Determination of Oligonucleotide Composition from Mass Spectrometrically Measured Molecular Weight

Steven C. Pomerantz, Jeffrey A. Kowalak, and James A. McCloskey
Departments of Medicinal Chemistry and Biochemistry, University of Utah, Salt Lake City, Utah, USA

Extensive calculations for molecular mass versus subunit composition have been made for oligonucleotides from RNA and DNA to determine the extent to which base compositions might be derived from mass spectrometrically determined molecular weights. In the absence of compositional constraints (e.g., any numbers of A, U, G, C), measurement of molecular weight leads to only modest restrictions in allowable number of base compositions; however, if the compositional value for any one residue is known, such as from selective chemical modification or enzymatic cleavage, the number of allowable base compositions becomes unexpectedly low. For example, hydrolysis of RNA by ribonuclease T_1 produces oligonucleotides for which G = 1, for which all base compositions can be uniquely specified up to the 14-mer level, solely by measurement of mass to within ±0.01%. The effects of methylation, phosphorylation state of nucleotide termini, and knowledge of chain length on the determination of subunit composition are discussed. *(J Am Soc Mass Spectrom 1993, 4, 204–209)*

The inference of composition of molecules from measurement of mass is one of the fundamental applications of mass spectrometry to the structural characterization of organic molecules and is perhaps best represented by the determination of elemental composition from exact mass [1, 2]. This approach can be extended to the subunit compositions of complex molecules, but success is limited primarily by the number of different subunit values to be considered in conjunction with the magnitude of the mass measured, which for large values of both can lead to an unmanageable number of compositional possibilities. A study was undertaken of the molecular mass versus subunit composition relationships in oligonucleotides from RNA and DNA to determine the extent to which base compositions might be derived from accurate measurement of oligonucleotide molecular weight. The potential for this approach is based on two favorable factors: the limited number of basic subunits in RNA and DNA considered in the mass-to-composition calculation (four each; Figure 1) and recent advances in methods for production of large gas-phase polynucleotide ions that can be measured by mass spectrometry [3]. The results, described in the following section, show that in the absence of compositional constraints [e.g., any values for A, U (T), G, C], the measurement of molecular weight leads to only modest restrictions in the number of allowable base compositions; however, if the number of any given residue is known, for instance, from experiments involving chemical modifications or selective enzymatic cleavage, the number of allowable compositions becomes unexpectedly low. For example, hydrolysis of RNA by ribonuclease (RNase) T_1, which cleaves preferentially on the 3′ side of G residues, produces oligonucleotides terminating in ...Gp-3′, for which G = 1. The base compositions of all such oligonucleotides, through the 14-mer level, can be uniquely determined by measurement of molecular weight within ±0.01%. Although the principles we describe can be applied to any class of polynucleotides for which selective cleavage or modification can be used, it is particularly advantageous in structural studies of RNA. In such cases, the corresponding gene sequence is often known but cannot be used to establish the presence of structural changes in RNA that result from processing events that occur after transcription, such as splicing or numerous forms of enzymatic modification [4]. The base compositions of all RNase T_1 hydrolysis fragments can be predicted from the gene sequence and compared with compositions determined by mass spectrometry, from which modifications are recognized by mass shifts associated with modification (14.03 u for methyl, etc.) [5]. In the

Portions of this work were presented at the 39th Annual ASMS Conference, Nashville, TN, May 1991.

Address reprint requests to James A. McCloskey, Department of Medicinal Chemistry, Skaggs Hall, University of Utah, Salt Lake City, UT 84112.

1044–0305/93/$6.00

Received August 21, 1992
Revised October 8, 1992
Accepted October 9, 1992

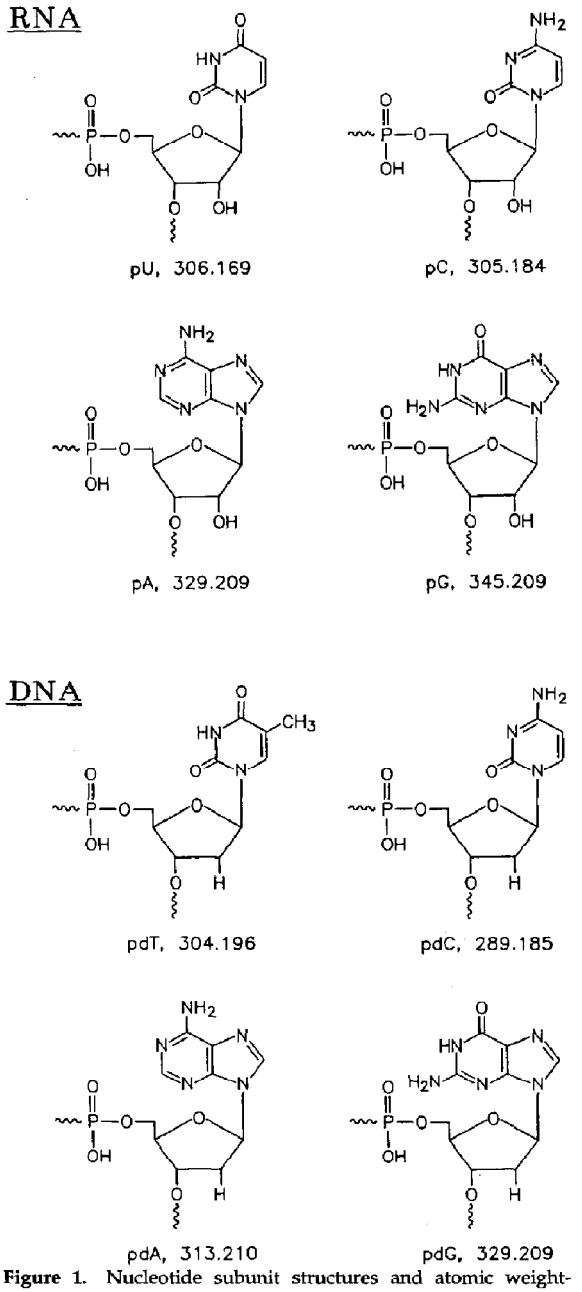

Figure 1. Nucleotide subunit structures and atomic weight-based residue mass values for RNA and DNA.

first application [6] of this approach, the thermostable nucleoside N^4-acetyl-2'-O-methylcytidine (ac^4Cm) was located at position 35 in 5S rRNA of the thermophilic organism *Pyrodictium occultum* by molecular weight measurement of the nonanucleotide in which it occurs, following its unexpected discovery by liquid chromatography/mass spectrometry in the dinucleotide ac^4CmpG. Because the corresponding gene sequence specified only unmodified C at position 35 [7], post-transcriptional modification at this (or any other) site was unanticipated and could not be determined from the DNA-inferred RNA sequence. Details of the full protocol for the detection and sequence locations of modified nucleotides in RNA based on the mass versus composition correlations we describe here will be published separately [8].

General Nucleotide Composition—Mass Correlations

Although there are only four principal constituent residues that need to be considered for either RNA or DNA, the relative molecular mass (M_r) values of these residues are sufficiently similar that isobaric compositions easily arise, even in relatively small oligonucleotides. Figure 2 illustrates the nearly exponential increase in the number of possible nucleotide compositions as a function of mass when base composition is not constrained in any fashion. All 1-u mass intervals (i.e., those within 1 Da) are not necessarily represented by one or more compositions owing to the quantized nature of M_r values of the residues. Every incremental mass value below 4900 Da is not necessarily populated by an allowable composition of an oligonucleotide,

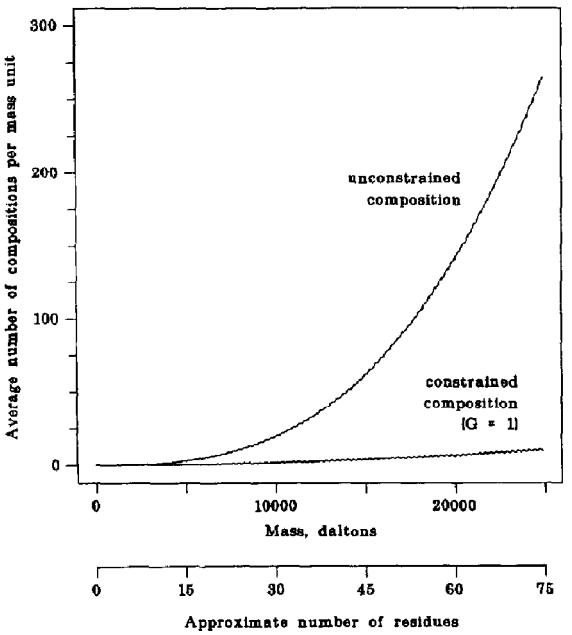

Figure 2. Average number of RNA oligonucleotide compositions per mass unit as a function of chain length. The unconstrained curve has no restrictions on allowable composition of the oligonucleotide chain. Compositions represented in the constrained curve are restricted to having one guanosine residue. Data points were determined by calculating the total number of compositions in each 100-Da interval. Analogous curves for DNA oligonucleotides (not shown) are similar.

although many mass intervals have more than one possible composition. Above 4900 Da, there is at least one composition associated with every mass value, whereas above 6000 Da there are at least two compositions at every 1-u mass interval. Table 1 presents a portion of the data used to generate Figure 2 and provides typical examples of oligonucleotide compositions within four different 1-Da mass windows. It is readily apparent that mass measurement accuracies of ±0.5 Da are not necessarily sufficient to uniquely specify the composition beyond approximately 2500 Da, although there are fortuitous cases up to mass 6000 that have only one allowable composition.

From the standpoint of correlation of nucleotide composition with mass spectrometrically determined molecular weight, accuracy of ±0.01% is considered within current experimental capability for high-quality measurements from oligonucleotides but obviously depends on factors such as sample quantity and type of mass analyzer used. Using the ±0.01% criterion, the listing in Table 2 shows all minimal nucleotide composition pairs that cannot be distinguished below mass 5000 for DNA and RNA. Additional compositions are possible, based on nucleotide extensions of each of the minimal subunits shown. For example, under RNA, the C_2G_3/A_5 pair is equivalent in terms of absolute mass difference to UC_2G_3/UA_5, and so on. Although mass–composition correlations at this level may suffice for some applications, additional constraints are necessary for determination of composition beyond approximately the tetramer level.

Table 1. RNA oligonucleotide compositions within selected mass intervals

Mass interval	Mass[a]	C	U	A	G	Chain length
1000 ± 0.5	None					
2500 ± 0.5	2499.512	7	0	0	1	8
	2500.497	6	1	0	1	8
5000 ± 0.5	4999.867	2	11	2	1	16
	4999.960	1	4	0	10	15
	5000.015	9	3	4	0	16
10000 ± 0.5	9999.687	2	24	3	3	32
	9999.739	0	24	8	0	32
	9999.780	1	17	1	12	31
	9999.783	11	16	0	5	32
	9999.835	9	16	5	2	32
	9999.928	8	9	3	11	31
	9999.931	18	8	2	4	32
	9999.980	6	9	8	8	31
	9999.983	16	8	7	1	32
	10000.021	7	2	1	20	30
	10000.024	17	1	0	13	31
	10000.032	4	9	13	5	31
	10000.073	5	2	6	17	30
	10000.076	15	1	5	10	31
	10000.079	25	0	4	3	32
	10000.084	2	9	18	2	31
	10000.125	3	2	11	14	30
	10000.128	13	1	10	7	31
	10000.131	23	0	9	0	32
	10000.177	1	2	16	11	30
	10000.180	11	1	15	4	31
	10000.232	9	1	20	1	31

[a]Values shown are molecular weights of neutral oligonucleotides with one external phosphate.

Effects of Constraints on Allowable Nucleotide Compositions

Fixed Base Composition

The most effective means of reducing the number of allowable compositions within a given mass range is by fixing the value for any one nucleotide species. The significant result of this form of constraint is shown in Figure 2 for G = 1. In principle, knowledge of composition of one base could come from selective chemical modification or cleavage reactions, but in the case of RNA can be effectively derived from hydrolysis using site-specific ribonucleases. RNase T_1 and U_2 cleave RNA selectively at G [9] and A [10, 11], respectively, to yield oligonucleotides terminating in Gp-3′ or Ap-3′. As shown in Table 3, when G or A are fixed at one residue, *all* oligoribonucleotide compositions can be uniquely specified solely by measurement of mass, to at least the 14-mer level. Although the smallest nucleotides that cannot be distinguished by mass measurement within ±0.01% are C_7U_7G versus $A_{13}G$ when G = 1, the reduction in number of allowable compositions is substantial for larger nucleotides, and many can still be unambiguously defined by mass.

An experimental example of how mass measurement used in conjunction with selective cleavage may be applied is shown by the electrospray mass spectrum in Figure 3. *Escherichia coli* 5S ribosomal RNA (M_r 38,855; 120 nucleotides [12]) was cleaved by RNase T_1 to yield a mixture of nucleotides, each containing one G residue as a Gp-3′ terminus. One of the hydrolysis products was isolated by anion-exchange chromatography and its molecular weight determined by mass spectrometry as 2267.65 (see Figure 3). This value corresponds to the sole nucleotide composition C_2UA_3Gp [M_r calculated (calc.) 2267.39], with the closest alternative compositions C_3A_3Gp (M_r calc. 2266.40) and CU_2A_3Gp (M_r calc. 2268.37). The determined composition corresponds to the 7-mer RNase T_1 fragment predicted from the RNA sequence [12], and the mass spectrum in Figure 3 therefore uniquely identifies the oligonucleotide as 45-5′-AACUCAGp-3′-51.

Chain Length and Phosphorylation State of the Termini

Measurement of molecular mass simply to the nearest integer mass unit permits unambiguous determination of chain length through the 7-mer (DNA) or 8-mer (RNA) level. Examination of the data in Table 2 reveals that above these chain lengths, nearly all of the isobaric pairs represent nucleotides of different chain

Table 2. Isobaric oligonucleotide compositions (within 0.01%) below mass 5000

DNA			RNA		
Residue[a]	Mass	Δm	Residue[a]	Mass	Δm
$d(C_2G_3)$	1566.016	0.052	C_2G_3	1645.994	0.052
$d(A_5)$	1566.068		A_5	1646.046	
$d(C_5G_3)$	2433.584	0.019	U_8G_1	2794.559	0.148
$d(T_8)$	2433.603		C_7A_2	2794.707	
$d(T_8)$	2433.603	0.033	$U_1A_1G_7$	3051.838	0.003
$d(C_3A_5)$	2433.636		C_{10}	3051.841	
$d(A_1G_7)$	2617.707	0.004	C_3U_7	3058.734	0.145
$d(C_8T_1)$	2617.711		A_3G_8	3058.879	
$d(C_{10}T_1)$	3196.090	0.047	$C_1U_7A_2$	3106.784	0.093
$d(A_6G_4)$	3196.137		G_9	3106.877	
$d(G_{10})$	3292.134	0.056	U_8A_3	3436.978	0.096
$d(C_6T_1A_4)$	3292.190		C_9G_2	3437.074	
$d(C_{13})$	3759.457	0.015	C_{12}	3662.209	0.049
$d(T_7A_1G_4)$	3759.472		$U_1A_6G_4$	3662.258	
$d(C_5T_9)$	4183.751	0.028	U_1G_{10}	3758.254	0.055
$d(A_6G_7)$	4183.779		C_8A_4	3758.309	
$d(T_7G_7)$	4433.899	0.037	U_7A_7	4447.645	0.042
$d(C_{11}A_4)$	4433.936		C_1G_{12}	4447.687	
			U_{15}	4592.532	0.241
			C_6G_8	4592.773	

[a]Internal phosphates not shown.

lengths. This is due to relatively purine-rich nucleotides of length n overtaking in mass pyrimidine-rich nucleotides of length $n + 1$. Often, the chain length can be established by other means, such as by anion-exchange chromatography or electrophoretic mobility. In many cases, over the 8-mer level this additional information is sufficient to allow a precise compositional assignment and almost always reduces the number of possible compositions that need be considered. For example, of the 22 possible oligonucleotides at mass 10,000 (Table 1), knowing the length of the oligonucleotide eliminates 50–80% of the possible compositions. In the case of RNase hydrolysis products in which one base value is fixed, knowledge of chain length permits unique assignment of all possible compositions up to at least the 25-mer level (Table 3).

Table 3. Smallest oligoribonucleotides that occur within 0.01% in mass when one base is fixed at one residue

	Residue[a]	Mass	Δm
Chain length unknown			
G = 1, 14 /15-mer	$C_7U_7G_1$	4642.694	0.249
	$A_{13}G_1$	4642.943	
A = 1, 15 /16-mer	$U_{15}A_1$	4939.756	0.241
	$C_6G_8A_1$	4939.997	
Chain length known			
G = 1, 25-mer	$U_{24}G_1$	7711.275	0.391
	$C_{23}A_1G_1$	7711.666	
A = 1, 30-mer	$C_3U_{10}G_{16}A_1$	9487.801	0.985
	$C_2U_{11}G_{16}A_1$	9848.786	

[a]One external phosphate and internal phosphates not shown.

In some instances, uncertainty in the phosphorylation state at either terminus of the oligonucleotide will lead to ambiguities in the conversion of mass to composition. The presence of terminal phosphate can readily be tested for by treatment of the sample by bacterial alkaline phosphatase, which removes terminal phosphates and results in a net mass shift of 80 u per phosphate group. 2′,3′-Cyclic phosphates at the 3-terminus of ribonucleotides cannot be hydrolyzed by phosphatase and so would result in a value 62 u higher than the unphosphorylated nucleotide or 18 u lower than with one phosphate terminus. Cyclic phosphate-containing termini can be opened by treatment with alkali to produce normal phosphates [13]. If the phosphorylation state of the terminus is uncertain, the experimentally determined mass value can be adjusted by addition or subtraction of 80 u, etc., to determine whether reasonable composition candidates will result.

Nucleotide Modification

The effect of any modification that changes the mass of any of the four basic nucleotide residues is generally to increase the number of allowable compositions if the number of such residues present is not known and if the total number of different subunits is greater than

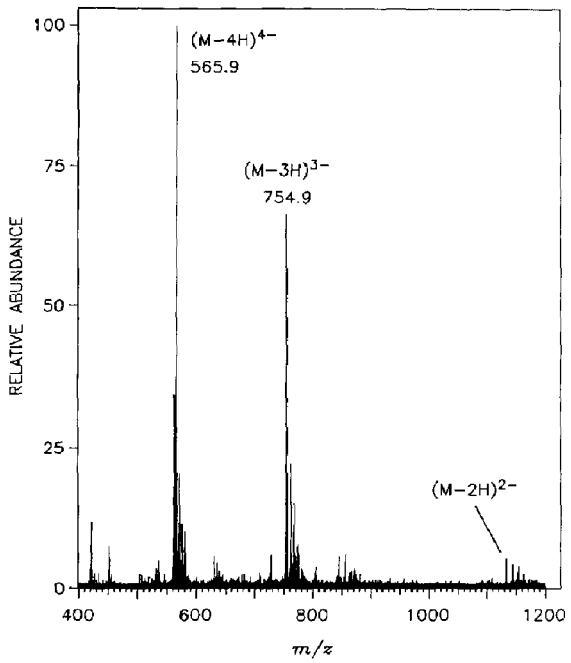

Figure 3. Negative ion electrospray mass spectrum of an oligonucleotide isolated from a RNase T_1 hydrolysate of *E. coli* 5S ribosomal RNA. The measured molecular weight M_r 2267.65, derived from single-scan centroid values shown for ions $[M - 3H]^{3-}$ and $[M - 4H]^{4-}$, uniquely specifies the composition C_2UA_3Gp (M_r calc. 2267.39).

the principal four species. However, as described below, this effect is minimal in the analysis of RNase fragments for detection of methylation, the most frequent form of modification. In any event, if the presence of a modified residue (e.g., methylcytosine) is known by independent means, such as chromatographic analysis of a total hydrolysate, it may serve to restrict the number of allowable compositions in the same sense as restriction of G in Figure 2.

In the case of natural modification, methylation is the single most common form, in both DNA [14] and RNA [15]. To ascertain the effect of methylation on the compositional uniqueness obtainable from a mass measurement, a data set of all possible oligoribonucleotide fragments from ribonuclease T_1 and U_2 digestion up through the 20-mer level was generated. This set of compositions was allowed to contain a maximum of four methyl groups and a maximum of two of the residues that are associated with the normal cleavage site (G for RNase T_1, A for RNase U_2). Two of the normal cleavage site residues G or A were permitted because methylation at the 2′-hydroxyl group prohibits cleavage and would in the case of RNase T_1 thus result in one internal G (e.g., 2′-*O*-methylguanosine) and one terminal G (...Gp-3′). Table 4 summarizes the results of this analysis and lists the compositional ambiguities resulting from the inclusion of methyl modification, all of which are otherwise absent; however, the complete set of mass values that were calculated (approximately 13,000 compositions below mass 7000) shows that even with methylation, the great majority of compositions can be assigned solely from mass measurement. For example, with 0.01% mass accuracy, the presence of one methyl can be unambiguously assigned through the 10-mer level in all RNase T_1 fragments, with no allowable fits for nonmethylated oligonucleotides (i.e., G = 1; A, U, C = any value). In the case of two methyl groups, only one case occurs at the 6-mer level that cannot be resolved by measurement of mass: $C_5Gp + 2CH_2$ (1917.198) versus U_4AGp (1917.108). At the 10-mer level, there are 15 cases in which dimethylated RNase T_1 fragments cannot be distinguished from unmethylated nucleotides (of which there are 55 possible compositional isomers) and no cases in which they cannot be differentiated from monomethylated composition candidates. Of note, if the mass accuracy tolerance is increased to 0.02%, a level which in our experience encompasses virtually all measurements, the number of compositional ambiguities associated with detection of one or two methyl groups through the 14-mer level is essentially unchanged from those at the 0.01% level. In practice [8], ambiguities resulting from potential modification can be readily resolved by an additional experiment involving high-performance liquid chromatographic analysis of the enzymatically hydrolyzed oligomer to determine the presence of methylated (or other) nucleosides [16].

Experimental

Calculations

All software for the calculation of oligonucleotide compositions was developed in the authors' laboratory. Molecular weight values were computed using the 1987 atomic weight data [17]. Computations were performed on a Sun Sparcstation 1 running SunOS 4.0.3 with Sun FORTRAN Version 1.2 or an IBM-compatible 80386DX computer under Microsoft MS-DOS 5.0 with Microsoft FORTRAN Version 5.

RNA Oligonucleotide

Five nanomoles of 5S rRNA was completely hydrolyzed with RNase T_1, as previously described [18]. The resultant oligonucleotide mixture was resolved by anion-exchange chromatography [8], and the oligonucleotide fractions were evaporated to dryness in a SpeedVac centrifuge.

Mass Spectrometry

The oligonucleotide sample analyzed for Figure 3 was prepared in 95% MeOH at a concentration of 5 pmol/μL, assuming quantitative RNA hydrolysis and

Table 4. Oligonucleotide compositions of ribonuclease hydrolysis fragments below mass 3500 within 0.01% of each other

Ribonuclease T_1 fragment			Ribonuclease U_2 fragment		
Residue[a]	Mass	Δm	Residue[a]	Mass	Δm
$U_2A_1G_1Me_2$	1314.809	0.003	G_2A_1	1019.626	0.084
$C_2G_2Me_1$	1314.812		$C_1A_2Me_4$	1019.710	
$U_2G_2Me_1$	1316.782	0.086	$U_2A_2Me_1$	1284.783	0.003
$C_3G_1Me_4$	1316.868		$C_2G_1A_1$	1284.786	
A_4G_1	1662.045	0.032	$U_2G_1A_1$	1286.755	0.087
$C_3G_2Me_4$	1662.077		$C_3A_1Me_3$	1286.842	
$U_4A_1G_1$	1899.093	0.090	$C_8U_1A_1Me_1$	3090.877	0.000
$C_5G_1Me_2$	1899.183		G_8A_1	3090.877	
$U_4G_2Me_2$	1943.146	0.058	$U_{10}A_1Me_2$	3418.951	0.151
$C_2A_3G_1$	1943.204		$C_9A_2Me_1$	3419.102	
$U_6G_1Me_3$	2224.302	0.061			
$C_4A_2G_1$	2224.363				
$U_{10}G_1Me_1$	3420.923	0.151			
$C_9A_1G_1$	3421.074				

[a] One external phosphate and internal phosphates not shown.

chromatographic recovery. The solution was continuously infused into the ion source with a Harvard Instruments (South Natick, MA) syringe pump at 1 μL/min. Full-scan mass spectra were acquired from m/z 400–1200 and averaged over 2 min. Ten picomoles were consumed for the spectrum shown, and 50 pmol of total sample was used for the overall experiment.

The electrospray mass spectrum shown was acquired on a Vestec model 201 (Vestec Corp., Houston, TX) quadrupole mass spectrometer (2000 m/z range) fitted with a Vestec electrospray ion source and a 10-kV postacceleration detector. Mass measurements were made on peak centroids derived from single scans acquired in the calibration mode of a Teknivent (St. Louis, MO) Vector/One data system.

Acknowledgments

This work was supported by grant GM 29812 from the National Institute of General Medical Sciences. The authors are grateful to Dr. Peter B. Moore for a sample of *E. coli* 5S RNA.

References

1. Beynon, J. H. *Mass Spectrometry and its Applications to Organic Chemistry*; Elsevier: New York, 1960; Chapter 8 and Appendix 1.
2. Biemann, K. *Methods Enzymol.* **1990**, *193*, 295–305.
3. (a) Smith, R. D.; Loo, J. A.; Edmonds, C. G.; Barinaga, C. J.; Udseth, H. R. *Anal. Chem.* **1990**, *62*, 882–899; (b) Standing, K. G.; Ens, W., Eds. *Methods and Mechanisms for Producing Ions from Large Molecules*; Plenum: New York, 1991; (c) for review and further references, see McCloskey, J. A.; Crain, P. F. *Int. J. Mass Spectrom. Ion Processes* **1992**, *118 / 119*, 593–615.
4. Apirion, D., Ed. *Processing of RNA*; CRC Press: Boca Raton, FL, 1984.
5. Kowalak, J. A.; Pomerantz, S. C.; McCloskey, J. A. *Proceedings of the 40th ASMS Conference on Mass Spectrometry and Allied Topics*; Washington, DC, May 31–June 5, 1992; pp. 1127–1128.
6. Bruenger, E.; Kowalak, J. A.; Kuchino, Y.; McCloskey, J. A.; Mizushima, H.; Stetter, K. O.; Crain, P. F. *FASEB J.* **1993**, in press.
7. Kaine, B. P.; Schurke, C. M.; Stetter, K. O. *System. Appl. Microbiol.* **1989**, *12*, 8–14.
8. Kowalak, J. A.; Pomerantz, S. C.; Crain, P. F.; McCloskey, J. A., in preparation.
9. Sato, K.; Egami, F. *J. Biochem.* **1957**, *44*, 753–767.
10. Arima, T.; Uchida, T.; Egami, F. *Biochem. J.* **1968**, *106*, 609–613.
11. Donis-Keller, H.; Maxam, A. M.; Gilbert, W. *Nucleic Acids Res.* **1977**, *4*, 2527–2538.
12. Liebke, H.; Hatfull, G. *Nucleic Acids Res.* **1985**, *13*, 5515–5525.
13. Uchida, T.; Egami, F. In *Procedures in Nucleic Acid Research*, Vol. 1; Cantoni, G. L.; Davies, D. R., Eds.; Harper and Row: New York, 1966.
14. Fasman, G. D., Ed. *Handbook of Biochemistry and Molecular Biology*, Vol. 2, 3rd ed.; CRC Press, Cleveland, OH, 1976; pp. 873–874.
15. Björk, G. R. In *Processing of RNA*; Apirion, D., Ed.; CRC Press, Boca Raton, FL, 1984; pp. 291–330.
16. Buck, M.; Connick, M.; Ames, B. N. *Anal. Biochem.* **1983**, *129*, 1–13.
17. Atomic weights of the elements 1987. *Pure Appl. Chem.*, **1988**, *60*, 841–854.
18. Brownlee, G. G. In *Laboratory Techniques in Biochemistry and Molecular Biology*, Vol. 3; Work, T. S.; Work, S., Eds.; North Holland: Amsterdam, 1972.