APPLICATION NOTE

SOS: A Simple Interactive Program for ab initio Oligonucleotide Sequencing by Mass Spectrometry

Jef Rozenski
Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium, and Department of Medicinal Chemistry, University of Utah, Salt Lake City, Utah, USA

James A. McCloskey
Departments of Medicinal Chemistry and Biochemistry, University of Utah, Salt Lake City, Utah, USA

Mass spectra of oligonucleotides derived from collision-induced dissociation following electrospray ionization provide an effective means of sequence determination, at the 20-mer level and below. An interactive, stand-alone computer program, Simple Oligonucleotide Sequencer (SOS) has been developed for rapid oligonucleotide sequencing from mass spectra, under user control on a residue by residue basis. Modifications can be defined in any combination for the base, sugar or backbone. Sequence ladders can be independently constructed in both the 5′ → 3′ directions and 3′ → 5′ directions, and graphically compared for homology and overlap. A particular advantage of this method is the ability to easily erase and rebuild alternate subsequences. The program can be used for ab initio sequencing of modified or unmodified oligonucleotides, for rapid verification of sequence, and in studies of fragmentation processes of model oligonucleotide derivatives. (J Am Soc Mass Spectrom 2002, 13, 200–203) © 2002 American Society for Mass Spectrometry

One of the principal advantages of mass-based sequencing of oligonucleotides [1, 2] is that it is generally applicable to a range of synthetic and natural modifications, for which conventional sequencing methods are poorly suited. Mass spectra derived from electrospray ionization and collision-induced dissociation (CID) are well-suited for direct sequencing in the 20-mer size range and below [3], in part due to the complexity, and thus wealth of structural information, of the spectra and the potential applicability to oligonucleotide mixtures, via LC/MS/MS.

The ion dissociation chemistry of polycharged oligonucleotides, on which the production of sequence-relevant fragment ions rests, was initially discovered and elucidated by McLuckey and his collaborators [4, 5], and has since been refined and extended in a number of directions using a variety of instrument configurations [6] most often in conjunction with electrospray ionization. The central approach to the interpretation of data [4, 7] requires the assignment of ions resulting from single cleavages of the backbone, generated in part by base loss at the chain cleavage site [5]. The *m/z* values of these ions permit mass ladders to be constructed independently from each terminus, moving 3' → 5' (e.g., w and y ion series) and 5' → 3' (a–B and d–H_2O series) (for ion series nomenclature see [4]).

Under favorable conditions, e.g., chain length ≤10, absence of unusual modifications, and mass spectra of good quality, sequencing by this method is rapid and straightforward, particularly in cases of sequence verification in which the structure is presumably known in advance (see additional comments below). However, in a number of practical circumstances the correct interpretation of data depends on a number of variables, for example:

- Recognition of unexpected effects on fragmentation that arise from modification or unusual sequence context, for example in the case of backbone or base substitution motifs that have not been previously examined.
- Difficulty in assessing the possibility of alternate mass ladders (i.e., sequences) derived from a given data set (for example in RNAs containing runs of U and C in which fragment ions are distinguished by the C versus U difference of 0.985/*z*).

Published online December 17, 2001

Address reprint requests to either Dr. J. A. McCloskey, Department of Medicinal Chemistry, University of Utah, 30 S. 2000 East, Room 311A, Salt Lake City, UT 84112-1115, USA. E-mail: james.mccloskey@m.cc.utah.edu, or Dr. J. Rozenski, Rega Institute, Katholieke Universiteit Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium. E-mail: jef.rozenski@rega.kuleuven.ac.be

1044-0305/02/$20.00
PII S1044-0305(01)00354-3

Received August 24, 2001
Revised November 9, 2001
Accepted November 9, 2001

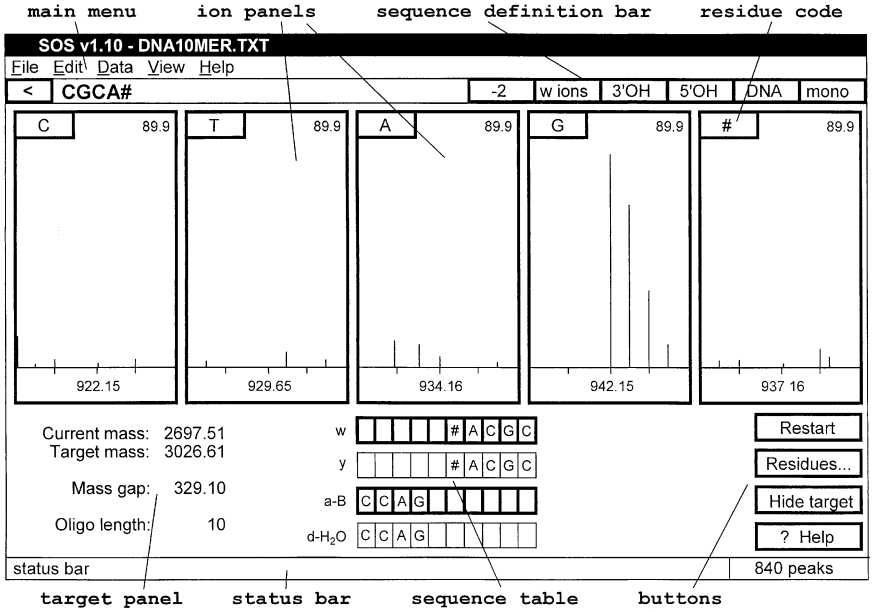

Figure 1. SOS graphic display generated during sequencing of the DNA oligonucleotide 5'-CCAGG#ACGC-3', where # is a RNA residue 2'-*O*-methylcytidine (rCm), which has been defined under residues. The ion panels display experimental ion patterns for candidates for the sixth residue from the 3' end, for doubly charged w ions in the monoisotopic mass scale. Cumulative assignments from the w and a–B ion series at the point of processing shown in Figure 1 account for 2697.51 Da, displayed in the target panel. From the user entered experimental molecular mass of 3026.61, the missing fifth nucleotide mass is calculated as 329.1 Da (the residue mass of dGp). If G is accepted as the next residue from ion panel data (also supported in the w_6^1 panels, not shown), the gap mass will be near zero. Further extension of the y ion series in the 3' → 5' direction by GG (not shown) is readily made, which provides limited sequence overlap in both directions, and leads to the full oligonucleotide candidate sequence 5'-CCAGG-rCm-ACGC-3'.

- Increasing ambiguities in assignment of sequence ions at longer chain lengths (e.g., 10–20 residues), or when low intensity spectra have been acquired, either of which is a particular problem toward the center of the molecule.
- The possibility of mass versus composition redundancies (see listing in reference [7]), such that an ion may have more than one rational assignment (for example, in RNA when the termini combination 5'-A. . .Gp-3' occurs, ion w_1 [*m/z* 442.201] cannot be readily distinguished from a_2-B_2 [*m/z* 442.301]).
- A priori use or input of incorrect initial parameters, such as phosphorylation state of the termini, or presence of unknown modifications.

We report here on the development of a user-interactive computer program, Simple Oligonucleotide Sequencer (SOS) for the determination of oligonucleotide sequence from CID mass spectra, including in-source (or nozzle-skimmer) type spectra, and is well suited for approaching the types of difficulties listed above. Execution of the program is simple and permits rapid inspection and assignment for each step of the mass ladder in either direction using multiple ion series and user-defined modifications for DNA or RNA. The data are presented for user interrogation such that alternate or seemingly less likely sequences can be explored, and the results from different ion series in the same or different chain directions visually compared and tested against experimental or calculated molecular mass values. The program permits residue-by-residue application of previously introduced rules of data interpretation to assemble and test sequence candidates, in particular when the sequence is not known in advance [7]. This point is important because of the relative ease with which misassignments can be made when large numbers of minor fragment ions are present, and are simply fitted to a preconceived sequence.

Experimental

The SOS program was written in Delphi, and will run on most Windows platforms such as 3.11, WIN95, WIN98, and WINNT/WIN2000. Data files utilized consist of text files containing mass/abundance lists of background-subtracted and centroided mass spectra. The identity of the operating system (e.g., PC- or MAC-based) used for data acquisition is of no consequence once the data have been converted to mass/abundance lists. Mass spectra used for program development and testing were obtained from triple quadrupole (Micromass Quattro II) and quadrupole/orthogonal acceleration time-of-flight (Micromass Qtof2; data in Figures 1 and 2) instruments, using

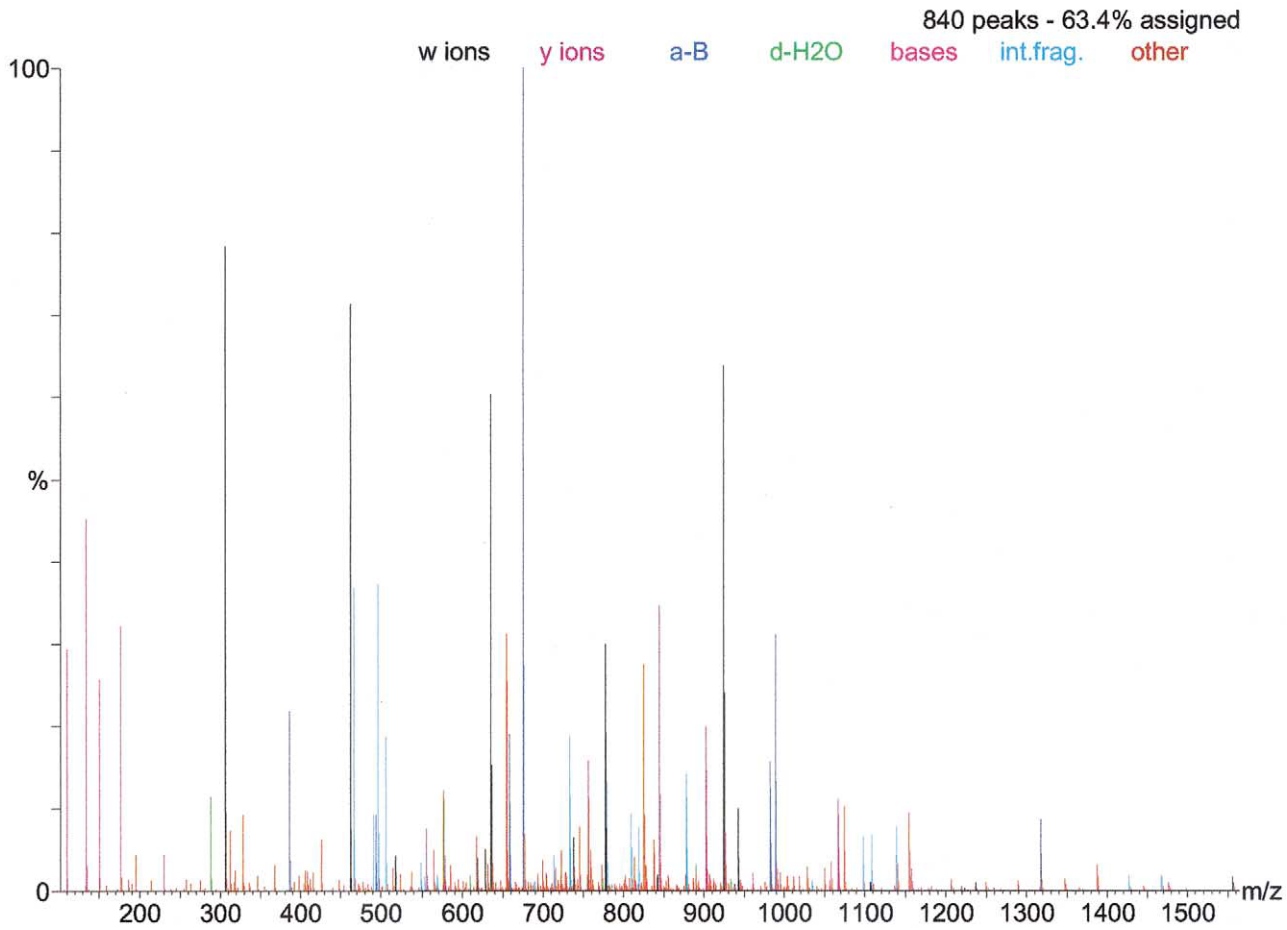

Figure 2. Display of the mass spectrum analyzed in Figure 1, in the 'spectrum view' mode, showing assigned ions that were color-coded at the time of assignments in the 'sequencer view' mode. Ion series: w, black; y, purple; a–B, blue; d–H_2O, green; bases (for example, *m/z* 134 for adenine), magenta; internal fragments of the type pNpf (where f is furanyl, from base loss [7, 8]), light blue; unassigned ions, red. The spectrum contains 840 peaks, of which the absolute intensity of the base peak is 891.9 counts, and 63.4% of the ion current has been assigned.

electrospray ionization in negative ion mode. Spectra were obtained from conventional collision-cell based CID, as well as from elevated nozzle-skimmer potential (in-source) fragmentation [8].

Discussion

The program contains two main parts: 'sequencer view' for construction of the sequence, and 'spectrum view' used to display the full mass spectrum, bearing color-coded peak assignments. An example of the 'sequencer view' presentation is shown in Figure 1 and of the 'spectrum' view in Figure 2 (both adapted from the tutorial section of the program).

The central operating mode of the program consists of the graphic presentation of limited segments of the mass spectrum, which have been computer-extracted from the full mass spectrum, and calculated for each step of sequence construction, in four (or more) panels representing four (or more) possible nucleotide residues in the ladder (see Figure 1). Using parameters that have been entered in the sequence definition bar and residues table (via the residues button) the theoretical *m/z* values corresponding to each possible residue for the next addition to the chain is defined on the mass axis in each ion panel. The degree of alignment of the observed mass spectral peak with the calculated value (e.g., *m/z* 942.15 in Figure 1) reflects the error of mass measurement relative to the putative assignment shown (G in Figure 1). Selection of the next residue in the chain is made by the user, based [7] primarily on the mass measurement error, the number of different charge states observed for the candidate assignment (−2 charge state is shown in Figure 1), and the abundances of those ions. The user can rapidly switch between charge states and ion type (e.g., w series versus y series) to aid in the selection of each residue in the chain. Assignment of residues is tabulated in the sequence table for comparison of results from different ion series, including bi-directional sequence overlap. The calculated mass of the growing chain is automatically updated in the target panel; and the gap mass,

representing unassigned residues compared to an experimental or calculated molecular mass (target mass) if available, is calculated and displayed. An important element of the program is the ability to define additional residues that are mass-modified in the base, sugar, or backbone. For example, in Figure 1 a new sugar residue, *O*-methylribose, has been defined in lieu of deoxyribose while keeping the unmodified base cytosine, as denoted by #. Modifications to the backbone include phosphorothioates, which are widely employed in studies of oligonucleotide therapeutics, and can be used alone or mixed with other backbone moieties such as phosphodiester. The modified residues as defined are then carried through all appropriate calculations for ion series and molecular mass.

A particular advantage of this method of user-directed data manipulation is the ability to easily erase and rebuild alternate sub-sequences for comparison. For example, if candidate peaks representing a given sequence position are absent or ambiguous, any of the four or more nucleotide possibilities can readily be entered and tested for suitability to anchor further extension of the mass ladder. As assignments are made in the sequence table the peaks in the spectrum can be color-coded, which serves two purposes. First they bring to the user's attention, as different ion series are investigated, the fact that an ion may have been previously assigned. Second, in the 'spectrum view' presentation (Figure 2) ion series assignments are readily distinguished and the presence of major unassigned ions in the spectrum easily recognized.

Although the principal reason for the development of SOS was the creation of a user-friendly tool for ab initio oligonucleotide sequencing, it is also useful in studies of fragmentation processes of model oligonucleotide derivatives, and can serve as a molecular mass calculator. The program, including installation instructions, HELP and TUTORIAL files with sample mass spectra, are available from either author.

Acknowledgments

This work was supported by the National Institutes of Health Grant GM29812 and a gift from NeXstar Pharmaceuticals. The authors thank P. F. Crain for her critique of the manuscript, software execution and associated text files.

References

1. Nordhoff, E.; Kirpekar, F.; Roepstorff, P. Mass Spectrometry of Nucleic Acids. *Mass Spectrom. Rev.* **1996,** *15,* 67–138.
2. Limbach, P. A. Indirect Mass Spectrometric Methods for Characterizing and Sequencing Oligonucleotides. *Mass Spectrom. Rev.* **1996,** *15,* 297–336.
3. Limbach, P. A.; Crain, P. F.; McCloskey, J. A. Characterization of Oligonucleotides and Nucleic Acids by Mass Spectrometry. *Curr. Opin. Biotechnol.* **1995,** *6,* 96–102.
4. McLuckey, S. A.; Van Berkel, G. J.; Glish, G. L. Tandem Mass Spectrometry of Small, Multiply Charged Oligonucleotides. *J. Am. Soc. Mass Spectrom.* **1992,** *3,* 60–70.
5. McLuckey, S. A.; Habibi-Goudarzi, S. Decompositions of Multiply Charged Oligonucleotide Anions. *J. Am. Chem. Soc.* **1993,** *115,* 12085–12095.
6. Murray, K. K. DNA Sequencing by Mass Spectrometry. *J. Mass Spectrom.* **1996,** *31,* 1203–1215.
7. Ni, J.; Pomerantz, S. C.; Rozenski, J.; Zhang, Y.; McCloskey, J. A. Interpretation of Oligonucleotide Mass Spectra for Determination of Sequence Using Electrospray Ionization and Tandem Mass Spectrometry. *Anal. Chem.* **1996,** *68,* 1989–1999.
8. Lotz, R.; Gerster, M.; Bayer, E. Sequence Verification of Oligodeoxynucleotides and Oligophosphorothioates Using Electrospray Ionization (Tandem) Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **1998,** *12,* 389–397.