Deconvolution of Combinatorial Oligonucleotide Libraries by Electrospray Ionization Tandem Mass Spectrometry

Steven C. Pomerantz,[†] James A. McCloskey,*[,†,‡] Theodore M. Tarasow,[§] and Bruce E. Eaton*[,§]

Contribution from the Departments of Medicinal Chemistry and Biochemistry, University of Utah, Salt Lake City, Utah 84112, and NeXstar Pharmaceuticals, Inc., 2860 Wilderness Place, Boulder, Colorado 80301

Received November 7, 1996[⊗]

Abstract: Studies were undertaken to explore the application of tandem mass spectrometry for the structure analysis of unfractionated mixtures of oligonucleotides. Limited combinatorial libraries were constructed of mixtures of 8-mers (**N**GACAC**N**G; nine compounds) and 12-mers (**N**GACT**N**AGAC**N**G; 27 compounds), where **N** is any of the 2′-deoxyribonucleotides of uracil, thymine, or 5-[*N*-(aminoethyl)-3-acrylimido]uracil. Molecular mass measurements of the mixture components (single mass analyzer) in the simplest cases or acquisition of collision-induced dissociation mass spectra (tandem mass analyzers) for mixtures of sequence isomers of the same molecular mass were used to establish guidelines for structure assignments. Although the construction of mass sequencing ladders from gas phase backbone cleavage reactions is notably more complex in the case of isomeric mixtures than for the single isomer case, assignment ambiguities are reduced by the presence of nonrandomized sequence positions and by recognition of likely fragment ion relative abundances. The approaches described offer a significant advance toward solving problems common to the development of oligonucleotide therapeutics using combinatorial synthesis techniques.

Introduction

Combinatorial chemistry for the discovery of new drugs has a major potential to greatly expand pharmacophore repertoires. However, these advances are accompanied by significant problems in the separation and structural characterization of active compounds in the combinatorial pool, a challenge for which approaches based on mass spectrometry are well suited (reviewed in ref 1). Chemical libraries can be assembled by a variety of techniques, including parallel synthesis, split synthesis, and mixed synthesis, performed in solution or, more typically, on solid-phase beads or pins.[2] Small molecule and oligomeric molecular libraries have been shown to produce "lead compounds" both as linear polymers[3] and as angular arrays.[4]

With respect to oligonucleotide libraries, the *s*ystematic *e*volution of *l*igands by *ex*ponential enrichment (SELEX) overcomes many of the selection and deconvolution problems associated with other combinatorial strategies. SELEX has the advantage that vast sequence libraries (ca. 10^{15}) can be selected, amplified, and reselected in an iterative process.[5] The ability to amplify allows for the screening and deconvolution of considerably larger libraries than can be surveyed using other combinatorial chemistries. Although obstacles to the use of oligonucleotides as therapeutics have been recognized, the reported affinity and specificity of SELEX-derived oligonucleotide therapeutic aptamers[5e] is remarkable, and a source of motivation to improve the pharmacokinetics and economics of oligonucleotide aptamers.[6] A general strategy to achieve these ends has been sequence truncation to render oligonucleotides economically feasible to synthesize, and incorporation of 2′-modified nucleotides that make aptamers sufficiently stable *in vivo* to be efficacious.[7] One possible approach to the derivation of smaller, high affinity aptamers, might include post-SELEX combinatorial exploration of the local shape space about the truncated oligonucleotide using nucleotides chemically modified in the base (pyrimidines and purines), ribose, or phosphate backbone.

As with any other combinatorial chemistry approach, a serious limitation to the implementation of a post-SELEX aptamer modification strategy is the partitioning and structural characterization of the most potent leads from a complex mixture. Conventional techniques to elucidate the composition and sequence of randomly modified oligonucleotide lead compounds have numerous problems and limitations that severely restrict the size and complexity of the libraries. For example, cloning and sequencing techniques are useless since the polymerases employed cannot distinguish between modified and unmodified nucleotides. Consequently, all modified bases would be translated by the polymerase into unmodified oligonucleotides, resulting in the loss of modified base identities and locations. An analytical method is required that is capable of discerning modified oligonucleotides with identical sequences (here sequence refers to U, C, G, A where any of the bases or sugars can be modified) in which the location of multiple modifications may be clearly identified.

* Address correspondence to J. A. McCloskey, 311A Skaggs Hall, University of Utah, Salt Lake City, UT 84112. Tel: (801) 581-5581. Fax: (801) 581-7457. E-mail: james.mccloskey@m.cc.utah.edu.

[†] Department of Medicinal Chemistry, University of Utah.

[‡] Department of Biochemistry, University of Utah.

[§] NeXstar Pharmaceuticals.

[⊗] Abstract published in *Advance ACS Abstracts,* April 1, 1997.

(1) Burlingame, A. L.; Boyd, R. K.; Gaskell, S. J. *Anal. Chem.* **1996**, *68*, 599R−651R.

(2) For a recent review on combinatorial chemistry see: Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. *J. Med. Chem.* **1994**, *37*, 1385.

(3) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P.; Gordon, E. M. *J. Med. Chem.* **1994**, *37*, 1233.

(4) Gordon, E. M. *Curr. Opin. Biotech.* **1995**, *6*, 624.

(5) (a) Lorch, J. R.; Szostak, J. W. *Acc. Chem. Res.* **1996**, 29. (b) Gold, L.; Polisky, B.; Uhlenbeck, O.; Yarus, M. *Ann. Rev. Biochem.* **1995**, *64*, 763. (c) Gold, L. *J. Biol. Chem.* **1995**, *270*, 13581. (d) Gold, L.; Tuerk, C.: Allen, P.; Binkley, J.; Brown, D.; Green, L.; MacDougal, S.; Schneider, D.; Tasset, D.; Eddy, S. In *The RNA World*; Gesteland, R., Atkins, J., Eds.; Cold Spring Harbor Laboratory Press: Plainview, NY, 1993; Chapter 19. (e) Lorch, J. R.; Szostak, J. W. *Biochemistry* **1994**, *33*, 973.

(6) Eaton, B. E.; Gold, L.; Zichi, D. A. *Chem. Biol.* **1995**, *2*, 633.

(7) Crouch, G. J.; Eaton, B. E. *Nucleosides Nucleotides* **1994**, *13*, 939.

S0002-7863(96)03875-9 CCC: $14.00

Table 1. Conditions for Acquisition of Mass Spectra

Figure	scan range (*m*/*z*)/no. of scans/dwell time (ms)/acquisition time (min)	precursor *m*/*z*/charge/collision energy (eV)[a]	sample concn[b] (μM)	sample consumptn[c] (pmol)
1	300−1400/5/3/2.8		5.6/50	215
2	50−1400/10/10/23	797.7/3-/30	5.6/50	189/1700
3	400−1400/10/3/5.0		3.7/100	773
4	50−1600/25/5/32	625.3/6-/78	0.74/20	36.3/982
5	50−1700/50/5/69	934.9/4-/40	3.7/100	390/10535
6	50−1800/25/5/37	934.9/4-/60	10/20	739/1477

[a] Laboratory frame of reference. [b] The first value is the concentration of an individual isomer assuming an equimolar mixture of the components; the second value is the total analyte concentration. [c] Single value represents total consumption. For two values, the first is the consumption of an individual isomer assuming an equimolar mixture of components; the second value is total consumption.

We report here an approach based on tandem mass spectrometry to the problem of sequence analysis of mixtures of oligonucleotides that are closely related (both in mass and sequence), without prior physical separation. The limiting case in terms of difficulty is that in which unresolved sequence isomers (same molecular masses) are present, resulting in added complexity in the interpretation of the resulting mass spectra.[8] The present study was made independent of the possible use of partitioning steps, which in most cases would reduce the difficulty of the problem. As a starting point, recently developed algorithms[9] were tested that had been designed for construction of sequencing "mass ladders" from mass spectra in which a given molecular mass represents only one, rather than multiple, sequence isomers. The extension of those principles to combinatorial oligonucleotide mixtures was undertaken to assess the potential for direct characterization of individual components and to indicate algorithm changes in the interpretation process needed to deal with this significantly more complex problem.

Experimental Section

Oligonucleotide Synthesis. All oligodeoxynucleotides were synthesized on an ABI 392 DNA synthesizer, dimethoxytrityl (DMT)-protected, using standard cycles and conditions except where indicated. Synthesis reagents, including the phosphoramidites of deoxyuridine (U), 5-methyldeoxyuridine (T), or 5-[*N*-(aminoethyl)-3-acrylimido]deoxyuridine (X), were obtained from Glen Research (Sterling, VA). The combinatorial library was prepared by using a mixture of all three phosphoramidites listed above at the appropriate steps of the synthesis for incorporation of variable residues, N. The convention adopted throughout this report will be to refer to specific oligonucleotides by the sequence order (5′ → 3′) of the variable positions. Thus, the oligonucleotide **T**GACT**U**AGAC**X**G (variable positions in bold) is abbreviated as TUX. All syntheses were performed on a 1 μmol scale using PerSeptive Biosearch 500 controlled pore glass (CPG). Oligonucleotide cleavage from the CPG and deprotection were carried out in concentrated NH_4OH at 55 °C for 12−16 h. All oligonucleotides were HPLC purified using a Rainin Dynamax HPLC instrument and Vydac C_4 reversed phase HPLC column, with triethylammonium acetate (TEAA) pH 7.0 and acetonitrile (CH_3CN) as eluents. A semipreparative Vydac column (1.0 × 25 cm, 3 mL/min) was used to purify the 8-mer and 12-mer pools, while a Vydac preparative column (2.2 × 25 cm, 9 mL/min) was used to purify the individual 12-mer oligonucleotides (TUX and XUT). DMT-protected oligonucleotides were purified using a gradient of 15−40% CH_3CN over 20 min. Fractions were collected and concentrated as indicated below. The DMT groups were removed by treating the oligonucleotides with 80% acetic acid at ambient temperature for 1 h, after which the samples were concentrated to dryness under reduced pressure. The fully deprotected oligonucleotides were HPLC purified using a 5−30% CH_3CN gradient over 20 min. Fractions were collected and concentrated as indicated below. Additional counter-ion exchange was performed as necessary by passing the HPLC-purified oligonucleotides through a Nap 5 column (Pharmacia) using 50 mM triethylammonium bicarbonate (TEAB) pH 7.5 as the eluent. UV spectroscopy was performed on a Milton Roy Spectronic Genesys 5 UV/visible spectrophotometer.

NGACACNG. The 8-mer pool was synthesized using an equimolar mixture of U, T, and X phosphoramidites, each at 33 mM in anhydrous acetonitrile, at each variable position, N. HPLC purification of the DMT-protected material yielded major peaks between 11.0 and 13.6 min. These fractions were collected, pooled, and lyophilized to dryness, and the DMT protecting group was removed. HPLC purification of the resulting material yielded one peak at 9.0 min which was collected and concentrated to dryness under reduced pressure. Purified oligonucleotide (293 nmole) was obtained as determined by UV absorbance ($\epsilon_{260} = 8.7 \times 10^4\ M^{-1}\ cm^{-1}$).

NGACTNAGACNG. The 12-mer pool was synthesized using a 1:1:3 molar ratio of U, T, and X phosphoramidites, respectively, at each N position. Phosphoramidite concentrations were 20 mM for U and T and 60 mM for X in anhydrous acetonitrile. The crude DMT-protected oligonucleotides were HPLC purified, with all fractions eluting between 11.7 and 15.0 min collected and pooled. After concentration to dryness and removal of the DMT protecting group the oligonucleotide mixture was HPLC purified again. Fractions which eluted between 9.7 and 12.6 min were collected, pooled, and concentrated to dryness. Purified 12-mer oligonucleotide pool (235 nmol) was obtained as determined by UV absorbance ($\epsilon_{260} = 1.32 \times 10^5\ M^{-1}\ cm^{-1}$).

TGACTUAGACXG. The TUX sequence isomer was synthesized using standard conditions as indicated in the general experimental procedure section. HPLC purification of the DMT-protected oligonucleotide produced a single peak with a retention time of 15.3 min. The oligonucleotide was collected and concentrated to dryness, and the DMT group was removed. HPLC purification of the fully deprotected molecule gave a single major peak at 10.3 min. Collection of this fraction, concentration under reduced pressure, and quantitation by UV absorbance ($\epsilon_{260} = 1.32 \times 10^5\ M^{-1}\ cm^{-1}$) yielded 128 nmol of the desired oligonucleotide.

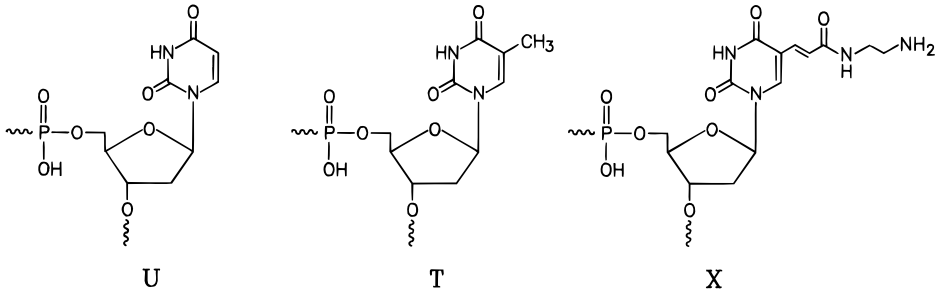
XGACTUAGACTG. The XUT oligonucleotide was synthesized using standard conditions as indicated in the general experimental procedure section. A single major peak was observed at 11.9 min upon HPLC purification of the DMT-protected molecule. Collection of this fraction, concentration to dryness, deprotection and HPLC purification produced a single peak with a retention time 10.5 min. Isolation and UV quantitation yielded 124 nmol of the desired 12-mer.

Mass Spectrometry. All mass spectra were recorded in the negative ion mode on a Sciex (Concord, Ontario, Canada) API III+ triple quadrupole mass spectrometer, with electrospray ionization. Scans were made in the MCA mode with 0.1 *m*/*z* steps. Concentrated aqueous solutions of the combinatorial libraries were diluted with water and methanol to yield a final solvent composition of 5:95 or 10:90 H_2O:CH_3OH. Samples were infused at 1.5−2.0 μL/min via a Harvard Apparatus (Cambridge, MA) Model 22 syringe pump into the electrospray ion source. The gas for collision-induced dissociation (CID) experiments was argon at a thickness of $\sim 2.5 \times 10^{15}$ atoms/cm^2. Resolution in the third quadrupole was adjusted to provide unambiguous charge state identification of fragment ions bearing as many as three charges by examination of the isotopic profile. The other pertinent experimental parameters are summarized in Table 1. In order to obtain high quality mass spectra, no attempt was made to minimize sample consumption or data acquisition time. Spectral interpretation of all

(8) This problem has direct analogy in the sequence analysis of complex oligonucleotide mixtures produced by enzymatic digestion of nucleic acids, for determination of modification sites. For example, hydrolysis of *E. coli* 23S ribosomal RNA (2904 nucleotides) by RNase T_1 is predicted from the corresponding gene sequence to liberate 46 6-mers, which are represented by only 15 unique molecular masses.

(9) Ni, J.; Pomerantz, S. C.; Rozenski, J.; Zhang, Y.; McCloskey, J. A. *Anal. Chem.* **1996**, *68*, 1989.

Chart 1



data was facilitated by comparison of data system-generated lists of experimental *m/z* vs ion abundance with precomputed values for expected fragment *m/z* values, for any assumed sequence (software developed in-house).

Results and Discussion

The analysis of combinatorial oligonucleotide mixtures by mass spectrometry involves two levels of difficulty. In the simplest case, every molecular ion in the mass spectrum of the mixture represents a single sequence, so that components of the mixture can in principle be individually sequenced by sequential selection of each molecular ion by the first mass analyzer MS-1, to produce a CID mass spectrum of each component in MS-2, e.g., refs 9−13. In the more complex situation the ion selected represents multiple sequences having the same molecular mass, leading to partially overlapping mass spectra recorded by MS-2. The interpretation process used for isomeric mixtures was based on algorithms developed for sequencing of single oligonucleotides[9] and on observations of fragment ion abundance patterns from earlier studies of model compounds.[10,14,15] Two combinatorial mixtures were investigated: an 8-mer library containing nine compounds and a 12-mer library containing 27 compounds. Variable nucleotide positions in both libraries were occupied by any of three 2′-deoxyribose pyrimidine derivatives: uracil (U), thymine (5-methyluracil) (T), or 5-[*N*-(aminoethyl)-3-acrylimido]uracil (X) (Chart 1). Selection of these structural moieties was based on several factors. Typically, the smallest unit that would require differentiation in any post-SELEX protocol would be 14 mass units, equivalent to one methylene (e.g., uracil vs thymine), since any non-hydrogen addition would be greater. The third modified uridine was chosen because it contains many functionalities that may be desirable in post SELEX modification. The amine and unsaturated side chain resemble reversible covalent cross-linking nucleosides amenable to post-SELEX protocols. In addition, the primary amine offers an opportunity to observe the mass spectral consequences of having zwitterionic nucleosides. Finally, X is a vinylogous derivative of C5-modified pyrimidines,[7,16] a rich source of combinatorial nucleotide building blocks. Electron deficient pyrimidine rings in T and X may alter the amount of base loss at cleavage sites in collision-induced dissociations, an important variable in the sequencing reactions.

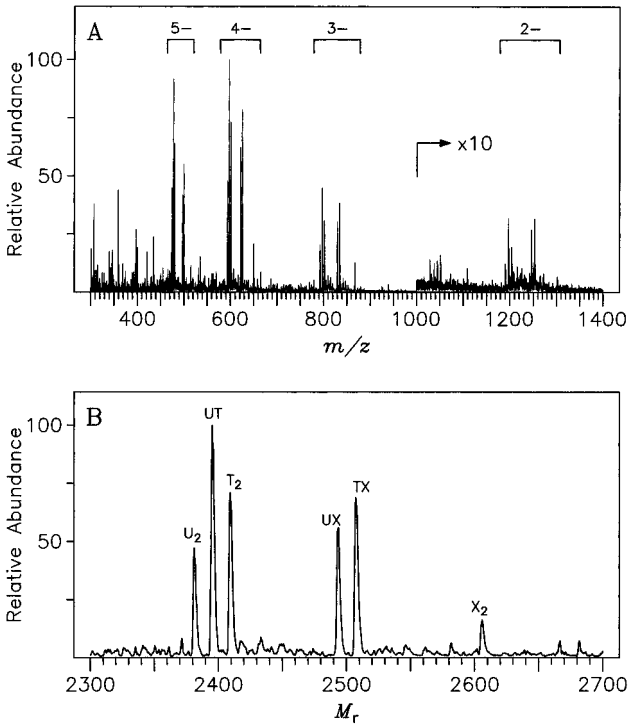


Figure 1. Mass spectrum of the 8-mer combinatorial library NGACACNG composed of nine oligonucleotides: (A) primary mass spectrum, with charge state regions indicated and (B) data from panel A transformed to a molecular mass scale. Peak assignments indicate identities of N residues.

8-mer Oligonucleotide Mixture. Initial experiments were carried out on the eight residue sequence **N**GACAC**N**G, containing two randomized positions N. The 8-mer library consisted of nine unique sequences, three pairs of which had identical base composition and molecular weights. The primary and mass-transformed spectra of the 8-mer combinatorial mixture are shown in Figure 1A,B, respectively. Table 2 summarizes the experimental mass measurement data from the primary spectrum, from which compositional assignments were made. The mass spectra in Figure 1 clearly indicate that at least one isomer of each of the expected sequences is present in the mixture. The primary spectrum unequivocally identifies only three of the nine sequences, UU, TT, and XX, because their masses are unique.

(10) McLuckey, S. A.; Van Berkel, G. J.; Glish, G. L. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 60.
(11) Barry, J. P.; Vouros, P.; Van Schepdael, A.; Law, S.-J. *J. Mass Spectrom.* **1995**, *30*, 993.
(12) Gentil, E.; Banoub, J. *J. Mass Spectrom.* **1996**, *31*, 83.
(13) Boschenok, J.; Sheil, M. M. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 144.
(14) McLuckey, S. A.; Habibi-Goudarzi, S. *J. Am. Chem. Soc.* **1993**, *115*, 12085.
(15) Crain, P. F.; Gregson, J. M.; McCloskey, J. A.; Nelson, C. C.; Peltier, J. M.; Phillips, D. R.; Pomerantz, S. C.; Reddy, D. M. In *Mass Spectrometry in the Biological Sciences;* Burlingame, A. L., Carr, S. A., Eds.; Humana Press: Clifton, NJ, 1996; pp 497−517.
(16) (a) Dewey, T. M.; Zyzniewski, C.; Eaton, B. E. *Nucleosides Nucleotides* **1996**, *15*, 1611. (b) Eaton, B. E.; Pieken, W. A. *Annu. Rev. Biochem.* **1995**, *64*, 837. (c) Dewey, T. M.; Mundt, A.; Crouch, G. J.; Zyzniewski, M. C.; Eaton, B. E. *J. Am. Chem. Soc.* **1995**, *117*, 8474.
(17) McLuckey, S. A.; Habibi-Goudarzi, S. *J. Am. Soc. Mass. Spectrom.* **1994**, *5*, 740.
(18) (a) McGee, D. P. C.; Sebesta, D. P.; O'Rourke, S. S.; Martinez, R. L.; Jung, M. E.; Pieken, W. A. *Tetrahedron Lett.* **1996**, *37*, 1995. (b) Sebesta, D. P.; O'Rourke, S. S.; Martinez, R. L.; Pieken, W. A.; McGee, D. P. C. *Tetrahedron* **1996**, *52*, 14385. (c) McGee, D. P. C.; Vaughn-Settle, A.; Vargeese, C.; Zhai, Y. *J. Organic Chem.* **1996**, *61*, 781. (d) Tu, C.; Keane, C.; Eaton, B. E. *Nucleosides Nucleotides* **1995**, *14*, 1631.

Table 2. Summary of Experimental Results from Primary Mass Spectra

composition	M_r calcd[a]	M_r exptl[b]	error (Da/%)	expected no. of isomers
NGACAC**N**G				
U_2	2381.58	2381.23 ± 0.16	0.35/0.015	1
U_1T_1	2395.61	2395.45 ± 0.42	0.16/0.007	2
T_2	2409.63	2409.30 ± 0.24	0.33/0.014	1
U_1X_1	2493.71	2493.73 ± 0.52	0.02/0.001	2
T_1X_1	2507.75	2507.60 ± 0.61	0.15/0.006	2
X_2	2605.84	2605.93 ± 0.73	0.09/0.003	1
NGACT**N**AGAC**N**G				
U_3	3618.37	3617.89 ± 0.29	0.48/0.013	1
U_2T_1	3632.39	3632.06 ± 0.47	0.33/0.009	3
U_1T_2	3646.42	3646.19 ± 0.43	0.23/0.006	3
T_3	3660.45	3659.81 ± 0.26	0.64/0.017	1
U_2X_1	3730.50	3730.04 ± 0.28	0.46/0.012	3
$U_1T_1X_1$	3744.52	3744.16 ± 0.25	0.36/0.010	6
T_2X_1	3758.55	3758.14 ± 0.29	0.41/0.011	3
U_1X_2	3842.63	3842.46 ± 0.05	0.17/0.004	3
T_1X_2	3856.66	3856.19 ± 0.50	0.47/0.012	3
X_3	3954.76	3954.44 ± 0.67	0.32/0.008	1

[a] M_r values are based on average chemical mass values (e.g., c = 12.011). [b] M_r ± standard deviation, derived from molecular mass measurement using four values of z.

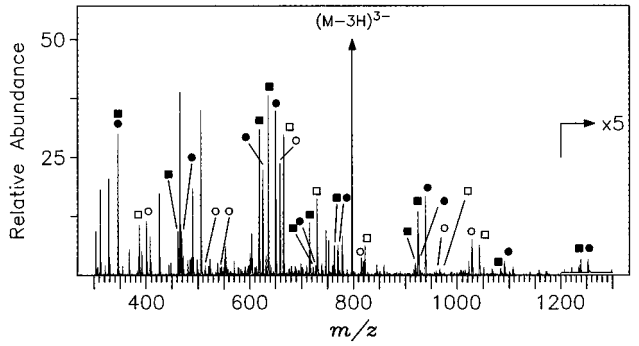


Figure 2. Product ion spectrum of two isomeric 8-mer sequences **U**GACAC**T**G (w series ions: ●; a-Base series ions: ○) and **T**GACAC**U**G (w series ions: ■; a-Base series ions: □).

Mass spectra resulting from collisional dissociation of the $(M\text{-}3H)^{3-}$ ion from each of the six molecular species represented in Figure 1B were acquired. In all nine cases complete sequences could be reconstructed using the w ion series (3′ → 5′ direction) and a-Base (5′ → 3′) series ions (see ref 10 for nomenclature), with the occasional exception of the a_8-Gua ion, since loss of the 3′ terminal base is somewhat disfavored.[14] Most n-mer fragments for $n = 3-6$ were observed at multiple charge states. A representative CID mass spectrum is shown in Figure 2 for the pair of sequence isomers **T**GACAC**U**G (TU) and **U**GACAC**T**G (UT), obtained by selection and dissociation of the m/z 797.7 ion (see Figure 1A). Besides the sequence-diagnostic ions mentioned above, the other ions[10] in Figure 2 (unlabeled) represent b, d, and y ion series and products of simple dehydration of the principal cleavage products (w, y, b, and d series ions) as well as products of double backbone cleavages and fragments which have subsequently lost a base moiety. With only a few exceptions, all fragment ions greater than ∼1% relative abundance in the nearly two dozen CID spectra examined could be assigned to one of the ion types listed above. The low m/z region of CID mass spectra (up to about m/z 400) is dominated by nonspecific ions which bear limited sequence information, including monoanions of the bases and mononucleotides.[15] Although not utilized in the present study, the characteristic base m/z values can serve to qualitatively indicate presence of the modified base in the oligonucleotide.[11,17] With the exception of the w_1 ion (m/z 346) representing ...G-3′ termini in both compounds, there are no w or a-B ions which

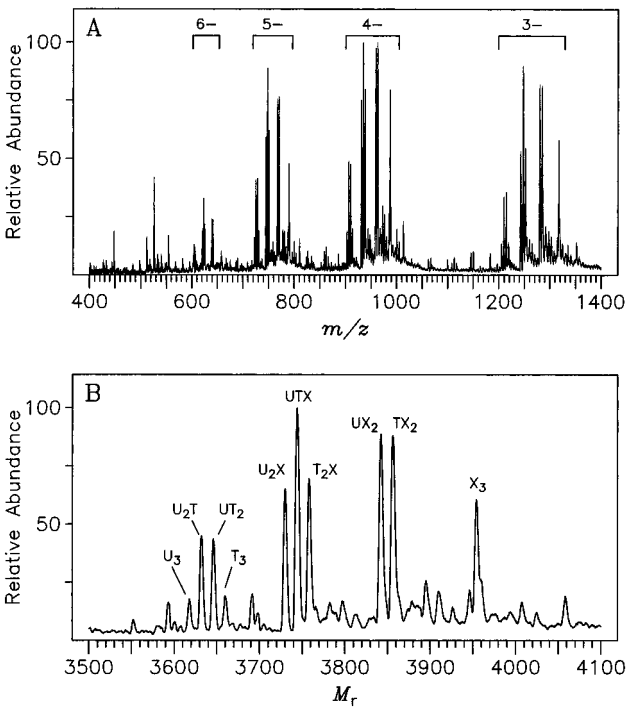


Figure 3. Mass spectrum of the 12-mer combinatorial library NGACTNAGACNG composed of 27 oligonucleotides: (A) primary mass spectrum, with charge state regions indicated and (B) data from panel A transformed to a molecular mass scale. Peak assignments indicate identities of N residues.

are common to both sequences and the spectrum can, in general, be readily interpreted. However, some mass ambiguities were encountered. For example, the prominent ion of m/z 730.0 could be assigned as either the $(a_3\text{-Ade})^{1-}$ ion (m/z 729.95) from the TU sequence or the y_5^{2-} ion (m/z 729.96) from the UT sequence. Close examination of the isotopic profile of this ion does not reveal any evidence for the UT doubly-charged ion, and so assignment was made to the a_3-B species as shown. In those instances where potential fragment ions are isobaric and of the same charge state, the ion is assigned to a principal series ion reflecting the fact that principal series ions are generally observed in much higher abundance than other ion types.[10,15]

12-mer Oligonucleotide Mixture. The complexity of the system was increased in the 12-residue pool, **N**GACT**N**AGAC**N**G, in both chain length and the number of variable positions, N. It was reasoned that a 12-mer would be of appropriate length to represent one-third to one-half of the typical SELEX-derived aptamer. For example, a 36-mer cleaved into three pieces of equal combinatorial complexity (Ω) would give Ω^3 full length modified aptamers. To make the size of the combinatorial library useful for lead enhancement, three positions within each oligomer could be varied with three different modified bases. In this case each truncated oligomer would give a library of 27 possibilities. The full length 36-mer would consist of a library with 19 683 members. A model oligonucleotide library was prepared consisting of three modified pyrimidines (U, T, and X), randomly substituted at three positions within the 12-mer. The mass spectral interpretive process was made more difficult because the 5′ and 3′ ends (NG) were made identical.[9] This second generation library is considerably more complex than the first. The 12-mer library contained 27 individual sequences and only 10 unique molecular masses. Thus, a great deal less definitive information can be gained from the primary spectrum compared with the 8-mer mixture. For example, a single molecular ion peak from the 12-mer library could contain anywhere from one to six different sequences in the most degenerate case, while for the 8-mer

Table 3. Ions Used for Identification of Six Isomeric Sequences (Figure 5)

ion type	*m/z*	interfering ions
$(a_4\text{-Cyt})^{1-}$	1028.9 (UTX, UXT)	none
	1043.0 (TUX, TXU)	$(y_7\text{-}H_2O)^{2-}$ (XUT, XTU); $(d_3\text{-}H_2O)^{2-}$ (XTU)
	1141.1 (XUT, XTU)	w_7^{2-} (XUT, XTU)
w_3^{1-}	925.0 (TXU, XTU)	$(y_9\text{-}H_2O)^{3-}$ (TUX); w_9^{3-} (XUT)
	939.0 (UXT, XUT)	b_9^{3-} (UXT, XUT); b_6^{2-} (XTU); $(y_3\text{-}H_2O)^{1-}$ (UTX, TUX)
	1036.9 (UTX, TUX)	none
$(a_7\text{-Ade})^{2-}$	962.5 (UTX, TUX)	w_9^{3-} (UTX, UXT); $(d_6\text{-}H_2O)^{2-}$ (UXT, XUT)
	1011.5 (UXT, XUT)	none
	1018.5 (TXU, XTU)	none

library any molecular peak could not represent more than two different sequences. The primary and molecular mass-transformed spectra of the combinatorial mixture are shown in Figure 3A,B, respectively. All ten allowable compositions were detected, based on measured molecular mass values (Table 2). It may be noted that the proportion of X phosphoramidite was increased for the 12-mer synthesis relative to the 8-mer synthesis, and this difference is reflected in the mass spectra in Figures 1B and 3B. The mass-transformed spectrum contains several peaks in addition to the six predicted peaks. These principally represent salt-adducted species, both sodium and potassium, as well as combinations thereof. For example, the peak at M_r 3691.8 in Figure 3B is assigned as the oligonucleotide 12-mer consisting of 2 U and 1 T adducted with one sodium and one potassium. In general, such peaks can be recognized, because the allowable molecular mass values are constrained by integral combinations of the residue masses of N.

Analysis of Three Isobaric Oligonucleotides. With the preliminary analysis of the mixture concluded and the homologous substitution (TTT, etc.) sequence isomers structurally confirmed by ES-MS/MS, the analysis of the isobaric components was carried out. A representative spectrum is shown in Figure 4 for the trio of sequence isomers **T**GACT**T**AGAC**X**G (TTX), **T**GACT**X**AGAC**T**G (TXT), and **X**GACT**T**AGAC**T**G (XTT) obtained by selection and dissociation of the $(M\text{-}6H)^{6-}$ ion. Unlike the binary isomer systems of the 8-mer library, the observation of a single ion is not necessarily sufficient for positive identification of a particular isomer in a *ternary* mixture. Only those ions which contain at least two of the variable bases can be potentially assigned to a specific isomer; however, this is not always a sufficient condition. Consider the ion at *m/z* 969.6 in Figure 4. The possible assignments are $(a_7\text{-Ade})^{2-}$ from TTX, which contains two variable positions, and $(d_6\text{-}H_2O)^{2-}$ from both TXT and XTT, which only contains one variable position. Differentiation of these ions is not possible on the basis of charge state identification. The relative abundance of the *m/z* 969.6 ion leads to the assignment of $(a_7\text{-Ade})^{2-}$ from TTX as the most likely interpretation but possibly with a small contribution from the dehydrated d_6 fragment, since d and d-H_2O ions are not typically observed at this relative abundance.[15] In general, within the design context of the library, at least three ions should be observed to deduce the sequence when three isomers are plausible: two ions which contain only one of the variable positions and one ion which contains two of the variable positions.

Analysis of Six Isobaric Oligonucleotides. The complexity of the interpretive task increases considerably as the number of possible sequence isomers increases. Figure 5 shows the mass spectrum derived from dissociation of the $(M\text{-}4H)^{4-}$ precursor from the six sequence isomers UTX, UXT, TXU, TUX, XUT, and XTU. Any fragment ion which contains only one of the

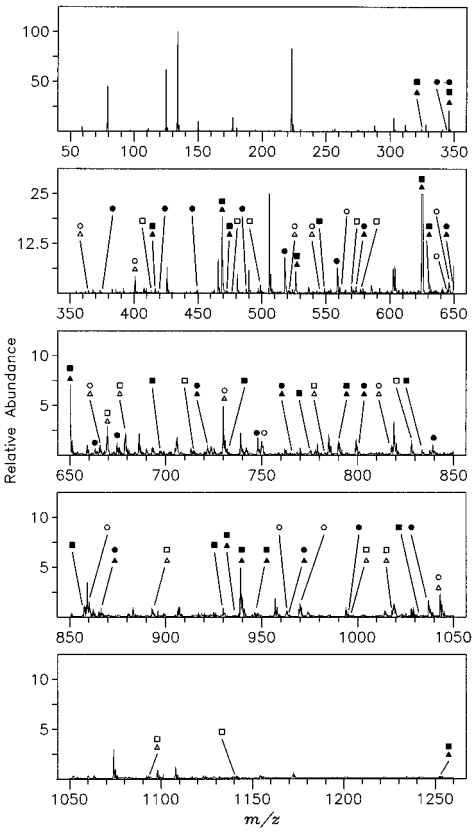


Figure 4. Product ion mass spectrum of three unfractionated isomeric 12-mers of the same molecular mass (3758.1), resulting from dissociation of $(M\text{-}6H)^{6-}$ ions of *m/z* 625.3. TGACT**T**AGAC**X**G (TTX) (w series ions: ●; a-Base series ions: ○), **T**GACT**X**AGAC**T**G (TXT) (w series ions: ▲; a-Base series ions: △), and **X**GACT**T**AGAC**T**G (XTT) (w series ions: ■; a-Base series ions: □).

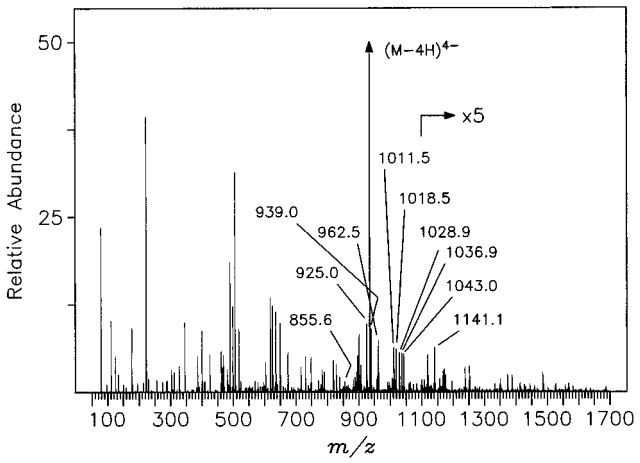


Figure 5. Product ion mass spectrum representing six unfractionated isomeric 12-mer sequences of the same molecular mass (3744.2), containing one residue each of T, U, and X. Assignments for annotated ions are listed in Table 3.

variable positions must be a member of at least two isomeric sequences. However, unlike the ternary systems, even fragments which bear two variable positions are not always unambiguously interpretable. As an example, the ion at *m/z* 855.6 may be assigned to five of the six possible sequences; each of these assignments contains two variable positions, and none of the possibilities may be excluded on the basis of charge state: $(a_9\text{-Ade})^{3-}$ from UTX and TUX, $(d_8\text{-}H_2O)^{3-}$ from UXT and XUT, and $(w_8\text{-}H_2O)^{3-}$ from TXU. In general, priority in assignment can be made on the basis of the likelihood of the dissociation process represented, based on earlier studies of model oligonucleotides.[9,10,14,15] Thus the primary assignment for the *m/z*

855.6 ion is $(a_9\text{-Ade})^{3-}$, since the loss of adenine at the backbone cleavage site is strongly favored.[14] Secondarily, contributions to the overall abundance of this peak are expected[14,15] to come from $(w_8\text{-}H_2O)^{3-}$ and to a lesser extent from $(d_8\text{-}H_2O)^{3-}$.

It is not possible to unambiguously assign the CID spectrum of Figure 5 in a manner that permits the identification of all six isomers conclusively. As previously noted, any ion arising from a single backbone cleavage must of neccessity be assigned to two sequence isomers. Within the context of the isomeric sequences, there are only three criteria which may be applied to the CID spectrum: two independent criteria arising from initial placement of two of the three variable positions and one dependent criterion which specifies the final variable position and is derived from the molecular mass measurement. Thus, without chemical prefractionation procedures, the system is underspecified, and the outcome would provide identification of three of the six isomers. This is still valuable information, and judicious selection of fragment ions is required to maximize the confidence of potential sequence assignments. *A priori*, reasonable choices for these ions would be a_2-Gua to indicate the 5′ variable base, w_2 to elucidate the 3′ variable base, and either a_7-Ade or w_7 to delineate the middle variable. These are, in fact, nonoptimal selections due to the inability to discern structural information arising from isobaric interferences, as follows. The a_2-Gua ion is an "Npf" ion, where f is the furan-like moiety formed at the 3′-terminus of an a-B ion.[14] Within the resolution capabilities of a quadrupole mass spectrometer, this ion is indistinguishable from the corresponding pNp ion ($\Delta m = 0.107$ u). While it has been previously noted that the w_1 ion (indicating the 3′ terminus) is always the most abundant of the monomeric ions regardless of the phosphorlyation state of the 3′ terminus, it is not always the case that the a_2-B ion (indicating the 5′ terminus) is the most abundant of the Npf or pNp ions.[9] Similarly, the w_2 ion (...NG-3′) cannot be distinguished from the d_2 ion (5′-NG...) when the variable residues are identical. While w_n series ions are in general much more abundant than d_n series ions for low values of n, it would not be possible to determine whether the species observed is the d_2 ion from an isomeric sequence present in high molar excess or the w_2 ion from an isomer that was not as well represented in the combinatorial mixture, on the basis of abundance alone when dealing with complex mixtures. Nevertheless, from consideration of expected ion abundance patterns,[14,15] fragment ions can be predicted that satisfy the established criteria as shown in Table 3. While the ions selected for this purpose are not without ambiguities on the basis of m/z, the interpretation is fairly straightforward on the basis of either charge state differences or expected relative abundances between the interferences and the characteristic principal series ions. However, as stated previously, the maximum information that can be derived from this data is the presence of three of the six possible isomers, where the subset must be represented by the presence of each of the variable residues at each of the variable positions. One particular set of isomers that satisfies these requirements is the set that contains the isomers UTX, TXU, and XUT.

Using the ES-MS/MS approach established above, identities of oligonucleotides in the 27 member library present were significantly constrained. Although this represents a considerable analytical accomplishment, data analysis was guided by the expectation that all 27 possible oligonucleotides were present in the mixture. Similar information would not be available during an actual combinatorial post-SELEX aptamer optimization. Identification of selected library members would be required without knowledge of which members were present or their relative abundances. In essence, this would constitute the analysis of an unknown mixture.

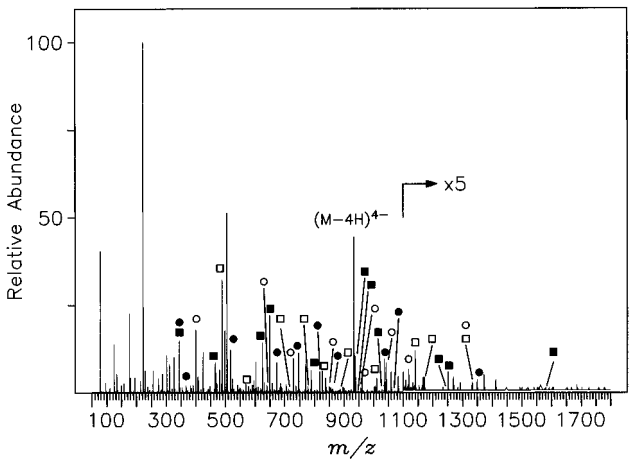


Figure 6. Product ion mass spectrum of "unknown" 12-mer binary mixture. Assignments: w series ions ● and a-Base series ions ○ for **T**GACT**U**AGAC**X**G; w series ions ■ and a-Base series ions □ for **X**GACT**U**AGAC**T**G.

Assignment of a Mixture of Two Unknown Isobaric Oligonucleotides. To evaluate the feasibility of tandem mass spectrometry as a tool for the complete identification of sequence isomers, an experiment was designed to emulate a less than favorable case in which two isomers have an equal affinity for the selection target. It was of interest to determine if an unknown mixture of two of the 27 possible molecules could be assigned unambiguously. It was reasoned that the most challenging sequences in the 12-mer mixture were the six isobaric molecules with one U, T, and X substitution each. As discussed earlier, the primary spectrum can significantly narrow the possibilities when identifying the components of a mixture. In the case of the 12-mer library, ten unique molecular masses were possible. The absence of any ion *from the primary spectrum* would eliminate all sequences corresponding to that molecular mass. The utility of the primary spectrum would obviously vary depending on the degree of degeneracy in the isobaric compositions represented. In the case of the 12-mer mixture, the absence of one, three, or six sequences could be concluded based on the absence of a single molecular mass value. ES-MS/MS sequencing would be required if the primary mass spectrum indicated compositions other than homologous substitution (TTT, etc.). It was reasoned that it would be particularly difficult to identify two sequences from the six isobaric species in which the 5′ and 3′ variable bases were switched leaving the same base in the center, and the 5′/3′ NG ambiguity would need to be reconciled. Since this experiment was intended to test the limits of the technique, a mixture of two oligonucleotides, TUX and XUT, was prepared in one of our laboratories; mass spectrometry experiments were performed blind in the second laboratory, with the knowledge that only two components of structure **N**GACT**N**AGAC**N**G were present. The observation of only one molecular mass, corresponding to the sequences containing one each of U, T, and X in the primary spectrum significantly narrowed the sequence possibilities from 27 to six.

The two component mixture whose CID mass spectrum is shown in Figure 6 was then manually assigned considering the six possible sequence isomers. The single sequence that best accounted for the data was that of the isomer XUT. This sequence was characterized by a complete set of w ions from w_1 to w_9 (and possibly $w_{11}\text{-}H_2O$), with the exception of the w_6 ion. Sequencing from the other terminus was accomplished via the a-B series, which was completely represented from bases 2 through 10, again with the exception of base 6. The lack of ions at position 6 was presumed to be a consequence of the difficulty of losing the uracil base within the sequence context

of the oligonucleotide and was certainly a result of using a relatively low charge state for the precursor ion. The low charge state was employed to mitigate the interpretative task by lessening the number of potential isobaric interferences which cannot be resolved by identification of product ion charge state. However, although the use of higher charge states, 5- and 6-, resulted in increased assignment ambiguities, complete bidirectional sequence ions were formed (data not shown).

The relatively abundant fragment ion at *m/z* 748.1 was the entry point for the sequence deduction of the second component. This ion was assignable as the XG dinucleotide with one external phosphate (pXpG, XpGp, pGpX, or GpXp). This ion eliminates X from the central variable residue and placed the X at either position 1 or position 11. X at position 1 would require XpGp, a d-series ion which is not generally seen in such high abundance. X at position 11 requires pXpG which is the w_2 ion, usually observed in high abundance.[14,15] The fragment ion at *m/z* 730.1 was used to place T at the 5′ terminus. Although this ion could be interpreted as $(w_2\text{-}H_2O)^{1-}$, the relative abundance, compared to w_2^{1-}, (*m/z* 748.1:730.1; abundance 1.15:1) is too large for this assignment. Typically, w:w-H_2O ion pairs are generally more in the range of 5−10:1.[15] However, if the base at position 1 was T, the 730.1 ion could also be interpreted as $(a_3\text{-Ade})^{1-}$, a favored cleavage, consistent with the abundance of this ion. This would account for the reduced w:w-H_2O ratio observed. This interpretation was further confirmed by the presence of the $(a_4\text{-Cyt})^{1-}$ ion at 1043.0, for which no assignment can be attributed to a simple cleavage from a sequence with U in the first position. Thus, the mixture composition was correctly deduced as **X**GACT**U**AGAC**T**G (XUT) and **T**GACT**U**AGAC**X**G (TUX).

The utility of low charge state parent ion selection has been previously emphasized; however, it is not without its attendant drawbacks. The use of low charge state precursors often has the undesirable effect of precluding backbone scission and product ion formation in the middle of longer oligonucleotides. The consequences of this effect would usually prevent mass ladder extension[9] and successful sequencing of the oligonucleotide. In the present case, the core sequence of the oligonucleotide is constrained, and the goal of the analysis is the correct assignment of the randomized nucleotides. When the identity of one or more interior positions may be fixed or otherwise constrained, they can be considered as invariant residue masses in the mass ladder extension process. For example, in the spectrum of the "unknown" sequence XUT, no ions were observed for the a_6-B cleavage, yet ions were observed for the seventh through tenth elements of this ion series. If the sixth residue were to be specified, then a partial sequence could be built that extended through the tenth position of the oligonucleotide, providing additional confirmation of the sequence. Although this feature is not currently implemented in the current version of the sequencing software, it is planned for inclusion.

Conclusions

The sequence elucidation of unfractionated, isobaric mixtures of oligonucleotides when the number of components is unknown presents a substantial analytical challenge. General guidelines have been developed as a result of the present study for the interpretation of product ion mass spectra of such oligonucleotides mixtures. Once a putative sequence has been determined using the w and a-B series ions in concert with the mass ladder extension techniques, assignment of the remaining ions in the spectrum to the other principal series ions (b, d, and y and the simple dehydrations of b, d, y, and w) should be made. The next step is the assignment of double backbone cuts and principal series ions that have lost a base moiety. Ions representing c, x, and z series are generally not observed with regularity or with high relative abundance; however, their existence should not be precluded. Any ions which remain unassigned above some arbitrary threshold should now be examined in the context of alternative sequences. While this threshold is variable and somewhat contingent upon experimental conditions (precursor ion charge, collision energy, etc.), an approximate level would be on the order of 3−5% relative abundance. Often the automated sequencing program[9] will generate lower ranked full-length sequences and/or partial sequences from both termini which will provide potential alternative sequences to be explored.

These guidelines are intended for the analysis of mixtures where the components are present in roughly equal proportion. When one component is present in high molar excess over other components of the mixture, the use of relative abundances for the identification of the minor constituents is compromised. In such cases the magnitude of the problem may be both recognized and reduced through prefractionation steps. Work is currently underway to develop rules for the prediction of ion abundance as a function of several parameters, including collision energy, charge state, position within the sequence, and sequence context (identity of adjacent bases). These rules would ultimately be embodied in algorithms and implemented within the automated sequencing program[9] to enable computerized identification of complex mixtures of oligonucleotides.

The ability to identify unique oligonucleotides in complex mixtures using ES-MS/MS provides a powerful tool for structural deconvolution. The work presented here, while preliminary, establishes a benchmark for the size of oligonucleotide libraries that can be analyzed. Post-SELEX combinatorial library syntheses could now be designed, where after selection for enhanced binding, longer oligonucleotides may be fragmented to a size amenable to ES-MS/MS deconvolution. Because of the vectoral nature of oligonucleotides, such an approach provides access to libraries with populations that equal the product of the fragment complexities. In addition, library diversity is not limited by the substitution of a single nucleotide base. Considerably larger, more complex mixtures can be assembled by including modified uridines, cytidines, adenosines, and guanosines. Clearly, library complexity is not a limiting factor since more than 80 modifications are easily accomplished by streamlined synthetic methods.[6,18] Rather, the design and deconvolution of these potentially enormous libraries will be the challenge. The application of ES-MS/MS for the identification of structurally complex post-SELEX modified aptamers is in progress.

Acknowledgment. This work was supported by Grant GM21584 from the National Institute of General Medical Sciences.

JA963875F