deaza modified purines to completely replace unmodified purines during PCR was under investigation in our laboratory. Substitution of a vinylinic carbon for the N-7 nitrogen lowers the mass by one Dalton. A shift in mass from the wild-type sequence can be used to determine the number of guanines or adenines depending on which deaza analog was present during PCR. For the SNP 22 locus amplified in the presence of 7-deaza guanine, the 4-Da mass shift observed indicates there are four guanines incorporated during amplification. This information in conjunction with the primer sequence would restrict the number of possible base compositions. The sequence for the c^7G-modified coding strand of the SNP 22 PCR product is shown in Fig. 2a. Bolded guanines indicate sites where c^7G is incorporated when the template is amplified with c^7dGTP in place of dGTP.

A small mass shift of the 7-deaza modified PCR amplicon from the unmodified product requires high mass measurement accuracy to unequivocally distinguish modified and unmodified sequences from one another. Furthermore, the determination of complete incorporation of modified bases was needed to elucidate optimal PCR conditions. For both of these reasons, we adopted the *averatide* approach described in Section 3.1 to analyze amplicons containing 7-deaza modified purines.

Initial attempts to generate PCR products containing 7-deaza guanine (c^7G) in our laboratory only proved successful after tandem PCR, and even then with poor yield (data not shown). Other researchers have noted that optimal amplification of DNA with *Taq* and dUTP (a modified dNTP) typically requires a higher concentration in comparison to the other dNTPs in the reaction. This is a result of the inefficiency of *Taq* polymerase incorporating modified bases. Applying this observation to the 7-deaza analogs, preliminary PCR optimization included varying the concentration of the modified base. By doubling the concentration of the deaza-modified base in comparison to the other dNTPs, PCR efficiency was dramatically increased. A 2× concentration was chosen as our upper concentration limit based on other researchers' work indicating that DNA could be amplified

(a) 5' GGG GCC ATC AGA GAA ATA TAC CTC TTA TTA CCA CCT **GG**A AAT TAC CAC CCT **GG** 3'

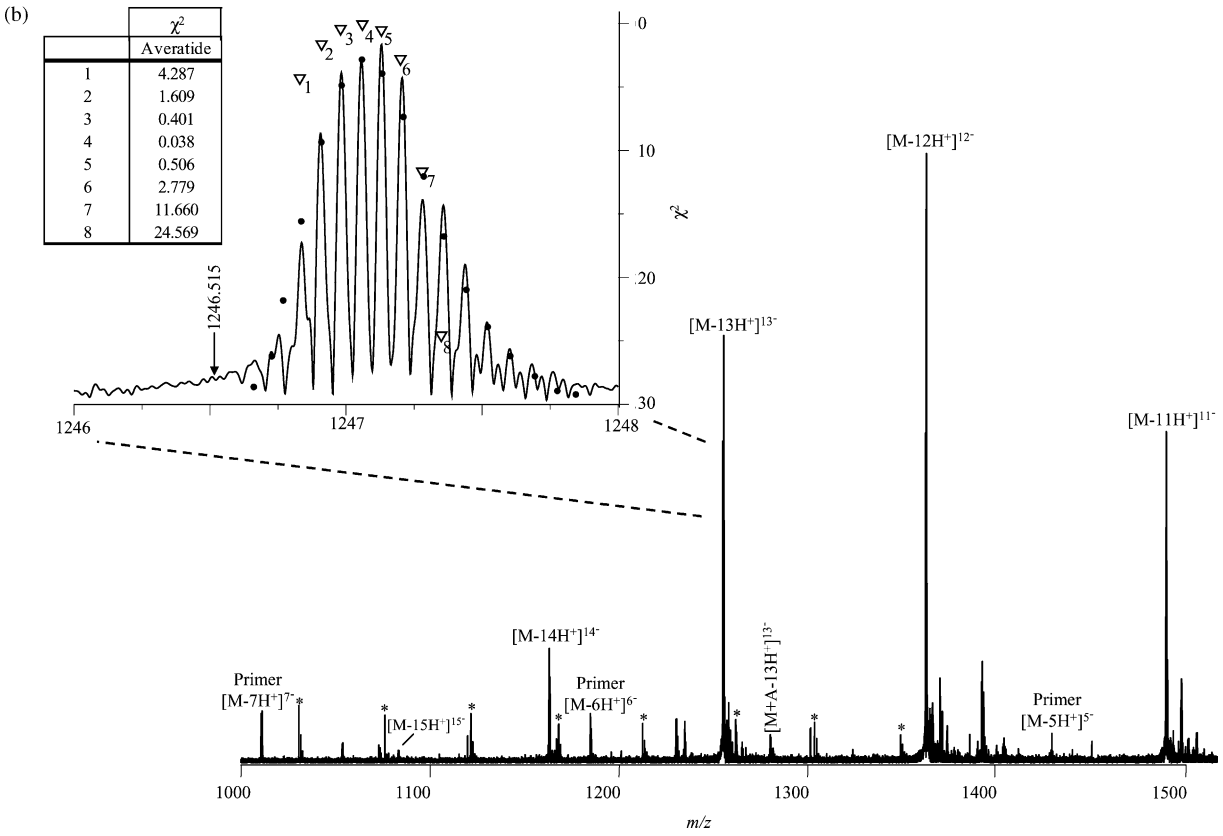

Fig. 2. (a) The c^7G-modified sequence is shown with c^7G sites in bold. The neutral monoisotopic mass is 16 217.758 Da, approximately a 4 Da shift from the unmodified sequence. (b) ESI FT-ICR mass spectrum of the coding strand of the 7-deaza-guanine SNP 22 PCR product. Notation is the same as in Fig. 1b. Some forward primer is present and the resulting peaks in the mass spectrum are labeled with their charge state. This FT-ICR mass spectrum was acquired with 512 k data points and digitized at rate of 1 MHz. An expansion of the 13^- charge state is shown, along with the algorithm output. The algorithm was run using the *averatide* sequence and the resulting χ^2 values are represented by triangles in the inset and listed in the accompanying table. Solid circles rfcommepresent the theoretical isotopic distribution. The monoisotopic *m*/*z* value is labeled with an arrow.

with a 2× concentration of the modified dNTP. Concentrations of 1× and 1.5× were also investigated. Amplicons from the three different reactions were quantified using UV absorbance and qualitatively examined on a 2% agarose gel; both methods indicated that the 2× concentration of the modified dNTP produced the largest amount of amplicon without producing artifacts (data not shown).

Once c^7G was incorporated into the PCR products, we attempted to generate single-stranded products. Efforts to denature double-stranded PCR products containing deaza-modified bases using both heat and solution composition [24] in our laboratory proved unsuccessful (unpublished data). However, we were able to generate single-stranded PCR products containing c^7dGTP by lambda exonuclease digestion. The resulting single-acquisition ESI-FT-ICR mass spectrum of the single-stranded species is shown in Fig. 2b. The double-stranded SNP 22 PCR product was not detected after lambda exonuclease digestion nor were truncated species clearly demonstrating the ability of lambda exonuclease to process the modified bases.

In Fig. 2b, peaks from the internal mass calibrant, PEG-1000, are labeled with asterisks. Internally calibrated data was used to generate an experimental monoisotopic mass for the 13^- charge state using the *averatide*-based chemical formula and the isotopic fitting algorithm. The experimental monoisotopic *m*/*z* value was then used to calculate the experimental neutral monoisotopic mass. Table 1 lists both the theoretical and experimental neutral monoisotopic masses along with the mass measurement accuracies obtained from both the *averatide* and sequence-specific approach. Data for the c^7G product was collected with 512 k data points.

The inset in Fig. 2b shows both the experimental and theoretical (solid circles) isotopic distribution of the 13^- charge state for the c^7G SNP 22 coding strand. Chi-square values are represented with triangles and listed in the table. Note that the most abundant peak in the experimental isotopic distribution does not correspond to the theoretical most abundant isotope. If the most abundant isotope from the experimental data had been used without the isotopic fitting algorithm, a ±1 Da error would have been incurred. However, the algorithm was able to correctly produce the expected monoisotopic neutral mass.

An average error of −2.0 ppm was determined for the 7-deaza guanine modified amplicon, which means the experimental neutral monoisotopic mass differed by −0.03 Da. Since this is less than the change in mass due to the substitution of a single c^7G for an unmodified guanine (i.e., 1 Da), we were confident that complete c^7G incorporation at all four sites in the amplified region occurred.

3.3. ESI-FT-ICR mass spectrometry of 7-deaza adenine products

The coding strand of the SNP 22 locus contains eight potential sites for 7-deaza adenine (c^7A) to be incorporated. This is double the number of potential sites for the c^7G product, so the mass difference between the c^7A modified and unmodified product is about −8 Da. Therefore, less stringent mass measurement accuracies are required to differentiate the c^7A modified and unmodified product. However, several authors have reported that amplification using *Taq* polymerase will not proceed solely in the presence of c^7dATP [45,46], so it was essential that we demonstrate that complete incorporation of c^7A was achieved in absence of dATP. To attain high mass measurement accuracies, the *averatide* approach using the isotopic fitting algorithm was applied to the c^7A system.

The method described in Section 2.2 was used to effectively amplify the SNP 22 locus containing c^7A. The sequence of the c^7A-modified product is shown in Fig. 3a. There are eight c^7A sites, indicated in bold.

Experimental neutral monoisotopic masses were determined from the isotopic fitting algorithm using the *averatide*-based chemical formula, described in Section 3.1, and the 13^- charge state of the single-stranded c^7A SNP 22 product internally calibrated using a C_5 oligonucleotide. Fig. 3b shows the aligned theoretical (solid circles) and experimental isotopic distribution. Chi-square values are indicated by triangles and the monoisotopic *m*/*z* value is denoted with an arrow. Both the experimental and theoretical monoisotopic masses for both the *averatide* and sequence-specific approach are listed in Table 1. Mass measurement accuracies of −2.1 ppm for the sequence-specific approach and −2.2 ppm for the *averatide* approach were obtained, which confirms PCR products containing c^7A were successfully amplified with complete replacement of dATP with c^7dATP.

In addition to amplifying c^7A-modified products, we successfully generated single-stranded species, indicating lambda exonuclease can digest modified bases. Even more interesting is the presence of three adjacent adenines outside of the primer region where c^7A was inserted during the PCR. The contiguous nature of these sites demonstrate that not only does *Taq* polymerase insert the modified bases but also that lambda exonuclease can process multiple, adjacent modifications. These two observations are validated by a mass comparison of the experimental and theoretical neutral monoisotopic masses, which are 16 213.813 and 16 213.777 Da, respectively with a resulting mass error of −2.2 ppm.

Another report indicates that PCR products can be amplified in the sole presence of c^7dATP using exo (−) *Pyrococcus furiosus* (*Pfu*) DNA polymerase [23]. The advantages of exo (−) *Pfu* are the lack of terminal transferase activity, which would result in only blunt-ended species, and a higher fidelity (i.e., lower misincorporation rate). The properties of several DNA polymerases have been examined elsewhere [47]. However, *Taq* polymerase does have a higher processivity (i.e., the number of bases extended before the enzyme falls off the template) and may be more amenable incorporation of 7-deaza purines.

(a) 5' GGG GCC ATC AGA GAA ATA TAC CTC TTA TTA CCA CCT GGA AAT TAC CAC CCT GG 3'

(b)

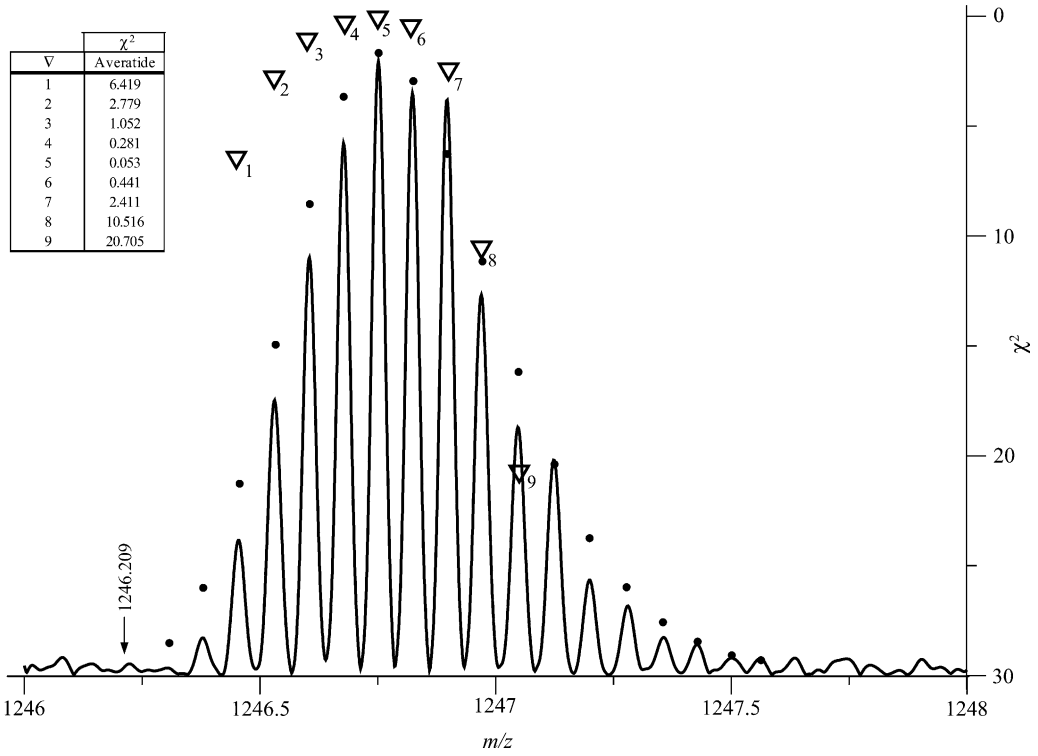

Fig. 3. (a) Sequence of the coding strand for the c^7A SNP 22 PCR product. The primer region is underlined and c^7A sites are in bold. (b) ESI FT-ICR mass spectrum of the coding strand of the 7-deaza-adenosine SNP 22 PCR product. The output of the algorithm using the *averatide* chemical formula is shown. Triangles indicate χ^2 values and numerical values are listed in the table. The lowest χ^2 value indicates the most abundant isotopic peak. The monoisotopic mass is indicated with an arrow. Solid circles correspond to the theoretical isotopic distribution.

After initial experiments with c^7A products were complete, gas-phase dissociation experiments with the excess product were performed using sustained off-resonance irradiation collision-induced dissociation (SORI-CID) [48]. SORI-CID results indicate that cleavage at 7-deaza sites is not observed which is in accord with our previous studies [19]. This approach could be used to reduce fragmentation pathways when sequencing large PCR products. Incorporating 7-deaza analogs in the PCR phase is a simple method to selectively reduce fragmentation from intact PCR amplicons and combined with high mass measurement accuracies and associated algorithms, this overall strategy will greatly improve top-down sequencing using tandem MS [19].

4. Conclusions

Improvements in mass measurement accuracies ultimately aid base composition analysis, which can be used to determine if sequence variation is present. Higher mass measurement accuracies can be achieved by a variety of approaches, which utilize aspects of both molecular biology and mass spectrometry. We have shown internal calibration using a dual-electrospray source of single-stranded species (generated from lambda exonuclease digestion) improves mass accuracy. Additionally, an isotopic fitting procedure that determines monoisotopic mass from isotopically resolved experimental data could further improve mass accuracy and precision. A deviation in base composition when compared with wild-type would warrant tandem mass spectrometry to determine the precise nature and location of the sequence variation. The incorporation of 7-deaza purines during PCR selectively eliminates dissociation pathways, thus reducing the number of product ion channels and improving signal-to-noise. These techniques in combination would improve tandem mass spectrometry experiments and allow gas-phase sequencing of larger amplicons.

Acknowledgements

The authors would like to thank Dr. Alan L. Rockwood (ARUP Institute for Clinical and Experimental Pathology) for the Mercury code. The authors gratefully acknowledge financial support received from the National Institutes of Health (R01HG02159), W.M. Keck Foundation, and Mayo Clinic College of Medicine.

References

[1] International Human Genome Sequencing Consortium, Nature 409 (2001) 860.
[2] J.C. Venter, et al., Science 291 (2001) 1304.
[3] P.F. Crain, J.A. McCloskey, Curr. Opin. Biol. 9 (1998) 25.
[4] D.C. Muddiman, R.D. Smith, Rev. Anal. Chem. 17 (1998) 1.
[5] P.A. Limbach, P.F. Crain, J.A. McCloskey, Curr. Opin. Biotechnol. 6 (1995) 96.
[6] J. Tost, I.G. Gut, Mass Spectrom. Rev. 21 (2002) 388.
[7] L.M. Polo, P.A. Limbach, J. Mass Spectrom. 33 (1998) 1226.
[8] U. Puapaiboon, J. Jai-nhuknan, J.A. Cowan, Anal. Chem. 72 (2000) 3338.
[9] N.R. Isola, S.L. Allman, V.V. Golovlov, Anal. Chem. 71 (1999) 2266.
[10] F. Greco, A. Liguori, G. Sindona, N. Uccella, J. Am. Chem. Soc. 112 (1990) 9092.
[11] K.B. Green-Church, P.A. Limbach, J. Am. Soc. Mass Spectrom. 11 (2000) 24.
[12] S.A. McLuckey, G.J. Vanberkel, G.L. Glish, J. Am. Soc. Mass Spectrom. 3 (1992) 60.
[13] J.W. Flora, J.C. Hannis, D.C. Muddiman, Anal. Chem. 73 (2001) 1247.
[14] J.S. Ni, S.C. Pomerantz, J. Rozenski, Y.H. Zhang, J.A. McCloskey, Anal. Chem. 68 (1996) 1989.
[15] K.X. Wan, M.L. Gross, J. Am. Soc. Mass Spectrom. 12 (2001) 580.
[16] K. Hakansson, R.R. Hudgins, A.G. Marshall, R.A.J. O'Hair, J. Am. Soc. Mass Spectrom. 14 (2003) 23.
[17] K.A. Sannes-Lowery, S.A. Hofstadler, J. Am. Soc. Mass Spectrom. 14 (2003) 825.
[18] D.P. Little, D.J. Aaserud, G.A. Valaskovic, F.W. McLafferty, J. Am. Chem. Soc. 118 (1996) 9352.
[19] J.C. Hannis, D.C. Muddiman, Int. J. Mass Spectrom. 219 (2002) 139.
[20] F. Kirpekar, E. Nordhoff, K. Kristiansen, P. Roepstorff, S. Hahner, F. Hillenkamp, Rapid Commun. Mass Spectrom. 9 (1995) 525.
[21] J. Gross, F. Hillenkamp, K.X. Wan, M.L. Gross, J. Am. Soc. Mass Spectrom. 12 (2001) 180.
[22] K. Schneider, B.T. Chait, Nucleic Acids Res. 23 (1995) 1570.
[23] C.W. Siegert, A. Jacob, K. H, Anal. Biochem. 243 (1996) 55.
[24] J.B. Mangrum, J.W. Flora, D.C. Muddiman, J. Am. Soc. Mass Spectrom. 13 (2002) 232.
[25] J.W. Little, J. Biol. Chem. 242 (1967) 679.
[26] R.G. Higuchi, H. Ochman, Nucleic Acids Res. 17 (1989) 5865.
[27] S. Tagaki, M. Kimura, M. Katsuki, BioTechniques 14 (1993) 218.
[28] A.P. Null, J.C. Hannis, D.C. Muddiman, The Analyst (Special Issue on Biol. Mass Spectrom.) 125 (2000) 619.
[29] D.C. Muddiman, G.A. Anderson, S.A. Hofstadler, R.D. Smith, Anal. Chem. 69 (1997) 1543.
[30] A.I. Nepomuceno, D.C. Muddiman, H.R. Bergen, J.R. Craighead, M.J. Burke, P.E. Caskey, J.A. Allan, Anal. Chem. 75 (2003) 3411.
[31] J.C. Hannis, D.C. Muddiman, Rapid Commun. Mass Spectrom. 15 (2001) 348.
[32] A.P. Null, L.T. George, D.C. Muddiman, J. Am. Soc. Mass Spectrom. 13 (2002) 338.
[33] R.A. Zubarev, P.A. Demirev, P. Hakansson, B.U.R. Sundqvist, Anal. Chem. 67 (1995) 3793.
[34] M.W. Senko, S.C. Beu, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 6 (1995) 229.
[35] A.P. Null, J.C. Hannis, D.C. Muddiman, Anal. Chem. 73 (2001) 4514.
[36] J.T. Stults, J.C. Marsters, Rapid Commun. Mass Spectrom. 5 (1991) 359.
[37] P.A. Limbach, P.F. Crain, J.A. McCloskey, J. Am. Soc. Mass Spectrom. 6 (1995) 27.
[38] C. Liu, D.C. Muddiman, R.D. Smith, J. Mass Spectrom. 32 (1997) 425.
[39] J.C. Hannis, D.C. Muddiman, Rapid Commun. Mass Spectrom. 13 (1999) 323.
[40] J.C. Hannis, D.C. Muddiman, Rapid Commun. Mass Spectrom. 12 (1998) 443.
[41] A.P. Null, A.I. Nepomuceno, D.C. Muddiman, Anal. Chem. 75 (2003) 1331.
[42] A.L. Rockwood, S.L. VanOrden, Anal. Chem. 68 (1996) 2027.
[43] D.M. Horn, R.A. Zubarev, F.W. McLafferty, J. Am. Soc. Mass Spectrom. 11 (2000) 320.
[44] R.A. Zubarev, P.A. Demirev, J. Am. Soc. Mass. Spectrom. 9 (1998) 149.
[45] F. Seela, A. Roling, Nucleic Acids Res. 20 (1992) 55.
[46] I.V. Yang, P.A. Ropp, H.H. Thorp, Anal. Chem. 74 (2002) 347.
[47] L.M. Benson, A.P. Null, D.C. Muddiman, J. Am. Soc. Mass Spectrom. 14 (2003) 601.
[48] J.W. Gauthier, T.R. Trautman, D.B. Jacobson, Anal. Chim. Acta 246 (1991) 211.